JCTC Journal of Chemical Theory and Computation

Coarse-Graining the Accessible Surface and the Electrostatics of Proteins for Protein−Protein Interactions

Francesco Pizzitutti,[†,‡] Massimo Marchi,*[,†] and Daniel Borgis*[,‡]

Commissariat à l'Énergie Atomique, DSV-DBJC-SBFM, CNRS URA 2096,
Centre d'Études Saclay, 91191 Gif-sur-Yvette Cedex, France, and Laboratoire
d'Analyse et Modélisation pour la Biologie et l'Environment, CNRS UMR 8587,
Université Evry-Val-d'Essonne, Boulevard François Mitterrand, 91405 Evry, France

Received May 21, 2007

Abstract: This study is concerned with the development and test of a coarse-grained representation specifically constructed for proteins and peptides, where each amino acid of the sequence is represented by a charged dipolar sphere. The model was parametrized from the physical properties of individual amino acids and applied to the study of the interaction between solvated proteins. Using an implicit solvent approach and our coarse-grained model, we computed the potential of mean force for the association of well-known proteins, such as the Cu−Zn superoxide dismutase, lysozyme, and basic pancreatic trypsin inhibitor, and a peptide, Aβ_7. The coarse-grained potentials of mean force were systematically compared with their all-atoms counterpart. For both the polar and nonpolar contributions to this potential, the results of our calculations show that the coarse-grained model provides a good approximation of the all-atoms potential when the distance between the molecule surfaces is greater than a solvent molecular diameter. For shorter distances and for specific interactions, like those found between the SOD monomers, the electrostatic desolvation effect appears to be underestimated by our coarse-grained representation. The possibility of a very short range all-atom refinement to better describe the interaction at close contact is explored. We find also that the most important contribution to the association free-energy comes from the hydrophobic solvent accessible surface area term, which is well reproduced by our coarse-grained approach.

I. Introduction

The development of molecular modeling has provided valuable insights to the understanding of interactions between biomolecules, proteins, and DNA. Because biological processes cover a broad range of time and length scales, progress needs to be made to adapt computational techniques to different levels of detail in the description of the systems. Indeed, in standard molecular dynamics (MD) simulations, solvent and biomolecules are represented at the atomic level. This implies that for an average size protein, the number of simulated atoms can easily reach a few tens of thousands atoms. Despite the ever increasing power of computers and the considerable effort undertaken to ameliorate molecular modeling algorithms, the time and length scales currently approachable in atomistic simulations are limited to a few hundred of angstroms and a few tens of nanoseconds, respectively. This means that even processes implying interactions between proteins and between proteins and DNA, crucial to life on earth, are currently out of reach of all-atoms (AA) computer simulations.

Although protein−protein interfaces have been widely studied[9,10] and calculations of free energy of association are now customary in the analysis of protein−ligand interac-

* Corresponding author e-mail: Massimo.Marchi@cea.fr (M.M.); Daniel.Borgis@univ-evry.fr (D.B.).

[†] Centre d'Études Saclay.

[‡] Université Evry-Val-d'Essonne.

10.1021/ct700121n CCC: $37.00 © 2007 American Chemical Society
Published on Web 08/24/2007

tions,[11] the study of protein−protein associations involving all-protein and solvent atomic degrees of freedom are not easily manageable by an all-atom simulation. Indeed, the partial desolvation of the molecular surfaces occurring in the process of protein association requires very long simulation times.

A possible strategy to increase the length and time scales spanned by molecular simulation is to reduce the level of detail of the atomic systems. Formerly, the most common approach has been to remove non-relevant degrees of freedom from the simulated system: for example, see ref 1. As an example, in some of these coarse-grained (CG) models the protein amino acids are represented with one to six centers only, whereas the interaction energies among residues is knowledge-based, procured from the database of the many native structures of proteins available today. Such a class of models has been used in docking[2] and folding[3] studies. In particular, most of the docking methods are based on some energetic scoring function relying on a simplified picture of the interactions involved. This goes from simple pattern recognition and simplified electrostatics to all-atom force fields at some final stage of the minimization process. In this context, Zacharias[2] recently developed an empirical coarse-grained representation of proteins in terms of a few effective sites per residue. Following a different, but related approach, Klein and co-workers developed effective CG interaction potentials for phospholipids,[4] on the basis of results procured from all-atom models.

Because most of the biology occurs in a wet milieu, other CG methods have been developed where the explicit molecular representation of the solvent is replaced by a static mean-field representation. These approaches provide a simple way to compute the solvation free energy for any given solute. In the most common Poisson−Boltzmann surface-area (PB-SA) approximations, the polar solvent is represented as a structureless continuum medium of dielectric constant ϵ_s, whereas the solute is well described by a point-charge distribution embedded in a medium with dielectric constant ϵ_p. In general, the electrostatic boundary between solvent and solute, in this view, regions of high and low dielectric constant, respectively, is defined as the molecular surface accessible to the solvent. The nonpolar (hydrophobic) contributions to the solvation free energy are assumed to be proportional to the solvent accessible surface area (SASA) of the solute.

In this work, we present a CG model of proteins, similar in the spirit to that developed by Song.[7,8] Our primary objective here is the study of the protein−protein interactions. Indeed, the understanding of the forces between solvated proteins and, more generally, between biomolecules is crucial to phenomena such as cell signaling and protein crystallization, which are driven by noncovalent specific protein−protein interactions. Thus, our approach does not aim, at least at first, at dealing with protein−protein docking.

Ours is a so-called bottom-up approach, which proceeds in much the same way that atomistic potentials are derived from quantum chemistry ab initio calculations: We compute the relevant microscopic properties of associated proteins from atomistic modeling and then fit the parameters of our CG model. This, we hope, will ensure the highest level of accuracy and transferability. Our final objective is to produce a CG residue-based force-field, which provides a good approximation of the potential of mean force (PMF) for protein−protein nonspecific interactions. Thus, we adopt a CG model where each amino acid residue is replaced by a van der Waals sphere, a charge, and a dipole. The corresponding parameters are fitted to reproduce relevant all-atom properties of the systems.

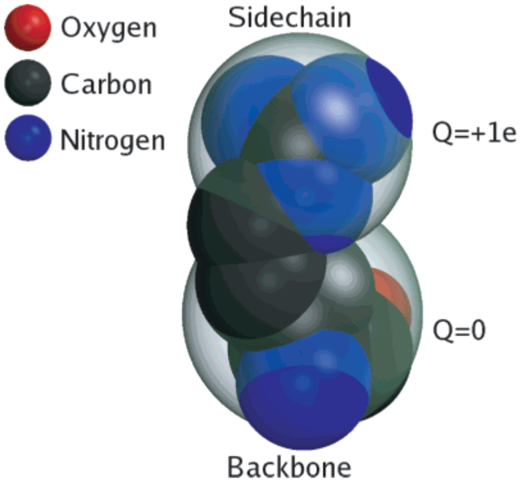

Figure 1. SASA of an arginine residue in its AA (red, black, and blue opaque balls) and CG (gray transparent balls) representations. Note that at pH 7 the arginine has a total net charge equal to −1e. Accordingly, this amino acid is represented in our CG approach by two CG spheres: The first one, labeled $Q = 0$, models the uncharged backbone part of the residue, whereas the second one, labeled $Q = +1e$, models side-chain atoms where the excess charge of −1e is distributed.

Our CG model is tested here on three widely studied proteins, Cu−Zn superoxide dismutase (SOD), lysozyme, and basic pancreatic trypsin inhibitor (BPTI), as well as on a small segment of the amyloid-forming peptide $A\beta_7$. The test consists of a comparison of the potential of mean force (PMF) for homodimer association, obtained using either a CG or AA representation with fixed internal geometries.

II. Materials and Methods

A. Coarse-Grained Model. We built the CG representation of a protein by associating one CG sphere to every uncharged residue and two CG spheres to every charged residue (Lys, Arg, Glu, and Asp residues and the protein terminals, which were considered to be charged at pH 7). For a neutral residue, the CG sphere was centered on the center of mass of the AA residue. For a charged residue, the first sphere was centered on the center of mass of the neutral part of the residue, whereas the second sphere was centered on the center of charge of the charged part (see Figure 1 for a representation of the CG model of an arginine residue). Every CG sphere was characterized by a radius, R_{CG}, a charge, Q_{CG}, and an electric dipole, $\mathbf{P}_{CG}$.

Coarse-Grained Radii. For any given protein, the R_{CG} values were chosen to procure a CG solvent-accessible surface area for each CG residue of the protein ($SASA_{CG}$)

Table 1. CG Parameters for the Aβ7 Peptide[a]

	residue	SASA$_{AA}$ (Å^2)	SASA$_{CG}$ (Å^2)	R_{CG} (Å)	p_{AA} (e Å)	P_{CG} (e Å)
1	Lys0	246.9	235.0	4.0	0.79	0.77
2	Lysq	95.7	121.1	2.4	0.00	0.00
3	Leu	139.0	104.9	3.6	0.66	0.68
4	Val	131.2	118.0	3.4	1.03	1.41
5	Phe	209.6	211.1	4.0	0.54	1.30
6	Phe	157.6	149.0	4.0	0.28	1.58
7	Ala	88.4	75.6	3.0	1.81	0.93
8	Glu0	123.1	127.7	3.2	0.98	0.89
9	Gluq	115.9	127.4	2.7	0.00	0.00

[a] The table presents the 9 CG radii (R_{CG}) and the absolute values of the 9 optimized CG dipoles (P_{CG}) for the Aβ_7 peptide. Note that Lys1 and Glu9, two charged residues, are represented by two CG spheres each, Lys0 and Glu0, with a dipole in their center and (Lysq and Gluq) with only a charge in their center. The AA and CG SASAs, SASA$_{AA}$ and SASA$_{CG}$, respectively, and the absolute values of the AA dipoles are also reported.

equal to its corresponding AA SASA$_{AA}$. The latter is defined as the SASA of the residue in the actual conformation it has in the native structure of the given protein. In particular, the SASA$_{AA}$ of a given residue was computed after the atomic radii of all remaining protein residues are switched off. Throughout this paper, SASA is defined as the surface traced by the center of a rolling probe sphere of a radius of 1.4 Å over the van der Waals surface of the solute molecule itself.[14] From the SASA$_{AA}$, we extract the R_{CG} through the relation

$$R_{CG} = \sqrt{\frac{\mathrm{SASA_{AA}}}{4\pi}} - 1.4 \tag{1}$$

where the radii are expressed in Å.

For the test proteins considered in this work, the obtained R_{CG} values in a range from 2 to 4 Å. As an example, the radii distribution for the Aβ_7 peptide are reported in Table 1.

Coarse-Grained Charges and Dipoles. The set of CG charges, $\{Q_{CGi}\}$, and CG dipoles, $\{\mathbf{P}_{CGi}\}$, was obtained by optimization of the electrostatic potential generated by the CG systems around the protein to reproduce, most accurately, the electrostatic potential generated by the AA system. More specifically a mean square deviation function, $\Psi(Q_{CG}, \mathbf{P}_{CG})$, depending on the set of the $\{Q_{CGi}\}$ and $\{\mathbf{P}_{CGi}\}$, was defined as

$$\psi(Q_{CG}, \mathbf{P}_{CG}) = \left(\sum_{i=1}^{N_{res}} \sum_{k=1}^{N_{grid}} \phi_{CG}\,(Q_{CGi}, \mathbf{P}_{CGi}, \mathbf{x}_k) - \phi_{AA}(q_{AAi}, \mathbf{x}_k)\right)^2 \tag{2}$$

where the electrostatic potentials $\phi_{AA}(q_{AAi}, \mathbf{r})$ and $\phi_{CG}(Q_{CGi}, \mathbf{P}_{CGi}, \mathbf{r})$, generated by the AA and CG systems, respectively, are discretized over a grid and evaluated at every point $\mathbf{x}_i$ of the grid. N_{res} is the total number of residues and N_{grid} is the number of grid points. The discretization grid was built inside an orthorombic box containing the AA system. The dimensions of the box were chosen in such way that the AA system was contained in 75% of the box volume. The distance between the grid points was always between 0.5 and 1.0 Å, depending on the system considered. The deviation function $\Psi(Q_{CG}, \mathbf{P}_{CG})$ was then minimized with respect to the set $\mathbf{P}_{CGi}$ through a conjugated gradient algorithm. We point out that the CG charges were not optimized, but were constrained to their assigned values.

B. Interaction Free Energy Calculations. To evaluate the interaction free energy between biomolecules, we used an implicit solvent approach. As noted by Roux and Simonson,[15] the total solvation free energy of a molecule in water can be expressed as the sum of a polar term, resulting from electrostatic interactions, and a nonpolar term, from the Van der Waals and hydrophobic contributions: $\Delta G_{solv} = \Delta G_{elect} + \Delta G_{np}$.

From this relation, we can obtain the interaction free energy, G_{int}, between two molecules, A and B, at a distance r, as the change in ΔG_{solv} when the two molecules approach each other to a distance r from infinity plus the direct electrostatic interaction V_{int}. The resulting expression for the interaction free energy is

$$G_{int} = \Delta\Delta G_{solv} + V_{int} = \Delta\Delta G_{elect} + \Delta\Delta G_{np} + V_{int} \tag{3}$$

Here, we have

$$\Delta\Delta G_{solv} = \Delta G_{solv}^{A+B}(r) - \Delta G_{solv}^{A} - \Delta G_{solv}^{B} \tag{4}$$

$$\Delta\Delta G_{elect} = \Delta G_{elect}^{A+B}(r) - \Delta G_{elect}^{A} - \Delta G_{elect}^{B} \tag{5}$$

$$\Delta\Delta G_{np} = \Delta G_{np}^{A+B}(r) - \Delta G_{np}^{A} - \Delta G_{np}^{B} \tag{6}$$

where $\Delta G_{solv}^{A+B}(r)$ and $\Delta G_{solv}^{A,B}$ denotes the solvation free energy of a system composed of two molecules, A and B, when the distance between the two molecules is r and when they are isolated, respectively. The notation is the same for the electrostatic and nonpolar contributions, ΔG_{elect} and ΔG_{np}, respectively. It should be noted that, in principle, the interaction free energy definition of eq 3 does not account for the totality of the free energy changes caused by the interactions between the two molecules. Specifically, the following contributions from the interaction between the two molecules are neglected: the cost of structural reorganization, the loss of configurational reorganization entropy, and the loss of the translational and rotational degrees of freedom of the two molecules. In what follows, these contributions will not be taken into account, which is equivalent to considering the two interacting molecules to be rigid bodies. Thus, our calculations consider only contributions to the interaction free energy coming from direct molecule−molecule interactions and desolvation effects.

Polar Interaction Energy Contribution. The electrostatic term, G_{elect}, is defined as

$$G_{elect} = \Delta G_{elect}^{A+B}(r) - \Delta G_{elect}^{A} - \Delta G_{elect}^{B} + \frac{1}{\epsilon_p} \sum_{i\epsilon A, j\epsilon B} \frac{q_i q_j}{|x_i - x_j|} \tag{7}$$

In this work, we calculated the first three terms in a continuum solvent context by numerically solving the Poisson−Boltzmann (PB) equation.[16] Following this approach, the water and the interior of molecules are treated as dielectric continuum of dielectric constant ϵ_s and ϵ_p, respectively. The volume occupied by the solute is defined by its molecular surface (obtained using a solvent radius of 1.4 Å). This volume contains, in full, the partial charges of the solute atoms.[17,18] The resulting electrostatic problem is treated numerically by solving the related Poisson equations with the ionic force set to zero. The last term in eq 7 is the pairwise Coulomb free energy between the two interacting molecules embedded in a dielectric medium of identical dielectric constant that their interior. The indexes i and j refers to molecules A and B, respectively; q and $\mathbf{x}$ are the charge and the position of the particles.

Non-Polar Interaction Energy Contribution. The non-polar term accounts for free energy differences resulting from both formation of cavities in the solvent and van der Waals interactions after the removal of the solvent from the interface between two interacting proteins. This term can be calculated according to eq 6. Every term on the right-hand side of this equation can be calculated using the well-established linear relation[19,20] that connects experimental hydration energies of small alkanes chains with their surfaces

$$\Delta G_{\text{np}} = \gamma \text{SASA} + b \tag{8}$$

where SASA is the solvent accessible surface of the molecule and γ and b are constants.

In summary, the final form of our interaction free energy is

$$G_{\text{int}} = G_{\text{elect}} + \Delta\Delta G_{\text{np}} = G_{\text{elect}} + \gamma\Delta\text{SASA} \tag{9}$$

C. System Setups and Calculation Parameters. In this work, we have tested systematically our CG protein model by computing the interaction free energies between proteins using either the AA or CG representation. The systems chosen to test the accuracy of the CG model are small biomolecules, such as the 16−22th residue segment in the amyloid forming peptide Aβ_7 associated to the Alzheimer's disease,[21] and three globular proteins, Cu−Zn superoxide dismutase (SOD) from *Photobacterium leiognathi*,[22] egg-white lysozyme from *Gallus gallus* (Protein Data Bank 2lym),[23] and basic pancreatic trypsin inhibitor (BPTI) from *Bos taurus* (Protein Data Bank 1bpi).[24] For each molecule, we have generated a homodimer at close contact, and we have progressively increased the distance between the two monomers. The interaction free energy was calculated approximately every 0.5 Å until a maximum interdimer edge-to-edge distance of 25 Å. All degrees of freedom of the system, other than the dimer separation, were frozen. For technical reasons, the CG dipoles on every uncharged residue were represented by two identical charges of opposite sign located at the residue center of mass and placed at a distance of 1 Å from each other.

Among the chosen biomolecules, only SOD and the Aβ_7 peptide are known to form stable aggregates in solution. Aβ_7 organizes itself in antiparallel β-sheet fibrils,[21] whereas SOD yields a highly stable homodimer. In the case of SOD, we started our calculations from the structure of the SOD dimer presented in ref 22 that was equilibrated through an MD run. For the other proteins, an initial homodimer structure was built by superimposition of two identical AA single molecules configurations and then translation of one of them along the x-axis until the minimum distance between two atoms belonging to different halves was greater than 1.8 Å. Figure 2 displays the AA initial configuration of the SOD dimer, superimposed with the corresponding CG model. For lysozyme and BPTI, the AA geometry of the individual proteins was taken from a X-ray configuration from the PDB database. The molecule orientations with respect to the reference axis were the same as those in the PDB files. For the Aβ_7 peptide, the starting structure was obtained as a result of the equilibration of a MD run performed with the ORAC[25] program, in the NPT ensemble, using the CHARMM27[26] force field. The total length of the run was 500 ps, and the peptide was solvated by 543 tip3[27] explicit waters. The longitudinal axis of the obtained relaxed structure was then oriented parallel to the reference z-axis.

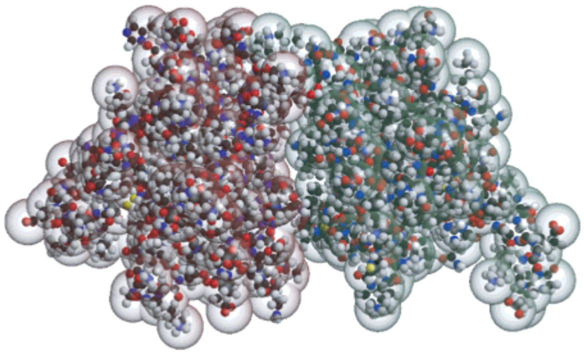

Figure 2. Pictorial representation of the coarse-grained SOD dimer. The AA SOD dimer (ball-and-stick) is superimposed to the CG representation (transparent, gray spheres).

Partial charges and radii for the AA systems were assigned according to the CHARMM27 force field.[26]

To eliminate numerical errors caused by the charge distribution interpolation used in the Poisson−Boltzmann solver, we computed the electrostatic contribution to the interaction energy in eq 5 through the relation

$$G_{\text{elect}} = \Delta E^{\text{A+B}}_{\text{elect}} - \Delta E^{\text{A}}_{\text{elect}} - \Delta E^{\text{B}}_{\text{elect}} \tag{10}$$

where, for a fixed dimer configuration, the three terms correspond to the total electrostatic energy E (solvation energy, plus direct Coulomb interaction, plus grid self-energy) of A + B, A, or B, respectively. This procedure is more time-consuming than the direct application of eq 7 because it involves a recalculation of the A and B isolated contributions for each distance, but it eliminates numerical errors resulting from finite grid effects. The solution of the linearized PB equation was calculated through the finite differences version of the APBS[28] program.

The radius of gyration, R_{g}, of the considered considered molecules ranges from ~9 Å for the small Aβ_7 peptide to ~20 Å for SOD, whereas the volume spanned by the pairs of interacting molecules also depends on the distance between the two partners. In the Poisson−Boltzmann calculations, for every molecular pair and distance, the dimensions of our orthorombic discretization grid a, b, and c are chosen such that the two molecules are contained in 75% of the volume occupied by the grid. Moreover, the distance between consecutive grid points is always kept below 0.5 Å.

In the PB calculations, the dielectric constants of solute and solvent are of crucial importance. The latter was set to that of water, $\epsilon_s = 78.5$, while the former, ϵ_p, crucially depends on the model of solute used in the calculation.[29] While a solute dielectric constant of 2 is usually used to account for the electronic polarizability of the atoms, larger solute dielectric constants are employed in PB calculations to account for the atomic and orientational polarizability of the protein matrix.[16] In the past, ϵ_p values ranging from 4 to 20 have been used to study protein−protein interactions,[30] and $\epsilon_p = 17$[31] was used in molecular dynamics simulations of proteins. We remind the reader that the purpose of this work is to assess the validity of our CG description of biomolecules. This is done by comparison with the corresponding AA model under identical dielectric conditions. Thus, in all our PB calculations, the dielectric constant of the solute was set to $\epsilon_p = 2$ for both AA and CG systems.

As far as the calculation of the nonpolar terms was concerned, the value of the constant γ was set to 0.24 kJ/mol Å^2 for both the CG and AA systems, as previously done by several authors.[32,33] The ionic strength was set to 0 in all PB calculations.

All the calculations were performed on a 2.2 MHz Pentium processor. For SOD, the overall calculations performed for all distances took about 7 h of CPU time.

III. Results and Discussion

A. Bare Electrostatic Energy. The first step needed to build the CG model of a biomolecule is the calculation of the charge and dipole for every CG residue. Through the optimization process described in the Materials and Methods section, we have obtained the complete charges and dipoles distribution of the four test CG systems. We note that amino acid charges depends on the residue ionization state. In our calculations, the molecular ionization state was fixed to the average ionization state at pH 7. Thus, during the optimization procedure, CG charges were constrained to their AA original values at pH 7, and the residue dipoles were considered as the free variables.

As an example, in Table 1 we have reported, the modulus of the nonoptimized Aβ_7 peptide dipoles, p_{AA}, as they come directly from the point-charge distribution of the CHARMM27 force field, and the corresponding optimized quantities, P_{CG}. We notice that the optimization procedure has already a non-negligible effect on the electrostatic representation. After dipole optimization, we calculated the bare electrostatic interaction in a medium of uniform dielectric constant $\epsilon_s = 2$, that is, V_{int} in eq 3. In Figure 3, we compare the CG and AA models by displaying V_{int} as a function of the monomer−monomer distance, r in the picture, for our four biomolecules. The dimers for BPTI, lysozyme, and Aβ_7 were formed by pulling away one of the two monomers along the x-axis direction from an initial configuration where the two monomers were superposed. Given that the SOD dimer structure is known experimentally, the dimer structures were formed from the initial structure by translation along the intermonomer direction of one of the two monomers. Thus, the zero point for the distance in the SOD case corresponds to the experimental relaxed structure, whereas for the remaining proteins, $r = 0$ is chosen as the conformation for which the minimum distance between two atoms belonging to two distinct monomers is 1.8 Å.

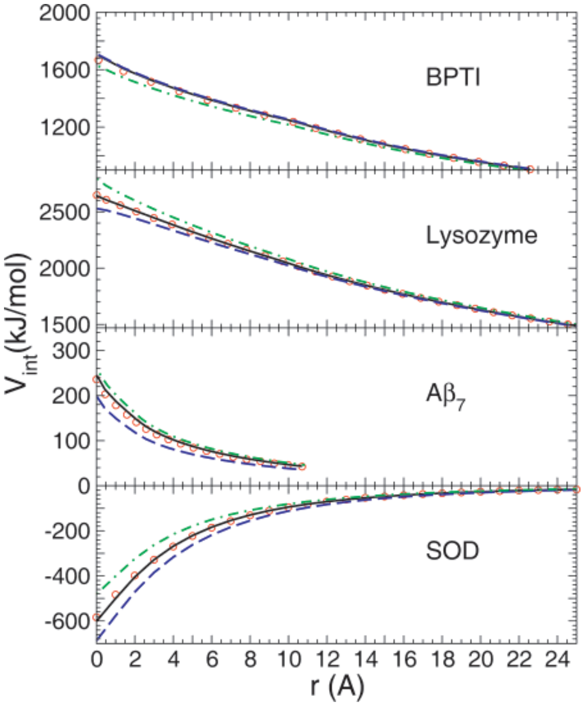

Figure 3. Dependence of the bare electrostatic interaction energy on the intermonomer distance, r. Starting from the initial reference configurations at $r = 0$ (see text), for all dimers, r was increased along a direction perpendicular to the dimer separation plane. In the pictures, the circles correspond to the AA systems and the solid curve to the full CG systems composed of charges and optimized dipoles. The other two curves corresponds to two models with nonoptimized dipoles (dashed line) and without dipoles (dashed-dotted line).

Figure 3 shows that the CG dipoles obtained by the optimization procedure described in section II are able to yield a very accurate representation of the bare electrostatic interaction energy. The same figure also shows the energy functions computed with nonoptimized dipoles and the ones obtained by setting all dipoles to zero. The latter corresponds to the electrostatic part of the coarse-grained models of proteins by Zacharias et al.[2] For the models including dipoles, we notice that both are able to yield the correct asymptotic behaviors, whereas the presence of the optimized dipoles is necessary to accurately reproduce the AA electrostatic fields at all distances. Curiously, the nonoptimized dipole curve comes out very close to the optimized one for BPTI, and the same observation is true for the results obtained with charge-only model for Aβ_7. Note also that the bare electrostatic interactions are repulsive for BPTI, lysozyme, and Aβ_7, whereas for SOD, they are quite attractive. Indeed, the SOD dimer is stabilized by many H-bonds and favorable electrostatic interactions between groups of opposite charge, located at the interface between the two SOD monomers.

B. Global Electrostatic Interactions. The next step in testing our CG models is to build solvated models of CG proteins and to calculate the overall electrostatic contribution to the interaction free energy as a function of the interdimer

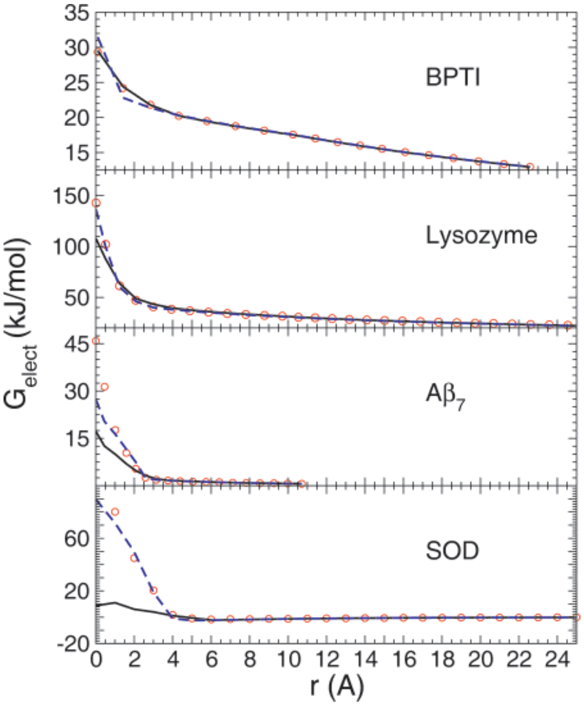

Figure 4. Total electrostatic interaction as a function of the intermonomer distance. We present results for the AA model (circles), the CG model (solid continuous line), and an intermediate model (dashed continuous line) with CG charges and dipoles, but with AA dielectric boundaries.

distances. To do this, it is necessary to add to the description of our CG systems an excluded volume component, in our case a van der Waals radius for every CG residue. This CG radii is computed as described in the Materials and Methods section. Typical radii are similar to those shown in Table 1 for the Aβ_7 case. For the other systems, the CG residue radii range from 2.3−2.5 Å for charged residues to 2.5−4.5 Å for non-charged residues.

With these excluded volume parameters to define the dielectric boundaries, the total electrostatic free energy, including solvation free energies and direct interactions as of eq 7, can be computed for the various dimer configuration using the PB solver. The resulting curves are reported in Figure 4. Two CG calculations are compared there to the AA reference (red circles). In one curve (dashed lines), the charge distribution is coarse-grained in terms of CG charges and dipoles, but the dielectric boundaries are those of the AA model. As expected, given the high accuracy of the CG charge distribution, the comparison is very favorable for all distances, even close to contact: the worst agreement involves the Aβ_7 peptide. The continuous curve accounts for the full CG model including the full coarse-graining of the electrostatic components and that of the molecular surface. Remarkably, this CG model reproduces very well the AA electrostatic interactions beyond a certain threshold distance. For the two proteins that are known to yield specific association, Aβ_7 and SOD, the threshold appears around 2 and 4 Å, respectively. For shorter distances, the CG curve underestimates the AA free energy. This discrepancy is caused by desolvation, which is accounted for in different ways in the AA and CG approaches. Even for identical electric fields, the different molecular surfaces leads to different surface polarization charges and thus different solvation free energies. For dimers where nonspecific interactions are dominant, such as BPTI and lysozyme, the accuracy of our CG model is larger at short distances. Indeed, for BPTI, there is a good agreement between AA and CG over the whole range of distances considered, whereas for lysozyme larger deviations are observed very close to contact.

Given the good agreement between the electrostatic energy of CG and AA in the homogeneous dielectric, the short-distance discrepancies, especially in the SOD dimer curve, are caused by desolvation effects. Thus, it is interesting to discuss those effects in more detail. When the distance between two atoms become smaller than the sum of the two atoms radii plus a solvent molecule diameter, the solvent molecules in the inner volume among the atoms start to be removed, and the atoms start to interact electrostatically through the low dielectric medium of the protein inner space. The change in the interaction free energy caused by this process can be separated into nonpolar and polar contributions. The nonpolar part, $\Delta\Delta G_{\mathrm{np}}$, will be treated in the next section. The polar part can be represented as the change in the electrostatic solvation energy of two molecules as they approach each other. Hence the desolvation polar part corresponds to the first three terms of eq 7.

A comparison between the curves in Figures 3 and 4, especially for SOD, shows that, even when the bare electrostatic interactions are favorable to the formation of the complex, the desolvation process can make the overall electrostatic interaction unfavorable. The net effect is more remarkable because the number of buried polar and charged groups located at the interface between the interacting molecules increases. Therefore, even when the flexibility of the protein matrix could rearrange the side-chains and the backbone to reduce the unfavorable electrostatic solute−solute interaction, behaviors like those in Figure 4 are to be expected: a sharp increase of the electrostatic effective interaction when two solvated molecules are closer than the desolvation threshold. In the case of SOD, the desolvation terms at $r = 0$ are equal to 396 and 310 kJ/mol for the AA and the CG models, respectively. These high energetic costs come from the desolvation of about 13% of the molecular surface in both models following the dimer formation. On the other hand, the direct Coulombic interaction energies with $\epsilon_s = 2$ amounts to −292 and −301 kJ/mol, respectively. Therefore, even though the CG gives a CG electrostatic field that differs by only 2% from the AA electrostatic field, the desolvation contributions are farther apart, by about 20%.

As we have seen earlier, using the CG charge distribution with the true AA dielectric boundaries does give excellent results, so that the deviation between AA and CG models is a direct consequence of the slight difference between the AA and CG molecular surface topologies. To increase the agreement between CG and AA, it is tempting to increase slightly the resolution and use a “finer” representation with, that is, 2−3 grains per residue instead of 1−2 in the present case. To test that proposition, we have used the finer CG residue definition of Zacharias[2] and have applied the same methodology as in section II.A to assign grain radii and grain

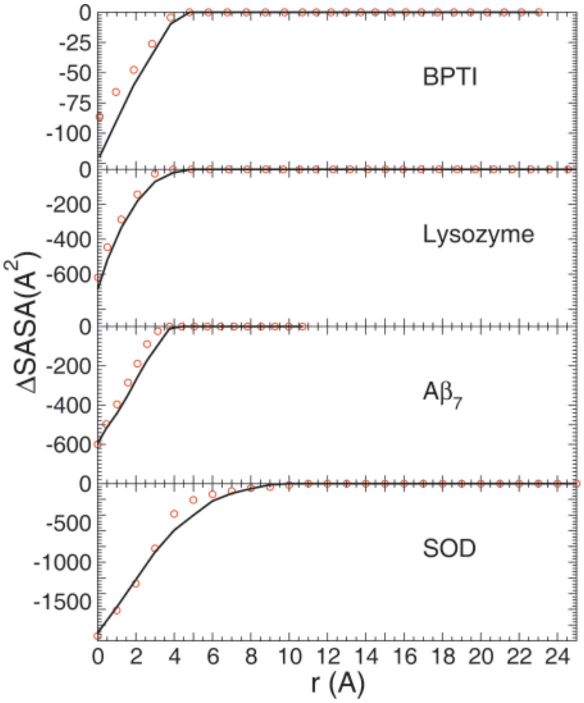

Figure 5. ΔSASAs as function of the intermonomer distance, obtained by taking the difference between the SASAs of dimers and the SASAs of the two isolated monomers. Results are reported for the AA (circles) and CG (continuous solid line) models.

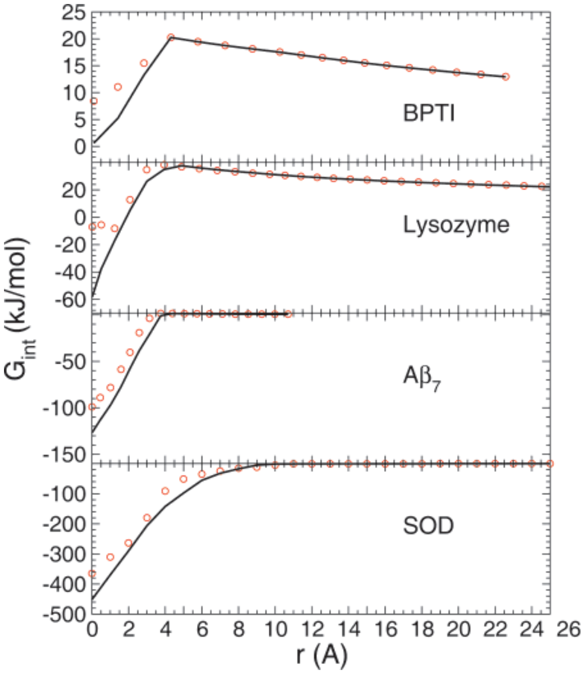

Figure 6. Total interaction free energy as a function of intermonomer distance. Circles are shown for the AA model, whereas the solid continuous lines are from the CG model.

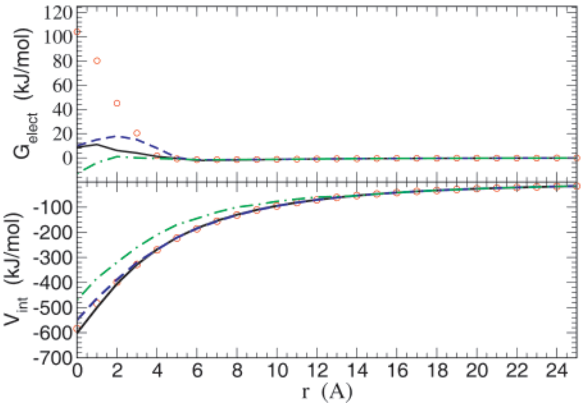

Figure 7. Bottom: Dependence of the bare electrostatic interaction energy on the intermonomer distance for the SOD dimer. The circles correspond to the AA system and the solid curve to the CG system composed of charges and optimized dipoles. The other two curves correspond to the two or three site Zacharias' model with natural charges (dot-dashed line) and optimized point charges (dotted line). Top: Same for the effective electrosatic interaction energy in solution.

partial charges. Every grain now carries a charge instead of a dipole, and all the charges are simultaneously optimized to best reproduce the AA electrostatic potential, as in eq 2, with the additional constraint of conserving the net charge of each residue (0 or ±1). This procedure improves upon the so-called *natural charge* representation of the original Zacharias' model, in which only charged residues carry one full charge on their extremal grain, all other grains staying neutral. We have carried out the same calculation described in Figures 3−6, comparing AA and CG dimer interaction energies as function of distance. Surprisingly, we find no significant improvement with respect to the 1−2-site model described before. For BPTI, lysozyme, and A$\beta_7$7, the results turn out more or less similar to those presented in Figures 3−6. The case of SOD is illustrated in Figure 7. For the vacuum electrostatics, it can be seen that the results are slightly deteriorated at short distances with respect to our initial one-dipole per residue representation. This is not surprising because a point-charge representation of the charge density is less flexible because it imposes the orientation of the residue dipoles. The electrostatic energy in solution, on the other hand, is slightly better, around 4−5 Å, as expected from the improvement of the cavity shape, but this improvement appears to no longer be true at shorter distances when interpenetration occurs. The natural charge representation of the original Zacharias' model does always give worse results, as can be expected from a less-accurate representation of the charge density.

The problem above clearly comes from the fact that the interaction between the two SOD monomers is highly specific and lateral chains from both partners become geometrically intricated at very short distances. A final possible refinement of our coarse-graining procedure is to stick to the AA representation at short residue−residue distances and switch to the CG model at only longer distances. To this end, given a certain threshold distance (for example 3 Å), we have computed the residue−residue interactions using the AA charge distribution and boundaries boundaries whenever the distance between any two atomic sites turns out below the threshold and the full CG representation otherwise. The results are presented in Figure 8 for a threshold at 3 and 6 Å. In the lowest panel, we have represented also the number of residues that are handled at

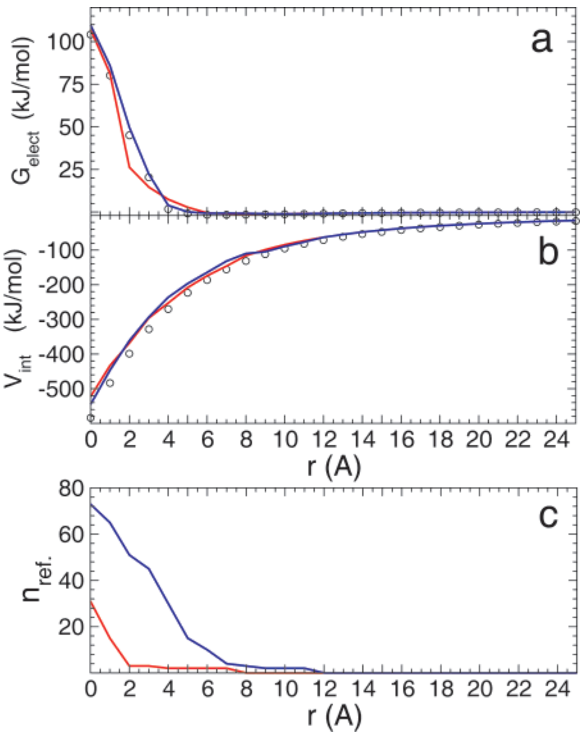

Figure 8. Electrostatic interaction energy in solution (a) and in vacuum (b) between the two monomers of the SOD dimer. Two short-ranged CG refiments are shown, with threshold at 3.0 Å (red curves) and at 6.0 Å (blue curves) and are compared to the AA results (black circles). In panel c, the number of refined residues, n_{ref}, as a function of distance are reported.

a AA level as a function of separation distance. It appears clearly that a threshold at 3 Å already gives a good description of the overall electrostatic interaction in solution, including desolvation, compared to the full AA calculation. The number of AA residues to consider remains relatively modest even at contact: about 30, that is, 15 per monomer (for a total number of 153). This number falls down to zero as soon as the intermonomer separation becomes greater than 2 Å.

C. SASA Term. For the four systems investigated, Figure 5 shows the variation of the differential solvent accessible surface area as a function of distance. As expected, when the minimum distance between the two interacting molecules is greater than a value $dist_{max}$, the difference between the sum of the SASA of two isolated molecules and the SASA of the interacting pair is zero. This $dist_{max}$ is the distance at which the solvent accessible surfaces of the two partners get in contact and the desolvation process starts. For lysozyme, BPTI, and Aβ_7 the ΔSASA term is vanishing for distances longer than ~4 Å. For SOD, the threshold distance reaches ~9 Å. From the visual analysis of the SOD solvent accessible surfaces, it is found that this early desolvation process is caused by the presence of specific residues, such as Lys25, Tyr26, and Ala112. These are placed at the border of the monomer separation surface and protrude out of one monomer toward the other, thus creating some steric hindrance between the two partners. Unexpectedly, the worst agreement between CG and AA is obtained for BPTI where no specific interaction between monomers is observed. Overall, the differential SASA are well reproduced by the coarse-grained model, which is certainly the result of our careful parametrization of the residue radii from their individual SASA. It is not yet clear why the CG model can provide such a good solvent-accessible surface but still somehow misses some of the fine electrostatic boundary effects.

D. Total Interaction Free Energy. From eq 3, we can obtain the interaction free energy profiles, corresponding to the potential of mean force (PMF) of the aggregate, by summing the electrostatic interaction free energy and the nonpolar term. The latter is obtained from the ΔSASA term and after multiplication with the γ proportionality constant introduced in eq 8. We have assigned to this proportionality factor a value of 0.24 kJ/mol Å^2, in agreement with previous studies of binding free energies from continuum dielectric methods.[32,35] It should be noted that a possible way to obtain the value of the γ parameter is to derive it from experimental results by fitting the calculated solvation free energies of small hydrophobic molecules to the experimental ones. This implies that the γ parameter takes different values if different sets of atomic radii and atomic charges are used. Furthermore, the γ parameter depends on the chosen solute and solvent dielectric constant ϵ_p and ϵ_s. The value of 0.24 kJ/mol for γ that we eventually chose was determined for the PARSE[34] parameter sets which reproduce the solvation free energies of proteins assuming $\epsilon_s = 2$. It should be stressed that the set of PARSE charges differ from the CHARMM27 set, used in this work for the AA models. For this reason, the sum of the electrostatic term G_{elect} and the nonpolar term $\Delta\Delta G_{np}$ in eq 3 might not give fully meaningful results compared to experiments, and this applies to both the AA and CG systems. Notwithstanding, we remind to the reader that the purpose of this first study is to explore the validity of CG models of biomolecules and *not* (yet) to reproduce, through an appropriate CG parametrization, experimental results. Thus, for our purposes the most adequate choice of γ is the simplest, that is, an identical γ for CG and AA.

In Figure 6, the profiles of the interaction free energy for the four biomolecules are shown. Remarkably, above ~4Å, the AA and CG free energy curves are in excellent agreement for Aβ_7, lysozyme, and BPTI. For SOD, deviations are noticeable only below 6 Å. For smaller molecule−molecule distances, the agreement is still quite good until 2−3 Å for all curves but degrades more strongly when the monomers are in closer contact.

This short distance discrepancy can be attributed to different contributions, the SASA term for BPTI, and desolvation effects in all other cases. It is noteworthy however that all the important physical features observed for the AA model are well reproduced by CG. BPTI and lysozyme are expected to have nonspecific interactions and to form very weak complexes if any. This is indeed what is observed, at least for the relative orientation for which the calculation were performed. At longer distances, the effective interaction potential is repulsive and dominated by like-charge repulsion; this part of the potential is perfectly reproduced by the CG description. The surface effects induce a rather shallow attraction at shorter distances, which is slightly overestimated by the CG model. For more specific interactions, Aβ_7, and SOD, the PMF shows a deep well at short distances and becomes completely flat afterward.

Bearing in mind the results in previous sections, we can conclude here that despute its not optimal parametrization, the CG model succeeds in reproducing the PMF of the AA systems. The agreement is semiquantitative when the two monomer are in contact and becomes quantitative at intermonomer distances larger than 4–6 Å.

IV. Conclusions

To efficiently characterize and model the interactions between biological macromolecules, it is necessary to consider a less-detailed description than that of the atomic scale. This can be done by modeling the system in term of elementary grains, coarse-grained description, such as the residues, the nucleotides, riboses, and even the higher molecular entities. Thus, this paper has developed a CG model of biomolecules that represents the amino acids in proteins and peptide chains as charged dipolar spheres. We have parametrized such a model based on the physical properties of individual amino acids and used it to study the interaction between solvated proteins and peptides. We have then computed the protein–protein potential of mean force for several selected systems and systematically compared results for the CG systems with those for the corresponding AA systems.

Despite the expected loss of atomic definition in the interactions, which also implies approximations in specific bonding such as hydrogen bonding, our CG model is capable of reproducing well the potential of mean force of the AA model until the intermonomer distance becomes too small. In particular, for conformations where specific interactions between monomers are unimportant, the CG interaction free energies are comparable to results from the AA model until the atom hard cores get into contact. This is the case of the lysozyme and BPTI homodimers. SOD and Aβ_7, instead, form stable aggregates in solution and presents specific electrostatic interactions. Hence, for the electrostatic interaction free energy, we obtain CG curves that deviate from those obtained for the AA models when the distance between the molecules are below the desolvation threshold. We also find that to get quantitative predictions at shorter distance a slight increase in the resolution of the model from 1–2 to 2–3 grains per residue is not sufficient. On the other hand, a mixed description where the interaction between grains at very short distances (<3 Å) is described at an atomic level is found to be quite accurate. In this latter approach, only a limited number of residues needs an atomistic representation, for example, 30 for the interaction between monomers in SOD, which will not degrade the computational efficiency of our coarse-grained approach too much.

To conclude, although the parametrization of our CG model is not optimal, the major result of this first investigation is that our CG model is very successful in reproducing the potential of mean force of its corresponding AA model. In a protein–protein interaction screening context,[12] our CG model will likely be useful to decide if two proteins can bind together or not, but it will not be sufficiently accurate to predict the association free energy. For our systems, this energy could deviate at full contact of 20% from the reference model.

We warn the reader that our protein–protein interaction picture is still missing some important contributions before a meaningful comparison with experimental data can be attempted. In particular, the steric repulsion effect should be considered via a (smooth) repulsive pairwise residue–residue potential. The induced dipole–induced dipole contribution emerging from both the electronic and atomic polarizability of the residues (the later being mainly due to the flexibility of the lateral chain) should also be considered. We point out that a set of atomic polarizabilities for all amino acids has been proposed recently by Song.[7] The underlying CG model, which involves one dipolar polarizable spherical site per residue, is the direct extension of the model explored in our study.

Finally, we point out that our CG model of proteins can be easily adapted coupled with Gaussian network models[36] to include a certain degree of interresidue flexibility.

Abbreviations: Coarse-grained (GC), all-atom (AA), molecular dynamics (MD), solvent accessible surface area (SASA), potential of mean force (PMF), Cu–Zn superoxide dismutase (SOD), basic pancreatic trypsin inhibitor (BPTI), Poisson–Boltzmann (PB).

Acknowledgment. The authors would like to thank Dr. L. Maragliano for providing the equilibrated structure of SOD homodimer. The authors also acknowledge support from the "Action Concertée Incitative IMPbio" project "Mésobio".

References

(1) Tozzini, V. Coarse-grained models for proteins. *Curr. Opin. Struct. Biol.* **2005**, *15*, 144.

(2) Zacharias, M. Protein–protein docking with a reduced protein model accounting for side-chain flexibility. *Protein Sci.* **2003**, *12*, 1271.

(3) Brown, S.; Fawzi, N. J.; Head-Gordon, T. Coarse-grained sequences for protein folding and design. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 10712.

(4) Nielsen, S. O.; Lopez, C. F.; Srinivas, G.; Klein, M. L. Coarse-grain models and the computer simulation of soft materials. *J. Phys.: Condens. Matter* **2004**, *16*, R481–R512.

(5) Baker, N. A. Improving implicit solvent simulations: A Poisson-centric view. *Cur. Opin. Struct. Biol.* **2005**, *15*, 137.

(6) Feig, M.; Brooks, C. L., III. Recent advances in the development and application of implicit solvent models in biomolecule simulations. *Curr. Opin. Struct. Biol.* **2004**, *14*, 217.

(7) Song, X. An inhomogeneous model of protein dielectric properties: Intrinsic polarizabilities of amino acids. *J. Chem. Phys.* **2002**, *116*, 9359.

(8) Song, X. The extent of anisotropic interactions between protein molecules in electrolyte solutions. *Mol. Simul.* **2003**, *29*, 643.

(9) Sheinerman, F. B.; Norel, R.; Honig, B. Electrostatic aspects of protein–protein interactions. *Curr. Opin. Struct. Biol.* **2000**, *10*, 153.

(10) Lo Conte, L.; Chothia, C.; Janin, J. The atomic structure of protein−protein recognition sites. *J. Mol. Biol.* **1999**, *285*, 2177.

(11) Simonson, T.; Archontis, G.; Karplus, M. Free-energy simulations come of age: Protein−ligand recognition. *Acc. Chem. Res.* **2002**, *35*, 430.

(12) Janin, J. Welcome to CAPRI: A critical assessment of predicted interactions. *Proteins* **2003**, *47*, 257.

(13) Mendez, R.; Leplae, R.; Lensink, M. F.; Wodak, S. J. Assessment of CAPRI predictions in rounds 3−5 shows progress in docking procedures. *Proteins* **2005**, *60*, 150.

(14) Lee, B.; Richards, F. M. The interpretation of protein structures: Estimation of static accessibility. *J. Mol. Biol.* **1971**, *55*, 379.

(15) Roux, B.; Simonson, T. Implicit solvent models. *Biophys. Chem.* **1999**, *78*, 1.

(16) Sharp, K. A.; Honig, B. Electrostatic interactions in macromolecules. *Annu. Rev. Biophys. Chem.* **1990**, *19*, 301.

(17) Gilson, M. K.; Honig, B. Calculation of the total electrostatic energy of a macromolecular system: solvation energies, binding energies, and conformational analysis. *Proteins Struct. Funct. Genet.* **1988**, *4*, 7.

(18) Honig, B.; Nicholls, A. Classical electrostatics in biology and chemistry. *Science* **1995**, *268*, 1144.

(19) Hermann, R. B. Theory of hydrophobic bonding II. The correlation of hydrocarbon. Solubility in water with solvent cavity surface area. *J. Phys. Chem.* **1971**, *76*, 2754.

(20) Chothia, C. The nature of the accessible and buried surfaces in proteins. *J. Mol. Biol.* **1971**, *105*, 1.

(21) Balbach, J. J.; Ishii, Y.; Antzutkin, O. N.; Leapman, R. D.; Rizzo, N. W.; Dyda, F.; Reed, J.; Tycko, R. Amyloid fibril formation by $A\beta_{16-22}$, a seven-residue fragment of the Alzheimer's β-Amyloid peptide, and structural characterization by solid state NMR. *Biochemistry* **2000**, *39*, 13748.

(22) Sergi, A.; Ciccotti, G.; Falconi, M.; Desideri, A.; Ferrario, M. Effective binding force calculation in a dimeric protein by molecular dynamics simulation. *J. Chem. Phys.* **2002**, *116*, 6329.

(23) Kundrot, C. E.; Richards, F. M. Crystal structure of hen egg-white lysozyme at a hydrostatic pressure of 1000 atmospheres. *J. Mol. Biol.* **1987**, *193*, 157.

(24) Parkin, S.; Rupp, B.; Hope, H. Structure of bovine pancreatic trypsin inhibitor at 125 K: Definition of carboxyl-terminal residues Gly57 and Ala58. *Acta Crystallogr. D* **1996**, *52*, 18.

(25) Procacci, P.; Darden, T. A.; Paci, E.; Marchi, M. ORAC: A molecular dynamics program to simulate complex molecular systems with realistic electrostatic interactions. *J. Comput. Chem.* **1997**, *18*, 157.

(26) MacKerell, S., Jr.; Bashfor, D.; Bellotan, M.; Dunbrack Jrand, R. L.; Evansecand, J. D.; Fieland, M. J.; Fischeand, S.; Gaand, J.; Guand, H.; Hand, S.; Joseph-McCarthand, D.; Kuchniand, L.; Kuczerand, K.; Laand, F. T. K.; Mattoand, C.; Michnicand, S.; Ngo, T.; Nguyeand, D. T.; Prodhoand, B.; Reiher IIand, W. E.; Rouand, B.; Schlenkricand, M.; Smit, J. C.; Stotand, R.; Strauand, J.; Watanaband, M.; Wiorkiewicz-Kuczerand, J.; Yin, D.; Karplux, M. All-atom empirical potential for molecular modelling in dynamics studies of proteins. *J. Phys. Chem.* **1998**, *102*, 3586.

(27) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926.

(28) Baker, N. A.; Sept, D.; Joseph, S.; Holst, M. J.; McCammon, J. A. Electrostatics of nanosystems: Application to microtubules and the ribosome. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 10037.

(29) Schutz, C. N.; Warshel, A. What are the dielectric "constants" of proteins and how to validate electrostatic models? *Proteins: Struct. Funct. Genet.* **2001**, *44*, 400.

(30) Dong, F.; Vijayakumar, M.; Zhou, H.-X. Comparison of calculation and experiment implicates significant electrostatic contributions to the binding stability of Barnase and Barstar. *Biophys. J.* **2003**, *85*, 49.

(31) Lu, B. Z.; Chen, W. Z.; Wang, C. X.; Xu, X. J. Protein molecular dynamics with electrostatic force entirely determined by a single Poisson-Boltzmann calculation. *Proteins* **2002**, *48*, 497.

(32) Froloff, N.; Windemuth, A.; Honig, B. On the calculation of binding free energies using continuum methods: Application to MHC class I protein−peptide interactions. *Protein Sci.* **1997**, *6*, 1293.

(33) Nicholls, A.; Sharp, K. A.; Honig, B. Protein folding and association: Insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins: Struct. Funct. Genet.* **1991**, *11*, 281.

(34) Sitkoff, D.; Sharp, K. A.; Honig, B. Accurate calculation of hydration free energies using macroscopic solvent models. *J. Phys. Chem*. **1994**, *98*, 1978.

(35) Trylska, J.; McCammon, J. A.; Brooks, C. L., III. Exploring assembly energetics of the 30S ribosomal subunit using an implicit solvent approach. *J. Am. Chem. Soc.* **2005**, *127*, 11125.

(36) Bahar, I.; Atilgan, A. R.; Erman, B. Direct evaluation of thermal fluctuations in proteins using a single parameter harmonic potential. *Folding Des.* **1997**, *2*, 173.

CT700121N