Implicit modeling of nonpolar solvation for simulating protein folding and conformational transitions

Jianhan Chen† and Charles L. Brooks III*

Received 13th September 2007, Accepted 5th November 2007
First published as an Advance Article on the web 14th November 2007
DOI: 10.1039/b714141f

Accurate description of the solvent environment is critical in computer simulations of protein structure and dynamics. An implicit treatment of solvent aims to capture the mean influence of water molecules on the solute *via* direct estimation of the solvation free energy. It has emerged as a powerful alternative to explicit solvent, and provides a favorable compromise between computational cost and level of detail. We review the current theory and techniques for implicit modeling of nonpolar solvation in the context of simulating protein folding and conformational transitions, and discuss the main directions for further development. It is demonstrated that the current surface area based nonpolar models have severe limitations, including insufficient description of the conformational dependence of solvation, over-estimation of the strength of pair-wise nonpolar interactions, and incorrect prediction of anti-cooperativity for three-body hydrophobic associations. We argue that, to improve beyond current level of accuracy of implicit solvent models, two important aspects of nonpolar solvation need to be incorporated, namely, the length-scale dependence of hydrophobic association and solvent screening of solute–solute dispersion interactions. We recognize that substantial challenges exist in constructing a sufficiently balanced, yet reasonably efficient, implicit solvent protein force field. Nonetheless, most of the fundamental problems are understood, and exciting progress has been made over the last few years. We believe that continual work along the frontiers outlined will greatly improve one's ability to study protein folding and large conformational transitions at atomistic detail.

1. Introduction

Recent years have witnessed remarkable progress in computational methodologies for modeling and prediction of protein structures and dynamics.[1–3] Central to these developments are the energy functions that describe the basic building blocks of the molecule and their interactions and simulation techniques. Such energy functions include knowledge-based ones that largely rely on statistics derived from known structures and sequence analysis.[4–6] These energy functions are often complemented by physics motivated terms, and have proven to be quite powerful in the challenging exercise of protein structure prediction.[2,7,8] For rigorous simulation of biomolecules (to explore the dynamics and understand biological function), the workhorse is the general purpose molecular mechanics force fields.[9,10] These empirical force fields are based on basic physical principles (*i.e.*, physics-based), and the associated

Department of Molecular Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, CA 92037, USA. E-mail: brooks@scripps.edu; Fax: (858) 784 8688; Tel: (858) 784 8035
† Current address: Department of Biochemistry, Kansas State University, Manhattan, KS 66506, USA.

Jianhan Chen received his PhD in Chemical and Materials Physics from the University of California at Irvine under the joint supervision of Professors A. J. Shaka and Vladimir A. Mandelshtam in 2002. After his postdoctoral training with Professor Charles L. Brooks, III at The Scripps Research Institute, he joined the Biochemistry Faculty of Kansas State University in 2007. His current research interests include developing accurate implicit solvent models and efficient sampling methods for studying biomolecular structure and dynamics.

Charles L. Brooks, III received his PhD from Purdue University with Professor Stephen A. Adelman in 1982, and was an NIH Postdoctoral Fellow with Professor Martin Karplus at Harward University between 1982 and 1985. He then joined the Chemistry Faculty of Carnegie Mellon University, and was promoted to Professor in 1992. Since 1994, Professor Brooks has been at The Scripps Research Institute. His research is focused on the application of statistical mechanics, quantum chemistry and computational methods to chemically and physically oriented problems in biology.

parameters are derived from a combination of experimental measurements, quantum mechanics calculations and empirical corrections. The accuracy of physics-based force fields has been consistently improved over the years.[10] Nevertheless, various limitations persist, especially when it comes to exploring large conformational transitions of proteins.[1,3,10,11] Many of these limitations are intimately related to the need for an accurate and efficient description of solvent that plays critical roles in the structure, dynamics and function of biomolecules.[12] Explicit inclusion of water molecules provides the most detailed description of solvent, but, at the same time, dramatically increases the system size and the associated computational cost. Furthermore, longer simulations are necessary to obtain statistically meaningful averages for protein structure, dynamics and thermodynamic properties. One should also remember that a higher level of detail does not automatically translate into higher level of accuracy, particularly in terms of describing protein conformational equilibria. For example, while it is recognized that there are systematic biases in the secondary structure preference in modern protein force fields with explicit solvent models,[9,13–15] progress towards general and transferable corrections has been very limited. This can be partially attributed to the expensive computational cost of explicit solvent simulations, besides the highly complex nature and large parameter space of the related optimization problem.

Implicit solvent models have emerged as a powerful alternative to explicit water for representing the solvent environment, where the mean influence of solvent molecules on the solute is captured *via* a direct estimation of the solvation free energy, the reversible work required to transfer the solute in a fixed configuration from vacuum to solution.[16,17] The total solvation free energy is often decomposed into nonpolar and electrostatic contributions, which correspond to the reversible work required to: first, insert the solute in the solvent with zero atomic partial charges; and second, switch the partial charges from zero to their full values.[17] Such a decomposition allows both components to be related to appropriate continuum descriptions of water, and is generally more accurate than fully empirical approaches where the total solvation energetics is estimated directly from either exposed surface area or solvent-excluded volume.[18–20] Continuum electrostatics is the most well-established model for electrostatic solvation, and the most popular nonpolar solvation model is based on solvent-accessible surface area (SA). In particular, the generalized Born (GB) approximation, complemented by an auxiliary SA-based nonpolar term, offers an excellent balance between efficiency and accuracy.[21] With continual enhancement and improvement over the last few years, the GB/SA approximation is now recognized as a prime choice for implicit treatment of solvent for biomolecular simulation.[9,16] Various implementations are now available in all major molecular modeling software packages, and have found applications in a wide range of biochemical and biophysical problems.[16]

Elimination of the solvent molecules by the implicit treatment substantially reduces the number of atoms needed to be simulated. More importantly, this can be achieved with only a moderate increase in the computational cost required for estimating the solvation free energy on-the-fly. Clearly, such a dramatic reduction in the system size does not come without a loss of detail and achievable accuracy. For example, implicit solvent models may yield considerable disagreement with explicit water simulations in short-range effects when the detailed interplay of a few water molecules (which are distinct from the bulk water) is important.[22,23] Examination of the potentials of mean force (PMFs) between model compounds also revealed a lack of fine structure in the implicit solvent PMFs due to continuum descriptions.[24,25] Nonetheless, the substantial reduction in the computational cost and extension of accessible simulation timescales with implicit treatment of solvent have opened a door to address many biological problems that are otherwise difficult with explicit solvent.[14,16] Particularly, it has also allowed development of new modeling techniques, such as constant pH simulations to study pH-dependent protein folding and unfolding.[26,27]

Another important, but easily overlooked, benefit of implicit solvent is that a reduction in the computational cost also facilitates careful re-parameterization of the force field, for example, to suppress systematic biases such as the above mentioned secondary structure preferences. The conformational equilibria of proteins is governed by delicate balance among sets of underlying competing interactions, *i.e.*, the solvation preference of side chains and backbones in solution *versus* the strength of solvent-mediated interactions between these moieties in a complex protein environment.[14,25] The extent to which a solvent model (explicit or implicit) can capture this delicate balance is a key to its success in describing conformational equilibria. Achieving sufficient balance of the competing interactions for complex heterogeneous systems is a challenging task. To a large extent, this is due to a severe lack of direct experimental measurements or reliable high-level quantum mechanics data. In practice, one has to resort to indirect experimental observables, such as thermodynamic stability and conformation equilibria of model peptides and proteins, in order to rebalance the force field.[10,14,15,25] However, reliable calculation of these thermodynamic quantities requires extensive folding and unfolding simulations, and is generally only accessible with implicit solvent. For example, it was recently demonstrated that a more consistent GB/SA force field could be achieved by carefully balancing solvation and intramolecular interaction, guided by PMFs between amino acid polar groups, and by conformational equilibria of model peptides.[25,28] The optimized force field was subsequently verified by folding of additional peptides and miniproteins that were not used in the optimization process,[25,26] indicating much improved robustness and transferability. Similar efforts have also been reported for other GB/SA models.[15,29] Note that reliable calculation of peptide conformational equilibria remains a challenging and expensive task even with implicit solvent, and that advance sampling techniques such as the replica exchange (REX) method[30,31] are indispensable in these developments.

Despite methodological advances and force field parameterization improvements, applications of various physics-based implicit solvent force fields to *ab initio* folding simulation (*i.e.*, without *any* additional knowledge besides the force field and sequence) of larger and more complex proteins (those with three or more secondary structure elements and non-trivial

tertiary folds) have demonstrated only limited success.[1,3,11,32] While individual successful folding simulations have been reported for a few mini-proteins and small natural proteins,[33–38] they appear to rely heavily on serendipitous cancellation of errors for the given combination of force field choice and protein sequence, and consistency in a force field's ability to fold a range of proteins with different topologies has yet to be demonstrated. Many factors might contribute to this, including limitations of the underlying protein force fields (*e.g.*, atomic charges and other nonbonded parameters) and insufficient sampling capability. It should also be emphasized that direct simulation of protein folding is one of the most challenging and stringent tests of the force field (and the sampling method). From the implicit solvent perspective, there exist additional intrinsic limitations due to the continuum approximations. Nevertheless, we believe that the full potential of implicit solvent force fields has yet to be reached, and there is plenty of room for further improvement. Specifically, many of the recent developments have been focused on electrostatic solvation, through both methodological improvements[16,39] and optimization of key (physical) parameters.[15,25,40,41] As a result, the electrostatic solvation free energy is now the most reliable and accurate aspect in most GB related implicit solvent models. In contrast, the nonpolar solvation free energy has been either largely ignored or described by simplistic SA models.[16,39] Indeed, nonpolar solvation is more complex in nature, and the associated energetics is generally of smaller magnitude than the polar counterpart. Nonetheless, it is well appreciated that hydrophobic association is one of the two principal interactions (besides hydrogen bonding) that determines biomolecular structures and assemblies.[42] The delicate balance between intramolecular dispersion interactions and nonpolar solvation is essential for the accurate description of protein conformational equilibria. Therefore, if one wishes to improve beyond the current level of accuracy in the implicit solvent force field such that it might be consistently applied to simulate protein folding and conformational transitions, further improvement in the treatment of nonpolar solvation is essential.

The main focus of this article is to investigate various drawbacks of popular SA models for treating nonpolar solvation, mainly in the context of modeling protein folding and conformational transitions. By examining interactions between model compounds in explicit and implicit solvent, and by simulating folding and unfolding of small peptides, we identify important physics that ought to be incorporated for more accurate and realistic modeling of nonpolar solvation. We conclude with a discussion of several promising solutions as well as remaining challenges in such an endeavor.

2. Surface area based nonpolar solvation models

2.1 First solvation shell approximation

The underlying statistical thermodynamics basis of implicit solvent and the formal decomposition of the total solvation free energy into nonpolar and electrostatic contributions have been described in detail previously.[17] Nonpolar solvation is mainly associated with short-range repulsive interactions (to create the solvent cavity where the solute resides) and solute–solvent dispersion interactions. Both interactions are dominated by the first solvation shell. Additional first-solvation-shell effects include entropic contributions due to changes in water structures near the solute.[43] The energetics associated with these effects should, in a first-order approximation, be proportional to the average number of water molecules in the first solvation shell. This is the physical basis of SA-based nonpolar solvation models, where the nonpolar solvation free energy, ΔG_{np}, is estimated as

$$\Delta G_{\mathrm{np}} = \sum_i \gamma_i A_i, \qquad (1)$$

with atomic effective surface tension coefficients, γ_i, and atomic solvent-accessible surface areas, A_i. In most GB/SA models, further simplification is made by assuming a universal γ value for all atom types, and eqn (1) is reduced to $\Delta G_{\mathrm{np}} = \gamma A$, where A is the total solvent-accessible surface area. Validity of the linear approximation of eqn (1) has been supported by examining the experimental solvation free energies of linear alkanes and other neutral organic compounds as a function of surface area,[18,44–46] as well as theoretical and computational studies on nonpolar solvation and hydrophobic interactions.[47–50]

Continuum descriptions of solvent break down at short range, due to a wide range of effects including nonlinear response of solvent to the local electric field near the solute and charge transfer to or from the solvent.[43] These "secondary" effects are also dominated by the first solvation shell. The associated energetics should also largely scale with the solute surface area, and can be in principle included in carefully parameterized SA terms. As such, the SA term often reflects more than just nonpolar solvation, and the parameter γ is largely empirical in nature, determined mainly based on the *total* solvation free energy of (neutral) model compounds. Depending on a range of factors including the underlying (solute) force field, accompanying electrostatic solvation model, and choice of the solute–solvent boundary (see below), the value of γ varies substantially, ranging from as low as 5–7 cal mol^{-1} Å^{-2} [21,45,51] to 40–70 cal mol^{-1} Å^{-2}.[52,53] Inconsistency in γ to some extent reflects the crudeness of SA models. For protein simulations, small γ has been empirically found to be optimal, with the most commonly used value being 5 cal mol^{-1} Å^{-2}.[25,54,55] With such a small γ, the SA term has minimal impact on distinguishing compact misfolded conformations from native-like folds. As it will be discussed in detail in the following sections, small γ (improperly) compensates for some of the artifacts of SA models.[32]

2.2 Solute–solvent boundary

The precise location of the solute–solvent boundary is a key physical property that governs both nonpolar and electrostatic solvation free energies. In principle, the solvent-accessible surface, defined as a continuous surface traced out by the center of a ball rolling over the solute,[56] is the most appropriate choice. Alternatively, van der Waals (vdW)-like and molecular surfaces are also commonly used.[57] As discussed above, the continuum approximations break down at short range, and the SA term contains empirical corrections to compensate for various first-solvation-shell effects and to

cancel errors in other force field terms. Therefore, the optimal choice of solute–solvent boundary in practice is no longer obvious. Furthermore, the optimal solute–solvent boundary for nonpolar solvation does not necessarily coincide with that for electrostatic solvation, even though the same boundary is used in all current GB/SA models. For a given surface definition, the exact location of the boundary is determined by the intrinsic atomic radii used to construct the surface (referred to as input radii hereafter). Input radii are key physical parameters in optimization of both electrostatic and nonpolar solvation. Many of the previously observed artifacts of continuum electrostatics models, such as over-stabilized salt-bridge formation, can be effectively suppressed through careful parameterization of input radii.[25,29] Similar optimization is expected to be necessary for improving the accuracy of the nonpolar solvation free energy. An important limitation, however, is that these optimized parameters are tightly coupled with the specific combination of protein force field and implicit solvent model, and are generally not transferable.

2.3 SA models systematically over-stabilize pair-wise nonpolar interactions

The linear approximation of eqn (1) is simple, can be implemented efficiently, and works reasonably well with careful parameterization. Such simplification also leads to certain caveats, many of which have been well recognized and extensively discussed in the literature, particularly concerning the difficulty in handling cyclic alkanes,[45] the limited ability to account for conformational dependence of hydration free energy,[49,58] and inconsistency in γ values.[53] These limitations have been mainly attributed to either an insufficient description of solvent screening of (medium-range) intramolecular dispersion interactions[58–60] or the dependence of γ on surface curvature and molecular shape.[45,53,61] For small molecules, reliable experimental data exist for the solvation free energy, and most of the potential limitations might be overcome with extensive parameterization and conservative application within the valid regimes.[43] The problem is much more complex for biomolecules because of the necessity to handle heterogeneous environments, complex molecular shapes, and different molecular length scales. The exact consequences of the above limitations have not been fully appreciated in terms of modeling peptide and protein conformational equilibria. While it is clear that critical limitations exist, as reflected in a lack of consistency in *ab initio* protein folding simulations discussed above, many factors contribute simultaneously and in highly convoluted ways, rendering it extremely difficult to pinpoint the exact origins of the problems.

Parameterization of a complex protein force field is a severely under-determined problem,[9,10] and the ability to reproduce the experimental solvation free energy of a few dozen of model compounds is often insufficient to enforce accurate balance of competing interactions that are critical for modeling conformational equilibria. Therefore, it is useful to directly examine the ability of an implicit solvent model to reproduce the strength of interactions between representative nonpolar side chain groups in arguably more accurate explicit solvent. Similar approaches prove to be effective for optimizing the electrostatic component.[24,25] In Fig. 1, we compare the strengths of interaction between several neutral amino acid side chains, including Phe, Ile, Pro, Tyr and Trp, in representative fixed configurations in three solvent models. The stability shown is simply defined as the free energy at the contact minimum with respect to that at large separation. The solute was described by the CHARMM22 all-atom force field.[62] The PMFs in TIP3P[63] explicit water were computed using the same free energy perturbation protocol described previously.[32] The convergence is on the order of 0.1 kcal mol^{-1}, estimated by comparing the stabilities using only the first and second half of sampling. PMFs in implicit solvent were computed by simply translating the solutes along the reaction coordinates. GBSW/SA[25,28] and GBMV/SA[64] are two of the latest GB models that have been shown to be particularly accurate.[57] GBSW employs a vdW-based surface with a smooth dielectric boundary, while a molecular surface is used in GBMV. In both GB

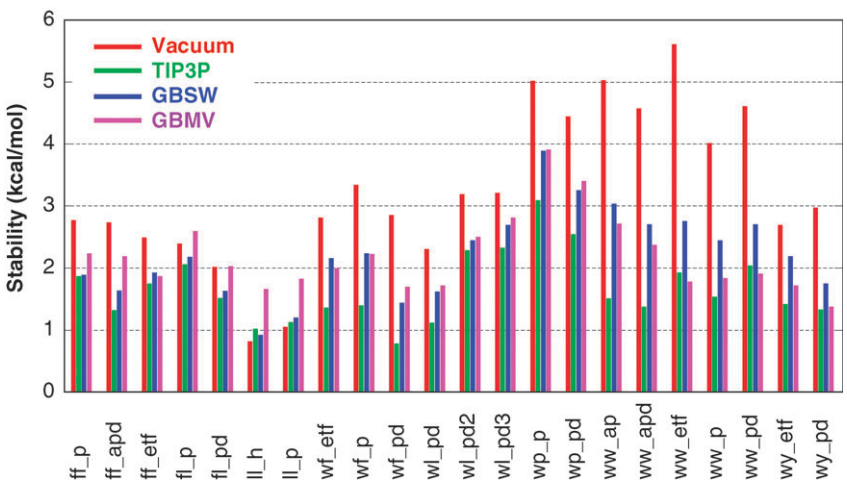

Fig. 1 Stabilities of pair-wise interactions between nonpolar amino acid side chain analogues in implicit and explicit solvents. The notations are: f (Phe), l (Leu), w (Trp), p (Pro), y (Tyr), p (parallel), pd (parallel displaced), ap (anti-parallel), apd (anti-parallel displaced), etf (edge-to-face). The same set of input radii[25] was used for both GBSW/SA and GBMV/SA. RMS deviations from the TIP3P results are 1.88, 0.68, 0.64 kcal mol^{-1} for vacuum, GBSW/SA and GBMV/SA, respectively; and the corresponding mean deviations are +1.55, +0.54, +0.53 kcal mol^{-1}, respectively.

models, Born radii are calculated by a rapid volume integration scheme that includes a higher-order correction term to the Coulomb field approximation.[64] Default GBSW/SA and GBMV/SA parameters were used, with the same set of input radii.[25] The nonpolar solvation energy was estimated with γ = 5.0 cal mol^{-1} Å^{-2}. The dimer stabilities in vacuum are also included, to illustrate the influence of solvent on the dimer stabilities. The results clearly show that both GBSW/SA and GBMV/SA systematically over-stabilize the dimer interactions in comparison to the TIP3P explicit water, with an average over-stabilization of 0.54 and 0.53 kcal mol^{-1}, respectively. The stabilities of particular pairs might be strongly influenced by the choice of vdW-like or molecular surfaces. When all pairs are considered, the difference is small and there is little indication of which surface is superior. Note that interactions with aromatic side chains, particularly with Trp, include substantial electrostatic contributions. While it is not obvious that the systematic over-stabilization can be largely attributed to limitations in the SA term, it will be illustrated further in the next section that there are several physical considerations that strongly link the over-stabilization to certain drawbacks of SA models. These results also provide a rationale for why small γ values have been empirically found to be optimal for protein modeling:[25,54,55] small γ alleviates the systematic over-stabilization. It is possible to improve the agreement with the TIP3P results by using atom or functional group specific γ values. However, such improvement is often not transferable to more complex interactions, such as those involving three or more nonpolar groups.[32]

In Fig. 2, we further examine the ability of the TIP3P, GBSW/SA and GBMV/SA solvent models to reproduce the experimental solvation free energy of all neutral amino acid side chain analogues except Ala. The root-mean-square (RMS) deviations from the experimental values are 1.41, 1.11 and 1.01 kcal mol^{-1} for TIP3P, GBSW/SA and GBMV/SA, respectively. The apparent better performance of implicit solvent is likely a result of more direct parameterization (*e.g.*, of the input atomic radii). Two important observations are made. Firstly, the TIP3P water model systematically under-solvates all the nonpolar side chains. This implies that interactions between these nonpolar side chains in TIP3P is systematically overestimated. Therefore, systematic over-estimation of dimer stabilities in implicit solvent (in comparison to the actual values) is likely even more severe than what the comparison with TIP3P results suggests. Secondly, the aromatic side chains (Trp, Phe and Tyr) are particularly problematic in current protein force fields (which is not limited to CHARMM22).[65] This is related to the difficulty of atom-centered fixed charge models in representing the out of plane π-electron densities.[66,67] Considering the importance of aromatic side chains in protein structure, this poses additional challenges in constructing a balanced (implicit solvent) force field.

2.4 Folding simulations of model peptides and proteins

Systematic over-stabilization of nonpolar interactions has important implications in simulation of protein folding and conformational transitions. It increases the roughness of the underlying potential energy surface and hinders rapid sampling of the conformational space. This is likely why spontaneous reversible folding and unfolding of weakly stable β-hairpins have proven difficult to simulate, even though reasonable agreement with experiments on important folding thermodynamics has been demonstrated for a range of peptides.[25] In particular, strong interactions between nonpolar residues and between large nonpolar side chains (*e.g.*, Phe and Trp) and the rest of the peptide chain (*e.g.*, backbone) render it difficult for the protein to escape from compact (misfolded or folded) conformational states. More importantly, the over-estimation of nonpolar interactions disrupts the delicate balance of underlying competing interactions and can shift the global energy minimum away from the true native basin.

For illustration, Fig. 3 summarizes the results of extensive *ab initio* folding simulations of two small helical proteins, including villin headpiece subdomain (residues 41–76; PDB ID: 1vii) (HP36),[68] and a 46-residue segment of

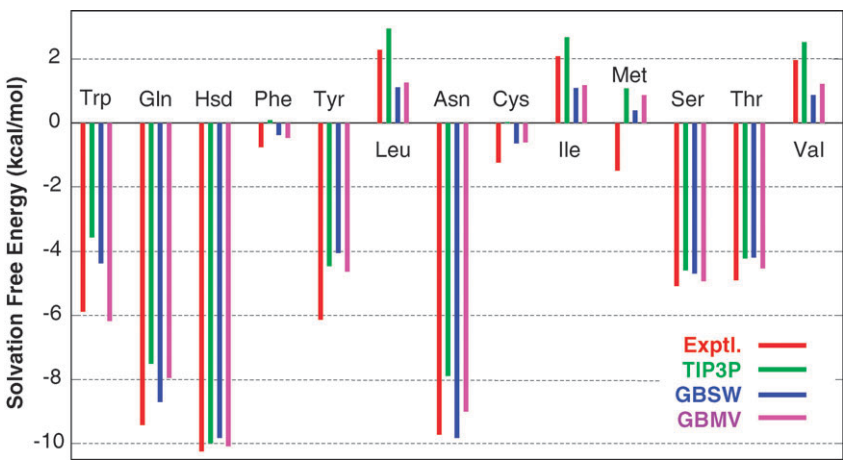

Fig. 2 Experimental and calculated total solvation energies of amino acid side chain analogues. The experimental and TIP3P results were taken from Shirts and Pande.[65] The implicit solvent results were computed using the same parameters as in Fig. 1, and all model compounds have the default geometries defined in the CHARMM22 fore field.[62] The RMS deviations from the experimental results are 1.41, 1.11 and 1.01 kcal mol^{-1} for TIP3P, GBSW/SA and GBMV/SA, respectively.

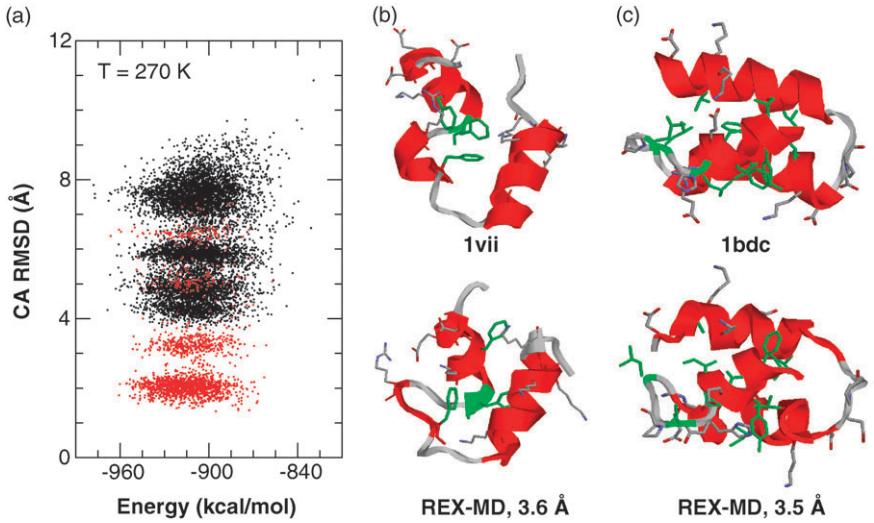

Fig. 3 (a) Total potential energy *vs.* CA RMSD plot from REX-MD folding (black dots) and control (red dots) simulations of villin headpiece subdomain (HP36). The folding simulation was initiated from a fully extended conformation, and the control simulation from the PDB structure. Snapshots were taken every 10 REX exchange steps at 270 K. The total simulation lengths were 160 and 20 ns for folding and control runs, respectively. (b) and (c) Representative near-native conformations from folding simulations in comparison with the experimental structures, (b) HP36, (c) protein A. The length of protein A REX-MD folding simulation is 100 ns. Representative conformations from the largest cluster at the lowest temperature (270 K) are shown, and their occupancies are 25 and 24%, respectively, during the last 20 ns of the REX-MD simulations. The helical segments (as defined in the PDB structures) are colored red, residues in the hydrophobic cores in green, and charged residues in CPK colors.

staphylococcal protein A fragment B (residues 10–55; PDB ID: 1bdc).[69] The CHARMM22/CMAP force field[62,70,71] with an optimized GBSW/SA solvent[25] was used. The REX molecular dynamics (MD) simulations were carried using the MMTSB Tool Set[31,72] to enhance the conformational sampling. 24 replicas spanning 270 to 600 K were used. Exchange of simulation temperatures was attempted every 2.0 ps. The simulations were initiated from fully extended conformations, and the total simulation lengths were 160 ns for villin headpiece (80 000 REX cycles) and 100 ns for protein A (50 000 REX cycles). For both proteins, near-native conformations, shown in Fig. 3b and c, were reached within the first 40 ns of REX-MD simulations, but little further progress toward the fully folded structures was made for the rest of the simulations. The near-native conformations are within 4 Å CA RMSD from the experimental structures. The secondary elements are largely correct, with helical content of about 60% of that for the native structures. The arrangements of helices are essentially native-like, but the packing is not as compact as in native states. The critical bottleneck to the fully folded conformations appears to be formation of the hydrophobic cores, highlighted in green color in Fig. 3b and c. For example, HP36 contains a mini hydrophobic core that consists of three phenylalanines, which is not formed in the simulated structure even at the end of the 160 ns REX-MD simulation. It is interesting that folding of secondary structures and formation of tertiary hydrophobic cores are clearly coupled, which is expected as both proteins are known to fold cooperatively.

The failure to reach fully folded states might be attributed to two reasons. First, the compact, near-native conformations are too stable in the current force field. To reach the true native fold requires searching through a large number of such compact conformations by breaking and reforming many (nonpolar) contacts. With systematic over-stabilization of nonpolar interactions, this becomes very slow and cannot be accomplished within the timescales simulated (even though these simulations are two of the longest REX-MD simulations reported up to date). Second, the underlying free energy surface is significantly distorted and the true native structure no longer corresponds to the global free energy minimum. To examine whether the native structure has lower potential energy, a 20 ns control REX-MD simulation was initiated from the PDB structure of HP36. The results were summarized in Fig. 3a. It shows that the fully folded conformations (*i.e.*, those with CA RMSD of about 2 Å) have lower energies on average. It indicates that the true global free energy minimum for HP36 might not have been severely distorted. Similar control simulations show that the native structures no longer have lower potential energies on average compared to other compact structures. Therefore, both reasons are responsible for the observed inability to fully fold HP36 and protein A. The fact that near-native conformations are reached in both cases is encouraging. It indicates that the force field has nearly proper balance of the underlying interactions, particularly with respect to electrostatic solvation and intramolecular interactions, and that a fine tuning of the nonpolar solvation model might be sufficient to fully fold both proteins. Recently, Duan and coworkers reported successful folding simulations of 35-resiude villin headpiece subdomain (HP35) using AMBER FF03 force field with a GB/SA model, reaching fully folded conformations with CA RMSD as low as 0.39 Å.[38,73] Again, there is little indication that this is a general case, and the success clearly hinges on the particular choice of force field (for a given sequence) as well as additional parameters such as cutoffs and γ.

3. Beyond SA models: important properties of nonpolar solvation that need to be explicitly modeled

Many of the solvation phenomena that give rise to the complex conformational dependence of nonpolar solvation are reasonably well understood, at least at a qualitative level.[17,49,53,58] Nonetheless, their importance in accurate modeling of protein conformational equilibria has not been fully appreciated. In this section, we discuss two physical properties that are believed to be critical for accurate modeling of nonpolar solvation.

3.1 Length-scale dependence of hydrophobic solvation

It has been recognized that there is a length-scale dependence of the free energy cost of solvating hydrophobic solutes.[42,74–76] Qualitatively speaking, small nonpolar solutes do not interrupt the hydrogen bonding network of water. The associated solvation free energy is largely entropic and depends on the solute volume. In contrast, large solutes induce the formation of an interface where the water molecules are involved in fewer hydrogen bonds on average. Therefore, the solvation free energy is proportional to the surface area. The crossover from small to large length scale regimes occurs roughly around 10 Å.[76,77] The effective sizes of exposed nonpolar amino acid side chains range approximately from about 3 Å (for Ala) to about 5 Å (for Trp), within the small length scale regime, whereas folded proteins typically fall in the large length scale regime. Therefore, such a length scale dependence of hydrophobic solvation is highly relevant in determining the equilibria of disordered, partially folded and folded protein conformations.

In the current SA models the length scale dependence is neglected and the parameter γ is conformationally independent. Such a simplification has been shown to result in two artifacts, over-stabilization of pair-wise interactions and failure to predict cooperativity in three-body hydrophobic associations.[32] These can be demonstrated through the following simple arithmetic. As the solvation free energy grows linearly with volume for small solutes,[42] effective surface tension coefficient of closely packed nonpolar n-mers ($n = 1, 2, 3$), defined as $\gamma^{(n)} = \Delta G_{\rm np}^{(n)}/A^{(n)}$, scales linearly with the effective size, *i.e.*, $\gamma^{(n)} \sim \gamma^{(1)} n^{1/3}$. The stabilities of dimers and three-body cooperativity contribution are defined as

$$W^{(2)} = \gamma^{(2)}A^{(2)} - 2\gamma^{(1)}A^{(1)} = \gamma^{(2)}\Delta A^{(2)} + 2\Delta\gamma^{(2)}A^{(1)}, \quad (2)$$

$$\begin{aligned}\delta F^{(3)} &= W^{(3)} - 3W^{(2)} \\ &= \gamma^{(3)}\Delta A^{(3)} - 3\gamma^{(2)}\Delta A^{(2)} + 3(\Delta\gamma^{(3)} - 2\Delta\gamma^{(2)})A^{(1)}, \quad (3)\end{aligned}$$

where $\Delta\gamma^{(n)} = \gamma^{(n)} - \gamma^{(1)}$ and $\Delta A^{(n)} = A^{(n)} - nA^{(1)}$. The second term in eqn (2) is positive considering $\Delta\gamma^{(2)} > 0$, but vanishes if γ is assumed to be independent of the solute size. In other words, the traditional SA models predict more negative $W^{(2)}$, *i.e.*, over-estimate the strength of pair-wise interactions. Such over-stabilization can be (improperly) compensated by choosing small values for parameter γ, and such choices were indeed found to give better results in simulating peptide folding.[25,54,55] Given that $\gamma^{(n)} \sim \gamma^{(1)} n^{1/3}$, eqn (3) can be reduced to $\delta F^{(3)} \sim 3(\Delta\gamma^{(3)} - 2\Delta\gamma^{(2)})A^{(1)} = -0.23\gamma^{(1)}A^{(1)}$, which is negative (*i.e.*, cooperative). Considering that the average nonpolar solvation free energy is on the order of 2–3 kcal mol^{-1} for amino acid side chains,[65] one can further estimate the three-body contributions to be about 0.5 kcal mol^{-1} per trimer. On the contrary, assuming a size-independent $\gamma = \gamma_0$ reduces eqn (3) to $\delta F^{(3)} = \gamma_0(\Delta A^{(3)} - 3\Delta A^{(2)})$, which is positive (*i.e.*, anti-cooperative) based on the simple geometric consideration that $\Delta A^{(3)} - 3\Delta A^{(2)} > 0$. These observations were confirmed by examining the PMFs of pair-wise and three-body interactions of Leu and Phe side chain analogues in various configurations in TIP3P and GBSW/SA solvents.[32] Other studies have also identified limitations of SA models in describing cooperativity and/or anti-cooperativity of hydrophobic associations.[78,79]

The implications of the above analysis in modeling protein folding and conformational equilibria is substantial. On one hand, a constant γ needs to be sufficiently large to compensate for lack of cooperativity in the model in order to maintain the stability of compact folded structures; on the other hand, small values of γ are desirable not to under-estimate the probability of extended conformations with exposed nonpolar side chains or not to over-stabilize loosely packed misfolded conformations. Even with the empirical choice of small γ (*e.g.*, 5 cal mol^{-1} Å^{-2} used in the examples discussed in the previous section), systematic over-stabilization of pair-wise interactions is evident. Furthermore, with such a small γ, the stability of fully folded (native) structures is likely under-estimated. These implications explain the observations derived from extensive REX-MD folding simulations of HP36 and protein A discussed in the previous section, and strongly argue that proper description of the length-scale dependence is critical in implicit modeling of hydrophobic interactions.

3.2 Solvent screening of dispersion interactions

SA models are based on the first-solvation-shell approximation, which assumes that energetics associated with various nonpolar solvation effects are largely dominated by short-range effects. However, these effects can have substantially different dependence on the solute composition and conformational state, and the simple linear relationship of eqn (1) can fail. For example, the total nonpolar solvation free energy can be further decomposed into a repulsive cavity term and an attractive solute–solvent dispersion interaction term,[60]

$$\Delta G_{\rm np} = \Delta G_{\rm cav} + \Delta G_{\rm vdW}, \quad (4)$$

where $\Delta G_{\rm cav}$ is the free energy cost of cavity formation in water, and $\Delta G_{\rm vdW}$ is the free energy for establishing the solute–solvent vdW dispersion interactions. $\Delta G_{\rm cav}$ and $\Delta G_{\rm vdW}$ have opposite signs and are anti-correlated with one another. It has been shown that only the cavity formation can be described by eqn (1) with a universal (*i.e.*, solute chemical composition independent) γ.[49,58,60] In contrast, the solute–solvent dispersion term depends strongly on the atomic composition of the solute,[58] and only approximately tracks the surface area. For example, while fitting the calculated solvation free energies of a series of n-alkanes in a fixed (all *trans*) conformation to eqn (1) yields $\gamma \sim 12.3$ cal mol^{-1} Å^{-2}, all conformational dependence of the solvation free energies of n-butane, n-pentane and n-hexane shows a much steeper

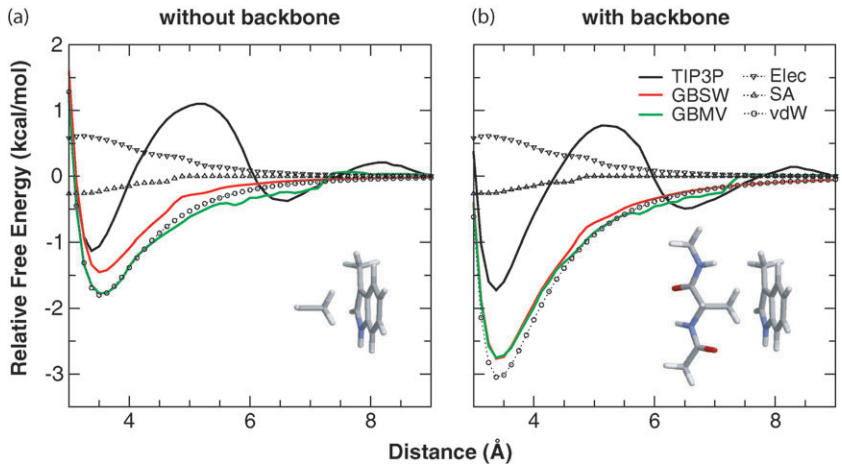

Fig. 4 PMFs of interactions between a Trp side chain and the alanine dipeptide (a) without and (b) with the peptide backbone (here the backbone atoms include the capping methyl groups), in three solvent models. The partial charge of C_α is reduced to −0.18 e.u. from −0.27 e.u. to maintain neutral total charge when the backbone is deleted. The backbone is in a fully extended (all *trans*) conformation. The reaction coordinate is defined as the distance from the C_β atom of alanine dipetide to the heavy atom plane of the Trp side chain. The same GBSW/SA and GBMV/SA parameters as in Fig. 1 and 2 were used. The stabilities (in units of kcal mol^{-1}) of dimers in TIP3P, GBSW/SA, and GBMV/SA are (a) without backbone: −1.13, −1.45, and −1.77; (b) with backbone: −1.73, −2.77, and −2.75, respectively. The (free) energy decompositions of the PMFs in GBSW/SA are shown by the dotted traces. The electrostatic component (Elec) includes both Coulomb interactions and GB electrostatic solvation free energy. The Elec, SA, and vdW contributions (in kcal mol^{-1}) to the dimer stabilities at the contact minima are (a) without backbone: +0.58, −0.23, and −1.80; (b) with backbone: +0.53, −0.25, and −3.05, respectively. Note that the contact minimum distance shifts slightly from 3.5 to 3.4 Å in all solvent models in the presence of the backbone atoms.

dependence on surface area, with a similar $\gamma \sim 110$ cal mol^{-1} Å^{-2}.[49] This can be explained by considering that ΔG_{vdW} is mainly determined by the number of carbons and is thus largely independent of the conformation state. As such, the conformational dependence of the solvation free energy is dominated by the cavity formation term. This highlights the necessity of the explicit decomposition of eqn (4) in order to properly describe the conformational dependence of nonpolar solvation. Furthermore, it has also been shown that buried atoms in proteins can contribute substantially to the solute–solvent dispersion interaction,[60,80] which can not be properly described by the solvent exposed area alone. Indeed, explicit inclusion of the solute–solvent dispersion interaction term has been shown to significantly improve the description of the conformational dependence of solvation free energy of small solutes[58,59] and proteins.[60,81]

The consequences of an insufficient description of the solute–solvent dispersion interaction by SA models on modeling protein conformational equilibria can be further understood by considering the influence on the balance of intramolecular vdW interactions and nonpolar solvation. Let's first consider the more realistic case of explicit solvent. Solutes are fully solvated at large separation, with extensive dispersion interactions with solvent. Upon contact, some solute atoms are buried and their dispersion interactions with solvent are substantially attenuated. The associated energy cost can be largely captured by the surface area reduction. While the rest of the solute atoms do not make any direct contact to contribute to the surface area change, there is a free energy cost due to displacement of solvent molecules by medium-range solute atoms. In other words, there is a non-negligible effective solvent screening of the medium-range solute–solute dispersion interactions. An insufficient description of such screening by SA models again over-estimates the strength of nonpolar interactions. This is illustrated in Fig. 4, where we compare interactions between a Trp side chain and the alanine dipeptide with and without the backbone atoms in TIP3P, GBSW/SA and GBMV/SA. The backbone atoms of alanine dipeptide do not make any direct contact with the Trp side chain even at the contact minimum. The presence of the backbone atoms increases the dimer stability by 0.6 kcal mol^{-1} in TIP3P, whereas the increases are 1.32 kcal mol^{-1} in GBSW/SA and 0.95 kcal mol^{-1} in GBMV/SA, larger than the TIP3P result. Free energy decomposition analysis reveals that the over-estimation of the stability increase is indeed due to neglecting solvent screening of the (medium-range) vdW dispersion interactions between the alanine dipeptide backbone and Trp side chain, which makes a full contribution of 1.25 kcal mol^{-1} to the stability increase in GB/SA models. Therefore, explicit inclusion of solute–solvent dispersion interaction, such as using the decomposition of eqn (4) is necessary to achieve a proper balance between nonpolar solvation and intramolecular vdW interactions.

3.3 Toward accurate implicit modeling of nonpolar solvation

The length-scale dependence of both hydrophobic solvation and solvent screening of solute–solute dispersion interactions need to be properly described for accurate implicit modeling of nonpolar solvation. It has been proposed that solvent screening of solute–solute dispersion interactions can be described using a continuum vdW solvent model.[60] The basic idea is to assume that the average water (oxygen) number density is constant outside of the solute volume, such that the atomic solute–solvent dispersion interaction energy can be evaluated

by volume integrals,

$$\Delta G_{\mathrm{vdW}} \sim \rho_w \sum_i \int_{\mathrm{solvent}} u_{\mathrm{vdW}}^{(i)}(|\boldsymbol{r} - \boldsymbol{r}_i|)\mathrm{d}\boldsymbol{r}, \qquad (5)$$

where ($\rho_{\mathrm{w}} = 0.0036$ Å^{-3}) is the water bulk number density at standard conditions. $u_{\mathrm{vdW}}^{(i)}$ is the solute–solvent dispersion interaction potential for atom i, which is typically defined based on the WCA decomposition of the Lennard-Jones potential.[60] In principle, the integral can be evaluated efficiently either using a surface integral approach,[59] or by a pair-wise descreening approximation.[82] It can also be directly evaluated using the same numerical quadrature techniques employed to calculate the Born radii.[64,83] The associated increase in the computational cost is either negligible or marginal. For example, less 1% increase in computational cost is observed when the numerical quadrature is used.[83] Note that both terms in eqn (4) are big and they largely cancel each other. The net nonpolar solvation energy, ΔG_{np}, is of at least one order of magnitude smaller than either ΔG_{cav} or ΔG_{vdW}. Therefore, a practical challenge is to achieve sufficient numerical accuracy for both components such that a reliable estimation of the net energetics can be made.

Proper description of the length scale dependence requires a reliable estimation of some effective local curvature, R_{c}. Theoretical relations can be then used to derive the local effective surface tension, such as in the scaled particle theory,[75]

$$\gamma(R_{\mathrm{c}}) \sim \gamma_{\infty}(1 - 2\delta/R_{\mathrm{c}}), \qquad (6)$$

where γ_{inf} is the limiting surface tension of a flat surface, and δ is the Tolman length. δ is of the order of the solvent size. Accurate estimation of the local surface curvature of complex molecular shapes is computationally expensive.[84] Alternatively, one might infer the effective curvature for each nonpolar group based on the number and type of contacts that it is involved in.[32] A local contact mass can be first computed,

$$M_c(i) = m(i) + \sum_{j \neq i} m(j)H(d_{ij}), \qquad (7)$$

where $m(i)$ is the effective mass of nonpolar group i. $H(d_{ij})$ is a contact switching function that decreases from one at short distance smoothly to zero at large distance. The inter-group average distance, $d_{ij} = (\sum\sum r_{\mathrm{nm}}^{-1})^{-1}$, allows some resolution of different contact poses for planar functional groups such as aromatic rings. The effective local curvature can then be estimated as

$$R_{\mathrm{c}} = R_0 + (\kappa M_{\mathrm{c}}(i))^{1/3}, \qquad (8)$$

where the constant κ is related the mass density and R_0 is related to the solvent size. eqn (8) can be viewed as the first term in a shape and density expansion, which represents a spherical shape with uniform mass density. Such a model allows an efficient estimate of the conformational dependence of γ, and is referred to as an SA model with varying γ (VGSA).[32] The main parameters include γ_∞ for each functional group. Additional, effective mass and solvent size parameters (R_0 and δ) might also be parameterized, such as to reproduce the solvation free energy of amino acid side chain analogues and stabilities of nonpolar interactions. It has been demonstrated that such a model has the ability to resolve the two main artifacts of SA models with fixed γ, suppressing over-stabilization of pair-wise interactions, and at the same time, correctly predicting cooperative three-body hydrophobic associations.[32]

To arrive at a working implicit solvent force field that can be reliably applied to model protein folding and conformational transitions, it is necessary to combine the explicit solute–solvent dispersion interaction term with a cavity term that includes conformation dependent γ's. Such a "complete" model should be carefully parameterized to capture the delicate balance between intramolecular vdW interactions and nonpolar solvation. Furthermore, co-optimization of the implicit solvent parameters together with the protein force field will be critical for achieving sufficient balance of electrostatic solvation, nonpolar solvation and various intramolecular interactions. Small model peptides and fast folding proteins with extensive experimental thermodynamic data will be extremely useful in such parameterization attempts.[15,25,29] It needs to be emphasized that even with a reduced representation of the system, achieving sufficient sampling of the protein conformational space remains one of the most challenging problems in computational biology. Advanced sampling techniques such as REX should continue to play an important role in improving the convergence of computed thermodynamic properties to allow direct comparison with experiments.

While conservatively optimistic, we recognize that actual improvement in modeling protein folding and conformational transitions with extended treatment of nonpolar solvation has yet to be demonstrated. In fact, there are many reasons to be pessimistic on whether a sufficiently balanced, yet reasonably efficient, implicit solvent force field for proteins can eventually be achieved. Some of main challenges include the following. First, an extremely high level of accuracy is required, considering that the average thermodynamic stability of proteins is only up to the order of 0.1 kcal mol^{-1} per residue.[85] Second, substantial limits exist in the modern protein force field underlying the implicit solvent models.[9,10] In particular, it is clear that there is a systematic tendency to under-estimate the solvation free energy of key protein functional groups[65] and to over-estimate solute–solute interactions.[86,87] The current limited success of implicit solvent force fields relies heavily on parameterization to achieve sufficient cancellation of errors. However, it seems that some aspects of the protein force fields, particularly regarding aromatic side chains,[66,67] will need to be improved along with the construction of the implicit solvent models. Third, conceptual difficulty remains in fast analytical estimation of the "local" surface curvature. The approximation of eqn (8) relies on contact order to infer the local curvature and has limited resolution of alternative packing geometries. Fourth, due to the short-range nature of dispersion interactions (with a r^{-6} dependence), the volume integral of eqn (5) is highly sensitive to the definition of surface and parameters such as solvent (probe) radius and input radii of solute atoms. Achieving sufficient (numerical) accuracy for both (large) terms in eqn (4) to arrive at a reliable estimation of the (small) net nonpolar solvation free energy is challenging in practice. Finally, further complications arise from breakdowns of the continuum approximation at short range. The

presence of backbone and other charged atoms can induce order or disorder in the local water structures, giving rise to nontrivial secondary contributions.[88] It has also been argued that polar and nonpolar solvation is coupled and one might need to solve an optimization problem to derive the most appropriate solvent accessible surface.[89] The implications of these secondary effects on the modeling of protein conformational equilibria are unclear, even though it appears that the current implicit solvent models are far too inaccurate to justify a meaningful attempt to incorporate these effects for protein simulations.

4. Concluding discussions

Implicit solvent has emerged as one of the most powerful techniques for classical simulation of proteins and other biomolecules in aqueous solution, offering a favorable compromise between speed and accuracy. Central to the implicit treatment of solvent is the estimation of the solvation free energy. The most accurate approaches require decomposition of the total solvation free energy into electrostatic and nonpolar components. Electrostatic solvation is usually described by the well-estabilished continuum electrostatics representation, and the associated energetics can be evaluated efficiently using the geneareralized Born (GB) approximation. The nonpolar solvation free energy is of smaller magnitude compared to the electrostatic component, and is often either largely ignored or simply estimated from the solvent accessible surface area (SA). As such, nonpolar solvation remains one of the least reliable aspects in most GB/SA models. One of the important limitations of SA models is the insufficient description of the conformational dependence of solvation. However, it is understood that capturing the delicate balance between intramolecular van der Waals (vdW) interaction and nonpolar solvation is critical for modeling protein conformational equilibria. Therefore, it is essential to further improve the implicit treatment of nonpolar solvation.

Comparison of the stabilities of nonpolar dimers in implicit and explicit solvents reveals a systematic bias to over-estimate the pair-wise nonpolar interactions in the current GB/SA models. Such over-stablilization does not only increase the roughness of the underlying potential surface and hinders rapid sampling of the conformational space, but also introduces severe distortion to the free energy landscape such that the global minimum may no longer correspond to the true native basin. Extensive folding and unfolding simulations of model peptides and proteins appear to support the above observations. These simulations also suggest that, with extensive optimization of the electrostatic solvation models, nonpolar solvation has become the critical bottleneck. For example, near native conformations could be reached for two small helical proteins, HP36 and protein A, with largely correct secondary elements arranged in native-like packings. The bottleneck to the fully folded conformations appears to be formation of hydrophobic cores for both proteins, highlighting the importance of re-balancing the nonpolar interactions.

We argue that two main physical properties of nonpolar solvation need to be properly described in order to improve beyond the current level of accuracy, the length-scale dependence of hydrophobic solvation and solvent screening of solute–solute dispersion interactions. Ignoring these two properties in the traditional SA models leads to systematic overstabilization of the pair-wise nonpolar interactions, consistent with the observations discussed above. Furthermore, the length-scale dependence of hydrophobic solvation gives rise to cooperativity in multi-body hydrophobic associations, while the popular SA models with a constant γ incorrectly predict anti-cooperativity. Promising solutions for explicit inclusion of both properties have been proposed. Solvent screening of dispersion interactions can be described by a continuum vdW solvent model, and the length-scale dependence of hydrophobic solvation might be captured by inferring the local curvature from the residue contact order. Extensive co-optimization of these models together with other components of the force field (*e.g.*, electrostatic solvation and intramolecular interactions) is necessary in order to eventually derive a fully balanced implicit solvent force field. The parameterization is expected to be a formidable task and substantial challenges exist. Nonetheless, we enthusiastically believe that it is critical to improve the implicit treatment of nonpolar solvation and that efforts along the directions outlined will substantially improve one's abililty to accurately model protein folding and conformational transitions.

Acknowledgements

This work was supported by the National Institutes of Health (RR12255).

References

1 S. Gnanakaran, H. Nymeyer, J. Portman, K. Y. Sanbonmatsu and A. E. Garcia, *Curr. Opin. Struct. Biol.*, 2003, **13**, 168–174.
2 O. Schueler-Furman, C. Wang, P. Bradley, K. Misura and D. Baker, *Science*, 2005, **310**, 638–642.
3 E. Shakhnovich, *Chem. Rev.*, 2006, **106**, 1559–1588.
4 L. G. Dunfield, A. W. Burgess and H. A. Scheraga, *J. Phys. Chem.*, 1978, **82**, 2609–2616.
5 P. Bradley, K. Misura and D. Baker, *Science*, 2005, **309**, 1868–1871.
6 T. Herges and W. Wenzel, *Biophys. J.*, 2004, **87**, 3100–3109.
7 Y. Zhang and J. Skolnick, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**, 1029–1034.
8 A. Kryshtafovych, C. Venclovas, K. Fidelis and J. Moult, *Proteins*, 2005, **61**(S7), 225–236.
9 J. W. Ponder and D. A. Case, *Adv. Protein Chem.*, 2003, **66**, 27–85.
10 A. D. MacKerell, Jr, *J. Comput. Chem.*, 2004, **25**, 1584–1604.
11 C. D. Snow, E. J. Sorin, Y. M. Rhee and V. S. Pande, *Annu. Rev. Biophys. Biomol. Struct.*, 2005, **34**, 43–69.
12 C. L. Brooks III, M. Karplus and B. M. Pettitt, *Proteins: A Theoretical Perspective of Dynamics, Structure, and Thermodynamics*, John Wiley and Sons, New York, 1987.
13 T. Yoda, Y. Sugita and Y. Okamoto, *Chem. Phys.*, 2004, **307**, 269–283.
14 W. Im, J. Chen and C. L. Brooks III, *Adv. Protein Chem.*, 2005, **72**, 173–198.
15 V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg and C. Simmerling, *Proteins*, 2006, **65**, 712–725.
16 M. Feig and C. L. Brooks III, *Curr. Opin. Struct. Biol.*, 2004, **14**, 217–224.
17 B. Roux and T. Simonson, *Biophys. Chem.*, 1999, **78**, 1–20.
18 D. Eisenberg and A. D. McLachlan, *Nature*, 1986, **319**, 199–203.
19 J. Wang, W. Wang, S. Huo, M. Lee and P. A. Kollman, *J. Phys. Chem. B*, 2001, **105**, 5055–5067.

20 P. Ferrara, J. Apostolakis and A. Caflish, *Proteins: Struct., Funct., Genet*., 2002, **46**, 24–33.
21 W. C. Still, A. Tempczyk, R. C. Hawley and T. Hendrickson, *J. Am. Chem. Soc*., 1990, **112**, 6127–6129.
22 C. J. Cramer and D. G. Truhlar, *Chem. Rev*., 1999, **99**, 2161–2200.
23 D. Bashford and D. A. Case, *Annu. Rev. Phys. Chem*., 2000, **51**, 129–152.
24 A. Masunov and T. Lazaridis, *J. Am. Chem. Soc*., 2003, **125**, 1722–1730.
25 J. Chen, W. Im and C. L. Brooks III, *J. Am. Chem. Soc*., 2006, **128**, 3728–3736.
26 J. Khandogin, J. Chen and C. L. Brooks III, *Proc. Natl. Acad. Sci. U. S. A*., 2006, **103**, 18546–50.
27 J. Khandogin, D. P. Raleigh and C. L. Brooks III, *J. Am. Chem. Soc*., 2007, **129**, 3056–3057.
28 W. Im, M. S. Lee and C. L. Brooks III, *J. Comput. Chem*., 2003, **24**, 1691–1702.
29 S. Jang, E. Kim and Y. Pak, *Proteins*, 2007, **66**, 53–60.
30 Y. Sugita and Y. Okamoto, *Chem. Phys. Lett*., 1999, **314**, 141–151.
31 M. Feig, J. Karanicolas and C. L. Brooks III, *J. Mol. Graph Modell*., 2004, **22**, 377–395.
32 J. Chen and C. L. Brooks III, *J. Am. Chem. Soc*., 2007, **129**, 2444–2445.
33 S. Jang, S. Shin and Y. Pak, *J. Am. Chem. Soc*., 2002, **124**, 4976–4977.
34 S. Jang, E. Kim, S. Shin and Y. Pak, *J. Am. Chem. Soc*., 2003, **125**, 14841–14846.
35 C. D. Snow, N. Nguyen, V. S. Pande and M. Gruebele, *Nature*, 2002, **420**, 102–106.
36 D. R. Roe, V. Hornak and C. Simmerling, *J. Mol. Biol*., 2005, **352**, 370–381.
37 A. Jagielska and H. A. Scheraga, *J. Comput. Chem*., 2007, **28**, 1068–1082.
38 H. Lei, C. Wu, H. Liu and Y. Duan, *Proc. Natl. Acad. Sci. U. S. A*., 2007, **104**, 4925–4930.
39 N. A. Baker, *Curr. Opin. Struct. Biol*., 2005, **15**, 137–143.
40 J. M. J. Swanson, S. A. Adcock and J. A. McCammon, *J. Chem. Theory Comput*., 2005, **1**, 484–493.
41 M. Nina, W. Im and B. Roux, *Biophys. Chem*., 1999, **78**, 89–96.
42 D. Chandler, *Nature*, 2005, **437**, 640–647.
43 C. J. Cramer and D. G. Trular, *Chem. Rev*., 1999, **99**, 2161–2200.
44 T. Ooi, M. Oobatake, G. Nemethy and H. A. Scheraga, *Proc. Natl. Acad. Sci. U. S. A*., 1987, **84**, 3086–3090.
45 T. Simonson and A. T. Brunger, *J. Phys. Chem*., 1994, **98**, 4683–4694.
46 C. Chothia, *Nature*, 1974, **248**, 338–339.
47 G. Nemethy and H. A. Scheraga, *J. Phys. Chem*., 1962, **66**, 1773–1789.
48 R. A. Pierotti, *Chem. Rev*., 1976, **76**, 717–726.
49 H. S. Ashbaugh, E. W. Kaler and M. E. Paulaitis, *J. Am. Chem. Soc*., 1999, **121**, 9243–9244.
50 T. M. Raschke, J. Tsai and M. Levitt, *Proc. Natl. Acad. Sci. U. S. A*., 2001, **98**, 5965–5969.
51 D. Sitkoff, K. A. Sharp and B. Honig, *J. Phys. Chem*., 1994, **98**, 1978–1988.
52 C. Tanford, *Proc. Natl. Acad. Sci. U. S. A*., 1979, **76**, 4175–4176.
53 K. A. Sharp, A. Nicholis, R. F. Fine and B. Honig, *Science*, 1991, **252**, 106–109.
54 J. Zhu, E. Alexov and B. Honig, *J. Phys. Chem. B*, 2005, **109**, 3008–3022.
55 A. Onufriev, D. Bashford and D. A. Case, *Proteins*, 2004, **55**, 383–394.
56 B. Lee and F. M. Richards, *J. Mol. Biol*., 1971, **55**, 379.
57 M. Feig, A. Onufriev, M. S. Lee, W. Im, D. A. Case and C. L. Brooks III, *J. Comput. Chem*., 2004, **25**, 265–284.
58 E. Gallicchio, M. M. Kubo and R. M. Levy, *J. Phys. Chem. B*, 2000, **104**, 6271–6285.
59 M. Zacharias, *J. Phys. Chem. A*, 2003, **107**, 3000–3004.
60 R. M. Levy, L. Y. Zhang, E. Gallicchio and A. K. Felts, *J. Am. Chem. Soc*., 2003, **125**, 9523–9530.
61 A. Ben-Naim and R. Mazo, *J. Phys. Chem*., 1993, **97**, 10829–10834.
62 A. D. MacKerell, Jr, D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher III, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin and M. Karplus, *J. Phys. Chem. B*, 1998, **102**, 3586–3616.
63 W. L. Joregensen, J. Chandrasekhar, J. D. Madura, R. W. Impey and M. L. Klein, *J. Chem. Phys*., 1983, **79**, 926–35.
64 M. S. Lee, F. R. Salsbury, Jr and C. L. Brooks III, *J. Chem. Phys*., 2002, **116**, 10606–10614.
65 M. R. Shirts and V. S. Pande, *J. Chem. Phys*., 2003, **119**, 5740–5761.
66 Z. Xu, H. H. Luo and D. P. Tieleman, *J. Comput. Chem*., 2007, **28**, 689–697.
67 C. M. Baker and G. H. Grant, *J. Chem. Theory Comput*., 2007, **3**, 530–548.
68 C. J. McKnight, D. S. Doerin, P. T. Matsudaira and P. S. Kim, *Nat. Struct. Biol*., 1997, **4**, 180.
69 H. Gouda, H. Torigoe, A. Saito, M. Sato, Y. Arata and I. Shimada, *Biochemistry*, 1992, **31**, 9665.
70 M. Feig, A. D. MacKerell, Jr and C. L. Brooks III, *J. Phys. Chem*., 2003, **107**, 2831–2836.
71 A. D. MacKerell, Jr, M. Feig and C. L. Brooks III, *J. Am. Chem. Soc*., 2004, **126**, 698–699.
72 M. Feig, J. Karanicolas and C. L. Brooks III, *MMTSB Tool Set, MMTSB NIH Research Resource*, The Scripps Research Institute, 2001.
73 H. Lei and Y. Duan, *J. Mol. Biol*., 2007, **370**, 196–206.
74 R. C. Tolman, *J. Chem. Phys*., 1949, **17**, 333–337.
75 H. Reiss, H. L. Frisch and J. L. Lebowitz, *J. Chem. Phys*., 1959, **31**, 369–380.
76 K. Lum, D. Chandler and J. D. Weeks, *J. Phys. Chem. B*, 1999, **103**, 4570–4577.
77 D. M. Huang and D. Chandler, *Proc. Natl. Acad. Sci. U. S. A*., 2000, **97**, 8324–8327.
78 S. Shimizu and H. S. Chan, *J. Chem. Phys*., 2000, **113**, 4683–4700.
79 C. Czaplewski, A. Liwo, D. R. Ripoll and H. A. Scheraga, *J. Phys. Chem. B*, 2005, **109**, 8108–8119.
80 J. W. Pitera and W. F. van Gunsteren, *J. Am. Chem. Soc*., 2001, **123**, 3163–3164.
81 Jason A. Wagoner and Nathan A. Baker, *Proc. Natl. Acad. Sci. U. S. A*., 2006, **103**, 8331–8336.
82 E. Gallicchio and R. M. Levy, *J. Comput. Chem*., 2004, **25**, 479–499.
83 W. Im, J. Chen and C. L. Brooks, unpublished work.
84 D. H. Ballard and C. M. Brown, *Computer Vision*, Prentice-Hall, Inc., Englewood Cliffs, NJ, 1982.
85 A. G. Cochran, N. J. Skelton and M. A. Starovasnik, *Proc. Natl. Acad. Sci. U. S. A*., 2001, **98**, 5578–5583.
86 M. Kang and P. E. Smith, *J. Comput. Chem*., 2006, **27**, 1477–1485.
87 J. Chen, C. L. Brooks and H. A. Scheraga, *J. Phys. Chem. B*, 2007, in press.
88 L. R. Pratt, *Annu. Rev. Phys. Chem*., 2002, **53**, 409–436.
89 J. Dzubiella, J. M. J. Swanson and J. A. McCammon, *Phys. Rev. Lett*., 2006, **96**, 087802.