PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

The VSGB 2.0 model: A next generation energy model for high resolution protein structure modeling

Jianing Li,[1] Robert Abel,[2] Kai Zhu,[2] Yixiang Cao,[2] Suwen Zhao,[3] and Richard A. Friesner[1]*

[1] Department of Chemistry, Columbia University, New York, New York
[2] Schrödinger, Inc., New York, New York
[3] Department of Pharmaceutical Chemistry, University of California, San Francisco, California

ABSTRACT

A novel energy model (VSGB 2.0) for high resolution protein structure modeling is described, which features an optimized implicit solvent model as well as physics-based corrections for hydrogen bonding, π–π interactions, self-contact interactions, and hydrophobic interactions. Parameters of the VSGB 2.0 model were fit to a crystallographic database of 2239 single side chain and 100 11–13 residue loop predictions. Combined with an advanced method of sampling and a robust algorithm for protonation state assignment, the VSGB 2.0 model was validated by predicting 115 super long loops up to 20 residues. Despite the dramatically increasing difficulty in reconstructing longer loops, a high accuracy was achieved: all of the lowest energy conformations have global backbone RMSDs better than 2.0 Å from the native conformations. Average global backbone RMSDs of the predictions are 0.51, 0.63, 0.70, 0.62, 0.80, 1.41, and 1.59 Å for 14, 15, 16, 17, 18, 19, and 20 residue loop predictions, respectively. When these results are corrected for possible statistical bias as explained in the text, the average global backbone RMSDs are 0.61, 0.71, 0.86, 0.62, 1.06, 1.67, and 1.59 Å. Given the precision and robustness of the calculations, we believe that the VSGB 2.0 model is suitable to tackle "real" problems, such as biological function modeling and structure-based drug discovery.

Proteins 2011; 79:2794–2812.

Key words: energy model; all-atom force field; protonation state assignment; side chain prediction; loop prediction.

INTRODUCTION

Knowledge of protein structure at atomic resolution is essential for modeling biological function and structure-based drug discovery approaches.[1–3] Although the generation of experimental structures, propelled by high throughput crystallography, continues to advance exponentially, the number of known protein sequences is growing even more rapidly. Furthermore, for any given sequence, there may be a significant number of biologically relevant conformations, not to mention possible structural reorganization associated with ligand binding or with protein–protein interactions. Hence, it is unlikely that the entire universe of biologically relevant protein structural data can be accessed by exclusively experimental means.

Computational modeling represents the logical approach to constructing protein structures that are not experimentally available. The coverage of protein families continues to increase rapidly in the Protein Data Bank (PDB), which implies that the vast majority of protein structure prediction problems involve perturbation of a known structure by a relatively small RMSD. Homology modeling, using sequence and profile-based approaches,[4–6] continues to make great progress, and models with the correct architecture and low RMSD can be built for substantial fraction of interesting cases, particularly for pharmaceutically relevant targets where substantial experimental work on the protein family to which the target belongs (e.g., kinases) has typically been performed.[7]

However, to predict relative protein conformation energetics and protein-ligand binding affinities, very high resolution structures are required, and current homology models are often not quite good enough for this purpose (although the suitability varies depending upon the target and the specific project for which the structure will be employed).[8] The technology that would address this problem is refinement of homology models, in which the RMSD of the homology model is progressively reduced until it is suitably close to the native structure. Such a refinement strategy in turn requires a sufficiently accurate potential energy function, including modeling of solvation effects and detailed physical chemistry of protein interactions, for example, hydrogen bonds. If the potential energy

Additional Supporting Information may be found in the online version of this article.
Grant sponsor: NIH; Grant number: GM-40526.
*Correspondence to: Richard A. Friesner, Department of Chemistry, Columbia University, New York, NY 10027. E-mail: rich@chem.columbia.edu.
Received 23 December 2010; Revised 3 May 2011; Accepted 13 May 2011
Published online 11 July 2011 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/prot.23106

surface has a free energy minimum that is distinct from the native structure, this implies a fundamental limitation on the RMSD that can be achieved. With an accurate potential surface, one is then left with the problem of sufficiently robust and comprehensive sampling of phase space, a challenging task given that the homology model may deviate from the native structure at any location in the protein.

There are various possible strategies that can be employed to carry out refinement. The most straightforward approach would be to perform a molecular dynamics simulation using an all-atom protein model and explicit representation of aqueous solvent. However, such simulations are very expensive computationally, and even assuming that the potential functions used in the simulation (which at present generally do not incorporate polarizability, for example) are adequate to yield the native structure as a free energy minimum, the effort required would be extremely large even for a small protein, and prohibitive for larger proteins and protein assemblies which constitute the great majority of biologically interesting systems.

An alternative approach to refinement is to utilize a continuum representation of solvent, along with an all-atom protein force field. Continuum approaches have two major advantages. First, it is not necessary to average over the positions of explicit water molecules, which generally requires very lengthy convergence times. Second, conformational search methods, as opposed to molecular dynamics, can conveniently be employed in conjunction with a continuum solvation model. Such methods can be many orders of magnitude more efficient than molecular dynamics for locating the global free energy minimum of the model, because much larger steps in phase space can be taken. The generalized Born continuum solvent model, in particular, is relatively inexpensive to evaluate, and is amenable to calculation of gradients, which are necessary for minimizations.

Although considerable progress has been made, in our group and others, in advancing the state of the art in continuum solvation calculations,[9–11] two principal problems still remain. First, as in any energy model applied to high resolution protein structure refinement, the model must be accurate enough to actually improve homology models beyond their current level of resolution. In considering the accuracy of a force field plus a continuum solvation model, what matters is the potential energy surface defined by the model as a whole, as opposed to the individual components. The problem is in some ways more challenging than that of constructing force fields for explicit solvent simulation, because development of an accurate continuum model requires guessing and experimentation, since the functional form represents a reduction of the true Hamiltonian of the system to a non-rigorous model approximation (as is the case, for example, in density functional theory in electronic structure). Hence, the accuracy of continuum models requires continuous improvement, either by comparison with explicit solvent simulations, experimental data, or both. Second, the sampling problem remains one of great difficulty, although amenable to a wide range of algorithmic acceleration due to the ability to employ conformational search methods of various types.

In previous study, we have made a series of improvements in both the energy model and sampling algorithms implemented in our continuum based Protein Local Optimization Program (PLOP).[12–17] These improvements have enabled reasonable results to be obtained for loop predictions up to 20 residues. However, a non-trivial fraction of test cases continued to exhibit large RMSDs, in some cases accompanied by large energy errors (defined as the energy gap between the predicted and minimized native structures). These results, while encouraging, reflected the fact that there was still important missing physics in our previous energy models.

In this article, we describe the most recent version of our energy model, the VSGB 2.0 model, which has been rigorously optimized by fitting to accurate experimental side chain and loop (11–13 residues) data, and contains a number of new terms not incorporated into the older functional form, as well as many reoptimized model parameters. We evaluate the performance of the VSGB 2.0 model by predicting structures for a set of 115 super long loops of 14–20 residues. At these lengths, alternative approaches in the literature uniformly display rapidly increasing RMSD errors,[18,19] a reflection of the greatly expanded conformational freedom associated with these very long loops; and in fact, there are very few prior studies in which loops of such lengths have been investigated systematically using a large data set. Remarkably, despite the exceptionally demanding test set (for which no parameter adjustment was made to improve agreement with experiment), a high degree of robustness, and small backbone and side chain RMSD, are demonstrated for 100% of the test cases. Achieving this level of accuracy requires other improvements besides the energy model, most prominently continued advances in the sampling algorithms, and application of a reliable approach to assigning protonation states, both of which are described in what follows. Given the precision and uniform robustness of the calculations, we believe that the VSGB 2.0 model and sampling algorithms, for the first time, are suitable for successfully tackling the "real" problem defined above, refining homology models (although augmented sampling algorithms will be required to perform what is clearly a more global sampling challenge than prediction of a single loop, no matter how long, and key protonation states will have to be sampled, rather than deduced from the native structure).

Very different philosophies have been used over the past decade to optimize and to evaluate atomic level protein models based on continuum solvation description.

Alternatives have included fitting generalized Born models to Poisson–Boltzmann (PB) results (note that the PB model itself has to be parameterized in some fashion),[20,21] exploring performance in the folding of small proteins,[22,23] and comparisons with explicit solvent simulations.[24,25] The present work is distinguished by the approach of fitting parameters to a large database of crystallographic single side chain and loop data (including a number of novel terms, one of which, the variable dielectric model, approximately incorporates polarization, and has proved to be extremely important in obtaining quantitatively useful results), and rigorously evaluating structural prediction accuracy for a large and demanding test set, the long loop data set described above. In our view, the use of these large training and test sets eliminates the possibility of overfitting, and provides confidence that the physics of the model is correct, and the right answers are being obtained for the right reason.

The article is organized as follows. We first discuss our selection of the training (side chain prediction, loop prediction) and test (super long loop prediction) sets. The use of an improved training set (as compared to that employed in Reference 15) turns out to be very important; our previous training set, despite the overall structural accuracy implied by the crystallography, contained side chains with ambiguous atom placement due to missing electron density in the crystallographic data. We then briefly review our algorithms for side chain and loop prediction, which have been described previously.[12–14] The VSGB 2.0 model is then discussed in detail, as well as the optimization protocol based on single side chain and 11–13 residue loop predictions. Results for the test set of super long loop prediction (length of 14–20 residues) are then presented, along with the discussion of the results. Finally, in the conclusion, we summarize the results and consider future directions.

MATERIALS AND METHODS

Selection of data sets

The rapid increase in the number of high resolution X-ray crystallographic structures has allowed us to build reliable data sets for training and testing the VSGB 2.0 model. To ensure the high quality, the data sets were selected based on the following criteria:

1. All structures are PDB X-ray crystallographic structures with low sequence identity (no more than 30% similarity) and high resolution (better than 1.0 Å for the side chain set and 2.0 Å for the loop sets).
2. Side chains or loops should not have multiple occupancy or missing heavy atoms.
3. Side chains or loops are not affected by ligands. The distance between the side chain/loop and a ligand is defined as the shortest heavy atom distance. The minimum distance allowed to any organic ligands is 4.0 Å and to any metal ions is 6.5 Å.
4. The average *B*-factor should be lower than 35.00.
5. The real space *R*-factor (RSR) of each residue should be lower than 0.200.
6. All atoms of the side chains or loops should be found to occupy well defined peaks in the experimentally determined electron density when visualized.

With all these criteria, we have collected 2239 single side chains from 45 proteins for the side chain training set, 100 loops (length of 11–13 residues) from 72 proteins for the loop training set, and 115 super long loops (length of 14–20 residues) from 97 proteins for the loop test set. The composition of proteins in these data sets is listed in Tables S1–S3 in the Supporting Information.

Preparation of all-atom models

As most crystallographic structures do not contain the hydrogen positions, it is necessary to add hydrogen atoms and determine the protonation states of ionizable residues for calculations at an atomic level of resolution. Additionally, the ambiguous orientation of Asn, Gln, and His due to the similar electron density of two alternative conformations rotated by 180° also impairs the correct physics of the models. Therefore, given the heavy-atom coordinates from X-ray crystallography, we created all-atom models for each protein using the Interaction Cluster Decomposition Algorithm (ICDA).[26] The ICDA assigns protonation states of ionizable residues, conformations of Asn, Gln, and His, and hydrogen positions of hydroxyls, by constructing clusters of potentially interacting side chains, enumerating a list of possible hydrogen bonding networks, and ranking these potential networks via energy evaluation. We should also note that in this work we improved the original ICDA algorithm reported in Reference 26 by using self-adjusted cluster sizes and more rigorous energy evaluation; the details of these improvements will be described elsewhere.

In addition to the protein, crystal environment, organic ligands, and metal ions were also taken into account for a fair comparison to the crystal structures obtained from experiments. Since the role of crystal environment and ligands in protein structure prediction has been extensively discussed,[13,27] their inclusion in our all-atom models was done in an automated fashion.

Single side chain prediction algorithm

Single Side Chain Prediction (SSCP) is defined as prediction of the conformation of one side chain with the rest of the protein fixed at the atomic positions of the native structure. The algorithm exhaustively samples side chain conformations with a residue-specific rotamer library at a high resolution.[28] Clash-free conformations are evaluated and sorted according to the single point

energy or the energy after minimization. The final prediction is determined by the lowest energy conformation, either with or without minimization. In this algorithm, all the conformations that survive after the steric clash check are kept for the evaluation stage. This is a modification to the original algorithm employed in previous publications,[12,15] which prescreens all the candidates with a reduced energy score and clusters the remaining conformations to only consider the cluster representatives. Since the total energy score is the only measure to evaluate all the conformations in the pool, the current version of SSCP algorithm is better able to provide a direct comparison of how well the energy models can distinguish the native from the non-native conformations in realistic applications such as loop prediction, where all of conformational space is considered, and minimization of the loop (which includes all side chain degrees of freedom) is employed. The optimized parameters obtained from SSCP fitting are discussed below.

Loop prediction (length of 11–13 residues) with decoys

Our loop prediction (length of 11–13 residues) was carried out with decoys. Conformations in the decoy set were generated from the loop predictions described in previous work.[14] Each loop case contains thousands of conformations in the decoy set, representing a wide spread of samples in the conformational space. For the purpose of optimizing the energy model (more specifically the hydrophobic term), the minimized conformation with the lowest energy was selected as the prediction. Loop prediction of 11–13 residues was employed to optimize the hydrophobic term, because the hydrophobic term is much larger in a loop than in a side chain and thus more sensitive in loop prediction. The use of decoys avoids the costly sampling in a vast conformational space, greatly reduces the sampling cost, and consequently allows fast optimization of the hydrophobic term. The optimized functional form and parameters are discussed below.

Super long loop prediction (length of 14–20 residues) algorithm

Our algorithm of Super Long Loop Prediction (SLLP) uses a hierarchical approach combined with an advanced sampling method to predict loops longer than 13 residues. Compared to the shorter loop prediction, super long loop prediction requires more intensive sampling effort. To improve both accuracy and efficiency, the algorithm uses a advanced sampling method based on a detailed dipeptide backbone rotamer library, which was first described in previous work[17] (with *trans* rotamers only) and improved in this work with addition of *cis* rotamers. In the SLLP algorithm, the loop candidates are first constructed without any constraint at the initial stage, whereas the rest of the protein remains the same as the native. An exhaustive search for possible loop conformations is carried out in two constrained refinement stages and a series of fixed stages.[14] At the end of each stage, the loop conformations generated in all the previously finished stages are ranked according to the energy after minimization, and the top ones without redundancy are sent to the next stage. Finally, the loop conformation with the lowest minimized energy is selected as the prediction.

To accurately predict protein loops longer than 13 residues, it is extremely important to generate conformations as close as possible to the native at the early sampling stages. During the initial stage, a large number of loop structures are constructed from five parallel calculations with overlap factor thresholds 0.45, 0.50, 0.55, 0.60, and 0.65. An overlap factor is defined as the ratio of distance between two atoms and the sum of their van der Waals radii. Loop conformations are considered with clashes and then abandoned, if any pair of atoms is found with an overlap factor lower than the threshold. A high overlap factor threshold sometimes causes sampling failure in a confined space or generates similar loop candidates. Therefore, multiple low overlap factors in the initial stage are important to generate a wide variety of loop candidates for the later stages.

The fixed stages, which sample a sub-region of the loop, are crucial in our hierarchical method for super long loop prediction. During the fixed stages, a large conformational space is searched and the native-like conformations are enriched progressively among the top loop candidates with the lowest energies. A "Fix N" stage only samples the residues that are outside of the fixed N residues in the loop. Since these N residues can start from either the C- or the N-terminal of the loop in question, all the N+1 combinations are considered in each "Fix N" stage. Sampling up to the "Fix 5" stage is adequate for the accurate prediction of many super long loops, but we extended to the "Fix 10" stage which successfully addressed a relatively small number of cases with serious sampling errors. The extension of the fixed stages up to 10 residues appears to be sufficient to essentially eliminate major sampling errors for loops up to 20 residues in length. It is likely though that additional sampling effort would benefit loops longer than 18 residues, where a jump in the average RMSD can be seen as compared to loops in the 14–18 residue length regimes (see the results below). Nevertheless, the current algorithm is sufficient to eliminate any errors greater than an RMSD of 2.0 Å even for the longer (19th and 20th residues) loops. On the other side, sampling that stops at "Fix 5" could have possible statistical bias, which we further discuss in the Discussion section.

Super long loop prediction method incorporating surrounding side chains

To further test the effectiveness of the VSGB 2.0 model, we also performed super long loop prediction

incorporating surrounding side chains (SLLP-SS). The method has been previously described by Sellers *et al.*,[29] in which the surrounding side chains that have heavy atoms within 7.5 Å from any C-beta atoms in the target loop are optimized simultaneously with the loop. In SLLP-SS, these surrounding side chains are temporarily removed when the backbone of the loop is being sampled. Then the side chains on the loop as well as those in the surroundings are put back and optimized by rotamer library sampling and minimization. Because the conformational phase space increases substantially when the surrounding side chains are optimized in addition to the loop, our most extensive hierarchical approach for long loop prediction is employed using the sampling stages "Fix 1" through "Fix 10."

The VSGB 2.0 model

Energy function

The VSGB 2.0 model provides a novel form of the energy function [Eq. (1)], which contains the OPLS-AA protein force field bonded and nonbonded terms, as well as a solvation term and a number of physics-based correction terms. The solvation free energy G_{sol} and the components of physics-based corrections $E_{corrections}$ are described in detail below.

$$\begin{aligned} G_{total} = & \sum_{bonds} k_b(r - r_0)^2 + \sum_{angles} k_\theta(\theta - \theta_0)^2 \\ & + \sum_{torsions} \frac{V_n}{2}[1 + \cos(n\phi - \delta)] + \sum_{impropers} k_\varphi(\varphi - \varphi_0)^2 \\ & + \sum_{electrostatics} \frac{q_i q_j}{r_{ij}\varepsilon_{in(ij)}} + \sum_{VDW} \left[\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}}\right] \\ & + G_{sol} + \sum E_{corrections} \end{aligned} \tag{1}$$

Solvent model

Solvation and desolvation effects are among the most important factors to determine a protein's global and local conformations in solvent.[30] The VSGB 2.0 model approximates the solvation free energy with an optimized implicit solvent model, which is based on Surface Generalized Born (SGB) model[10,31] and the variable dielectric (VD) treatment of polarization from protein side chains.[15] The SGB model, as an approximation to the Poisson-Boltzmann Equation (PBE), has been widely used in protein structure modeling. The variable dielectric treatment further improves the accuracy of the SGB model by varying the internal dielectric constants from 1.0 to 4.0 to incorporate the polarization effects. In this study, the VD-SGB implicit solvent model, as an important component of the VSGB 2.0 model, has been further optimized via fitting to single side chain predictions.

In a typical GB model, the solvation free energy (G_{sol}) is expressed as the sum of a cavity term (G_{cav}), a van der Waals term (G_{vdw}) and a polarization term (G_{pol}) [Eq. (2)].[32,33]

$$G_{sol} = G_{cav} + G_{vdw} + G_{pol} \tag{2}$$

The nonpolar solvent–solute interaction is usually represented by the sum of the cavity term and the van der Waals term, which is considered proportional to the solvent-accessible surface area (SASA) [Eq. (3)].

$$G_{cav} + G_{vdw} = \sigma \cdot SASA \tag{3}$$

However, such a surface area model has been found insufficient to account for the nonpolar solvation effect (e.g., dispersion) in many previous studies.[34,35] Therefore in our VSGB 2.0 model, the nonpolar contribution of solvation is calculated by a parameterized hydrophobic term described in detail below. The polar solvent-solute interaction is represented by the polarization term which depends on the solvent and internal dielectric constants, the partial charges, and f_{GB} [Eq. (4)].

$$G_{pol} = -\frac{1}{2}\left(\frac{1}{\varepsilon_{in(ij)}} - \frac{1}{\varepsilon_{sol}}\right)\sum_{i<j}\frac{q_i q_j}{f_{GB}} \tag{4}$$

f_{GB} is a function of the distances between two atoms (r_{ij}) and their generalized Born radii (α_i and α_j) of the form described in Reference [32].

$$f_{GB} = \sqrt{r_{ij}^2 + \alpha_{ij}^2 e^{-D}}\ , \quad \varepsilon_{in(ij)} = \mathrm{Max}(\varepsilon_{in(i)},\ \varepsilon_{in(j)}) \tag{5}$$

$$\alpha_{ij} = \sqrt{\alpha_i \alpha_j}\ , \quad D = \frac{r_{ij}^2}{(2\alpha_{ij})^2} \tag{6}$$

In Eq. (5), the internal dielectric constant $\varepsilon_{in\ (ij)}$ can vary from 1.0 to 4.0 as the maximum value of the internal dielectric constants of atom *i* and atom *j*. Table I shows the assignment of internal dielectric constants in the optimized VD-SGB model and in the original VD-SGB model.

The optimized VD-SGB model reduces the values of internal dielectric constants for Lys, Glu, protonated His, and neutral His while increases the values for Asp and Arg. These changes, although derived from parameterization, actually incorporate a better physical picture from two main aspects: first, the dielectric constant assigned to the neutral His is adjusted to be identical to other noncharged amino acid residues; second, the internal dielectric constants for the acidic amino acid residues (Asp and Glu) are tuned to have closer values, more consistent with their chemical similarity.

Table I
Internal Dielectric Constants of the Original and Optimized VD-SGB Model[a]

Residue	Lys	Glu	Hip[b]	Asp	Arg	His[b]	Other
Original VD-SGB model	4.00	3.00	3.00	2.00	2.00	2.00	1.00
Optimized VD-SGB model	3.85	2.78	2.86	2.44	2.11	1.00	1.00

[a]The assignment of internal dielectric constants is based on an atom-based scheme: only the charged atoms and the atoms adjacent to the charged ones are assigned with values >1.00.
[b]Hip, protonated histidine; His, neutral histidine.

Hydrogen bonding correction

An accurate description of hydrogen bonds is critical to predicting protein structure at high resolution. While a conventional fixed charge force field such as OPLS-AA does a reasonable job of getting the magnitude of hydrogen bonding interactions right, it is limited by having an atomic point charge description of electrostatics, as opposed to a more accurate multipole or lone pair description. One approach would be to improve the electrostatics by explicit addition of such higher order terms, as has been done in the AMOEBA force field.[36] An alternative is to use an empirical functional form to enforce hydrogen bond angles and distances, fitting to experimental PDB data. We have chosen to use the latter approach, following work of Baker and coworkers[37] as in the spirit of exploiting the large amount of experimental structural data in the PDB to achieve the highest possible accuracy for protein structure specifically. The new terms are added as a correction to the existing OPLS-AA force field and as part of the VSGB 2.0 model, and the parameters are optimized by fitting to improve single side chain prediction accuracy, as is discussed below.

The hydrogen bonding correction E_{HB} term is a function of distances, angles and atom types [Eq. (7)]. As the implicit solvent model is used, this correction is not applied to protein-solvent hydrogen bonding, which can cause underestimation of the interaction between protein and the first-shell solvent (see discussion below).

$$E_{HB} = \sum_i \sum_j \alpha_i \alpha_j \exp[-(r^{HA} - r_0)^2] \cos(\theta^{DHA} - \theta_0) \quad (7)$$

where i and j are heavy atoms involved in a hydrogen bond. The parameters of hydrogen bonding geometry r_0 (optimal distance between the hydrogen atom and the acceptor atom, 1.94 Å) and θ_0 (optimal angle of the donor atom, the hydrogen atom and the acceptor atom, 160°) were adopted from the Density Functional Theory (DFT) optimized formamide dimer and acetamide dimer.[37] α_i and α_j are coefficients related to the roles of the heavy atoms playing in a hydrogen bond, one as a donor while the other as an acceptor. For an atom i, a_i is assigned based on the following rules:

1. Positively charged nitrogen atoms in the side chains of Lys, Arg, charged His, and the N-terminal backbone are strong donors, $a_i = 1.5$
2. Negatively charged oxygen atoms in the side chains of Asp, Glu, and the C-terminal backbone are strong acceptor, $a_i = -1.5$
3. Polar atoms in the side chains of Ser, Thr, Asn, Gln, neutral His, Tyr, and Trp can be either weak donors or acceptors. The assignment is dependent on the paired atom j: if atom j is a strong donor, atom i is a weak acceptor; if atom j is a strong acceptor and atom i has at least one bonded hydrogen atom, atom i is a weak donor; otherwise, atom i is counted twice as a weak donor and a weak acceptor whereas atom j as a weak acceptor and a weak donor, respectively (as long as both have hydrogen atoms). For the weak donor, $ai = 0.5$; for the weak acceptor, $ai = -0.5$.
4. A neutral backbone oxygen atom is considered as a weak acceptor while a neutral backbone nitrogen atom is considered as a weak donor.

The hydrogen bonding correction depends on the distance between hydrogen atom and the acceptor atom r^{HA}, as well as the angles formed by donor atom, hydrogen atom and acceptor atom θ^{DHA} (Figure 1). Such a design involves the most relevant atoms in a hydrogen bond while being easy to implement and inexpensive to calculate.

π–π packing correction

π–π stacking is one of the main driving forces to stabilize vertical base stacks in DNA and the hydrophobic cores in proteins.[38–41] It also plays important chemical and biological roles in processes like self-assembly[42] and molecular recognition.[43] In contrast to covalent bonds and hydrogen bonds, π–π interactions are slightly directional with diverse preferred configurations, such as parallel stacking (sandwich and parallel-displace) and T-stacking (also known as T-shape),[44,45] which are generally referred to as π–π packing interactions in this work.

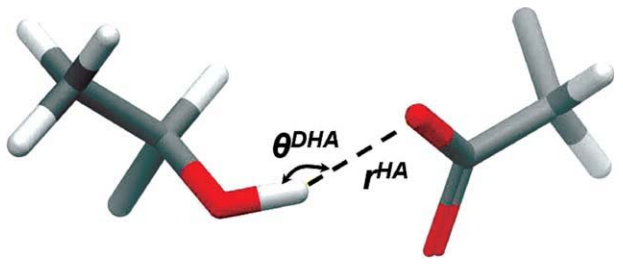

Figure 1
Illustration of geometry variables in hydrogen bonding correction. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

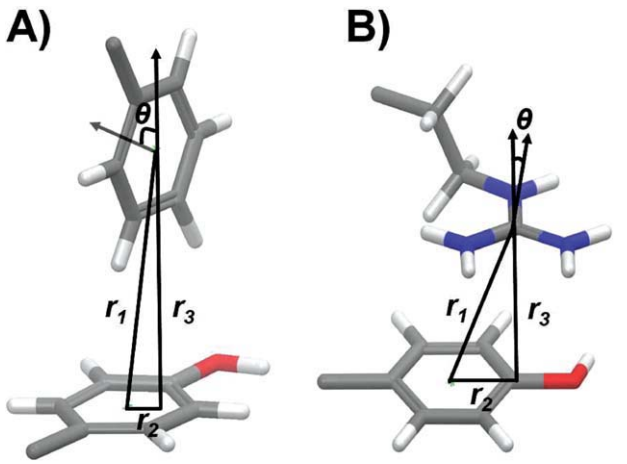

Figure 2

Illustration of geometry variables in π–π packing correction. **A**: T-stacking. **B**: Parallel stacking. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

However, it is difficult to rigorously separate packing interactions from other nonbonded terms such as van der Waals interactions. Two approaches can be used to provide estimates of the magnitude of the packing effects; observation of π–π interactions in native protein structures, via crystallography, and quantum chemical calculations of such interactions, using high level theories that adequately capture electron correlation effects.[38,46] For example, Burly and Petsko found that the preferential distance for π–π packing interaction is 4.5–7.0 Å and the preferential dihedral angles are close to 90° in 34 protein crystallographic structures.[38] An estimation of the free energy contribution from Burly and Petsko is between −0.6 and −1.3 kcal/mol. The quantum level study of Jurecka *et al.* gave a higher estimation as −2.5 and −7.0 kcal/mol for aromatic amino acid dimers.[46]

Our own investigations involve an analysis of our side chain and loop data sets, and comparison of predicted and native structures using our previous energy model, which did not contain an explicit π–π packing term. When such a model is used, the native structures contain a systematically higher percentage of π–π interactions (estimated by geometrical criteria) than is seen in predicted structures. These observations have motivated the development of a stacking term, empirically optimized to reproduce side chain structures, and then tested via loop prediction.

Given the insufficient treatment of π–π packing interaction in standard force field methods,[47] it would be useful to design a new π–π packing correction to improve the accuracy of the previous energy model. Here, we present an explicit π–π packing correction in the VSGB 2.0 model for pairs of amino acid side chains including the conventional aromatic ones such as Phe, Tyr, His, and Trp as well as the Y-aromatic structures such as Arg, Asn and Gln.[48] To reduce the complexity of the algorithm, only side chain-side chain packings are considered, although π–π stacks also exist in interactions involving protein backbone. In Eq. (8), the π–π packing correction is expressed as a function of distances r_1, r_2, and r_3 as well as dihedral angle θ (see Figure 2), which are defined as follows:

r_1: distance between the centers in aromatic planes of two side chains. r_1 should be within 0.0 to 8.0 Å.

r_3: distance from the center in one aromatic plane to the other aromatic plane.

r_2: horizontal displacement between the aromatic planes.

θ: dihedral angle between the aromatic planes. The range is from 0° to 90°.

$$E_{\text{packing}} = \sum C \cdot f(r_1) \cdot f(r_2) \cdot f(r_3) \cdot f(\theta) \qquad (8)$$

where

$$f(x) = \exp[-A \cdot \exp[-B \cdot (x - x_0) \cdot (x - x_1)]] \qquad (9)$$

$f(x)$ is a normalized continuous function which shape is similar to a step function. In Eq. (9), A and B are constants with values 2.0, while x_0 and x_1 represent the boundaries for the variable x which values are shown in Table II. The coefficient C for the packing correction is −3.0 kcal/mol, in the range of Jurecka's estimation.[46]

In the present VSGB 2.0 model, we use the same parameters for all of the π–π interactions enumerated above. Preliminary investigation suggests that there is not a large sensitivity to the specific value of, for example, the value of C, and that the various functional group interactions yield reasonable results with the single value of −3.0 kcal/mol specified above. However, this needs to be looked at in more detail in future works, as one would expect some systematic differences, particularly when the chemistries are significantly different, as in the case of two guanidinium groups interacting as compared to two phenyl rings.

Self-contact correction

It is common in proteins to find side chains of Asn, Gln, Ser, and Thr interacting with their own backbone nitrogen or oxygen atoms. Such an interaction depends on both the side chain conformation and the secondary structure, so that it is more complicated than a normal hydrogen bond. It was found by Pal *et al.* that self-contact interactions have specific roles in protein local environments, while most of them have tertiary interactions,

Table II

Geometry Parameters for π-π Packing Correction

	r_1 (Å)	r_2 (Å)	r_3 (Å)	θ (°)
x_0	3.5	0.0	0.0	0.0 (θ < 45°), 80.0 (θ ≥ 45°)
x_1	6.5	3.5	5.0	20.0 (θ < 45°), 90.0 (θ ≥ 45°)

Table III
Parameters of Self-Contact Correction

	Asn	Gln	Ser, Thr
A (kcal/mol)	−4.7	−2.5	−4.8
B	3.2	4.0	3.2
C	1.0	0.0	0.8
r_0 (Å)	3.2	3.8	3.2

saturated by hydrogen bonds from side chain-backbone and side chain-solvent.[49] Therefore, self-contact interactions are considered as a special case of hydrogen bonding in the VSGB 2.0 model.

The correction is represented as a sum of Gaussian functions dependent on the distance r between two heavy atoms with self-contact interactions [Eq. (10)]. One atom is the polar atom from the side chain of Asn, Gln, Ser, or Thr, while the other one is the backbone nitrogen or oxygen atom in the same residue. On the basis of our statistical study with high resolution PDB structures, Asn, Ser, and Thr are most likely to form self-contact interactions, so that the correction is stronger for these three amino acid residues. The value of r_0 was taken from the most populated distance for the corresponding amino acid residue. Ser and Thr are treated with identical parameters due to their similarity in side chain hydroxyl group. It is worth mentioning here that the coefficient C is used to rescale the correction if the amino acid residue is in a regular secondary structure, since the self-contact interaction disturbs the hydrogen bonding network in a helix or sheet. Parameters for the self-contact correction are presented in Table III.

$$E_{\text{self-contact}} = \sum A \cdot \exp[-B \cdot (r - r_0)^2] \cdot C \qquad (10)$$

Hydrophobic term

The hydrophobic term was introduced by us in a previous version of our energy model.[14] The original term was taken from a scoring function employed in docking calculations, ChemScore.[50] The hydrophobic term rewards contacts between nonpolar heavy atoms and stabilizes hydrophobic contacts. As the effect of hydrophobic term is much smaller in one side chain than in a loop, in the VSGB 2.0 model we replaced the linear function with a polynomial and refit the parameters based on the loop predictions at lengths of 11–13 residues.

$$E_{\text{hydrophobic}} = \text{coeff} \cdot \sum_{ij} E^{ij}_{\text{hydrophobic}} \qquad (11)$$

$$E^{ij}_{\text{hydrophobic}} = \begin{cases} 0.0 & (1 \le \text{scale}) \\ 0.25 \cdot \text{scale}^3 - 0.75 \cdot \text{scale} & \\ \quad +0.5 (-1.0 < \text{scale} < 1.0) & \\ 1.0 & (\text{scale} \le -1.0) \end{cases} \qquad (12)$$

where

$$\text{scale} = 2.0 \cdot (r_{ij} - r_i^{\text{vdw}} - r_j^{\text{vdw}} - 2.0)/3.0 \qquad (13)$$

The coefficient in Eq. (11) was fit by a line search algorithm and the optimal value we obtained is −0.30. (Comparison of the linear function and polynomial is shown in Figure S1 in Supporting Information.)

The hydrophobic term is intended to model the interaction of hydrophobic surfaces presented by various protein and ligand groups with water; when these groups make contacts (in the extreme case forming the hydrophobic core of the protein), unfavorable interactions with water are eliminated, and this effect drives hydrophobic packing. The atom–atom contact term described above represents an alternative to models which attempt to directly compute cavitation and van der Waals interactions of the solute atoms with the solvent. Our empirical investigations, initially described in Reference 14 but continued over the past several years with extensive experiments on large data sets, suggest that the model of Eqs. (12) and (13) provides superior predictions as compared to more standard approaches penalizing exposed hydrophobic surface area, which are generally derived from small molecules in bulk solution, a very different situation from hydrophobic groups embedded in an active site cavity, or otherwise positioned in confined spaces within a protein environment. Therefore, in the VSGB 2.0 model, we have eliminated the G_{cav} and G_{vdw} terms discussed above in Eqs. (2) and (3), and replaced them with the hydrophobic term presented in this section. The quality of results for long loop prediction, which involve no further adjustment of the energy model, will serve as an unbiased test of the validity of this approach.

Optimization of the VSGB 2.0 model

In the development of the VSGB 2.0 model, we fit the methods and parameters to 2239 single side chain and 100 11–13 residue loop predictions. The goal of optimizing the parameters in the VSGB 2.0 model was not only to improve the results of single side chain or loop prediction, but also to give a more accurate physical description that is transferable to tackle practical problems such as homology model refinement. As we mention in the introduction, fitting to large high-quality experimental data sets helps to capture the correct physics in proteins and reduces the risks of overfitting.

The VSGB 2.0 model was optimized by carrying out single side chain prediction. As mentioned previously, the prediction of a single protein side chain can be determined by the lowest energy conformation either without minimization (single point) or with minimization. We tried to maximize the performance of the VSGB 2.0 model on both single point and minimization selection during the procedure to optimize our model. Thus the

Table IV
Summary of Single Side Chain Prediction Results for 11 Polar or Charged Residues

		VSGB 2.0		VSGB 1.0		Zhu et al.[b]	
Residue type	Number of cases	SP (%)[a]	MIN (%)[a]	SP (%)[a]	MIN (%)[a]	Number of cases	SP (%)[a]
Arg	144	85.4	84.0	83.3	82.6	171	77.8
Asn	252	92.5	91.7	90.5	88.5	237	85.7
Asp	293	94.9	94.9	94.2	92.5	254	91.7
Cys	92	100.0	100.0	100.0	100.0	49	93.9
Gln	161	88.8	83.2	85.7	77.6	161	85.7
Glu	152	88.8	86.2	88.2	84.9	193	79.3
His	83	95.2	95.2	91.6	91.6	132	86.4
Lys	121	91.7	90.1	95.9	88.4	198	76.8
Thr	404	95.8	94.3	94.8	92.6	302	92.4
Tyr	221	99.5	99.1	99.1	98.6	184	89.7
Ser	316	88.9	88.0	88.3	86.1	297	79.1
All	2239	93.0	91.6	92.0	89.6	2178	85.0

All the single side chain predictions were performed with ICDA prepared structures.
[a]A "successful" prediction is defined as one where the heavy atom RMSD <1.5 Å to the native side chain conformation. The percentages reported in here are the ratio of the number of accurate predictions to the total number of predictions. "SP" stands for the single point energy evaluation; "MIN" stands for the minimized energy evaluation.
[b]This method uses the original single side chain prediction algorithm, a different data set and the VSGB 1.0 model.[15]

parameter optimization is based on the ability to select the lowest energy structure, as well as the ability to make a good approximation of the energy funnel. Except the hydrophobic term which was fit to 11–13 residue loop predictions with decoys, all the parameters of the solvent model and physics-based corrections were optimized by a script based on a Monte-Carlo (MC) algorithm. The MC script generates parameters, recalculates the single point energy for each side chain candidate in the pool, and determines the prediction by the lowest energy candidate. We selected two sets of optimal parameters with the lowest average RMSDs suggested by the MC script, and performed the actual side chain predictions to determine the final parameters as the ones leading to the highest success rate with and without minimization.

RESULTS

Single side chain and loop (11–13 residues) prediction

The results of single side chain prediction obtained by the VSGB 2.0 model are shown in Table IV in comparison to our previous energy model (herein referred to as VSGB 1.0), which employs the OPLS-AA energy function with the original variable dielectric solvent model and the earlier implementation of the hydrophobic term. Eleven polar or charged amino acids were included for the fitting to compare the results from previous work. The overall accuracy of the 2239 single side chain predictions is as high as 93.0% with single point energy evaluation (SP) and 91.6% with minimized energy evaluation (MIN) respectively. Generally, the VSGB 2.0 model improves the prediction accuracy for 1.0% (SP) and 2.0% (MIN). The most significant improvements come from Gln and Asn, at 3.1% (SP), 5.6% (MIN) and 2.0% (SP), 3.2% (MIN). Another remarkable improvement advanced by the VSGB 2.0 model is to reduce the large differences in accuracy between single point energy evaluation and minimized energy evaluation. With the VSGB 1.0 model, the difference in success rate is 7.5% for Lys and 8.1% for Gln; with the VSGB 2.0 model, it is reduced to 1.6% and 5.6%, respectively. Given the high success rate of the VSGB 1.0 model, such an improvement is non-trivial. It is evident that the VSGB 2.0 model improves not only the energy score but also the gradient of the energy score.

The hydrophobic term was optimized with loop predictions of 11–13 residue length, results of which are presented in Table V. Although the change from linear form to polynomial form did not significantly improve the accuracy, we still believe that the basic physical arguments can easily justify our reformulation, which here allows for the introduction of a continuous first derivative and the uniform application of the term throughout all of space. In particular, a continuous first derivative would take the force into consideration and allow a smoother minimization. Likewise, to our knowledge, there is no experimental evidence to suggest that the hydrophobic forces exerted between crystal mates should be any different than those hydrophobic forces exerted within the unit cell. Thus, the parsimony principle seems to motivate the change, even in the absence of strong data suggesting the polynomial form adopted here is fundamentally more accurate than the earlier linear ramping function.

It should be noted that the average RMSD of ~0.8 Å reported in Table V arises from selection of loops from a suite of decoys generated with an older energy model from Reference 14, as opposed to full scale optimization of loop prediction with the current energy model and sampling algorithms. In view of the results to be reported below for longer loops, it is likely that improved results would be obtained for 11–13 residues loops with the latter, more rigorous sampling protocol.

Super long loop prediction

To validate the effectiveness of the VSGB 2.0 model, a super long loop set was used to test the accuracy of the

Table V
Summary of 11–13 Residue Loop Prediction Results with Different Forms of Hydrophobic Term

Hydrophobic term	Included the crystal environment	Average backbone RMSD (Å)	Accuracy (%)
Linear	No	0.81	91.0
Polynomial	No	0.83	91.0
Polynomial	Yes	0.85	91.0

Table VI
Summary of Super Long Loop Prediction Results

		VSGB 2.0 with ICDA				VSGB 1.0 without ICDA			
Loop length	Number of cases	Median backbone RMSD (Å)	Average backbone RMSD (Å)	Average side chain RMSD (Å)	Percentage of cases with RMSD <2 Å	Median backbone RMSD (Å)	Average backbone RMSD (Å)	Average side chain RMSD (Å)	Percentage of cases with RMSD <2 Å
14	36	0.38	0.51	1.67	100.0	0.67	1.19	2.51	91.7
15	30	0.54	0.63	1.85	100.0	0.75	1.55	3.07	73.3
16	14	0.43	0.70	1.85	100.0	0.80	1.43	3.20	78.6
17	9	0.57	0.62	1.84	100.0	1.92	2.30	4.25	66.7
18	16	0.60	0.80	1.78	100.0	3.45	4.18	5.59	37.5
19	7	1.60	1.41	3.46	100.0	1.31	2.65	3.87	57.1
20	3	1.68	1.59	2.88	100.0	1.12	1.43	2.71	66.7
All	115	0.52	0.69	1.91	100.0	1.04	1.89	3.37	73.0

RMSD calculation: A predicted structure is superimposed to the native protein structure excluding the target loop. Backbone RMSDs are calculated with N, C_a, and C atoms; side chain RMSDs are calculated with heavy atoms.

energy function described herein. A summary of the results is shown in Table VI and Figure 3, whereas the detailed results are given in Table S4 in the Supporting Information.

Testing on the same data set, we are able to make a direct comparison of the relative performance of the VSGB 2.0 model and the VSGB 1.0 model. Using the VSGB 2.0 model, 100.0% of loop predictions have backbone RMSDs below 2.0 Å from the native conformations, where only 73.0% of the predictions made with the VSGB 1.0 model would be similarly accurate. Furthermore, the better performance of the VSGB 2.0 model was independent of any particular chosen success criteria, as depicted by Figure 4, where the percentage of predicted

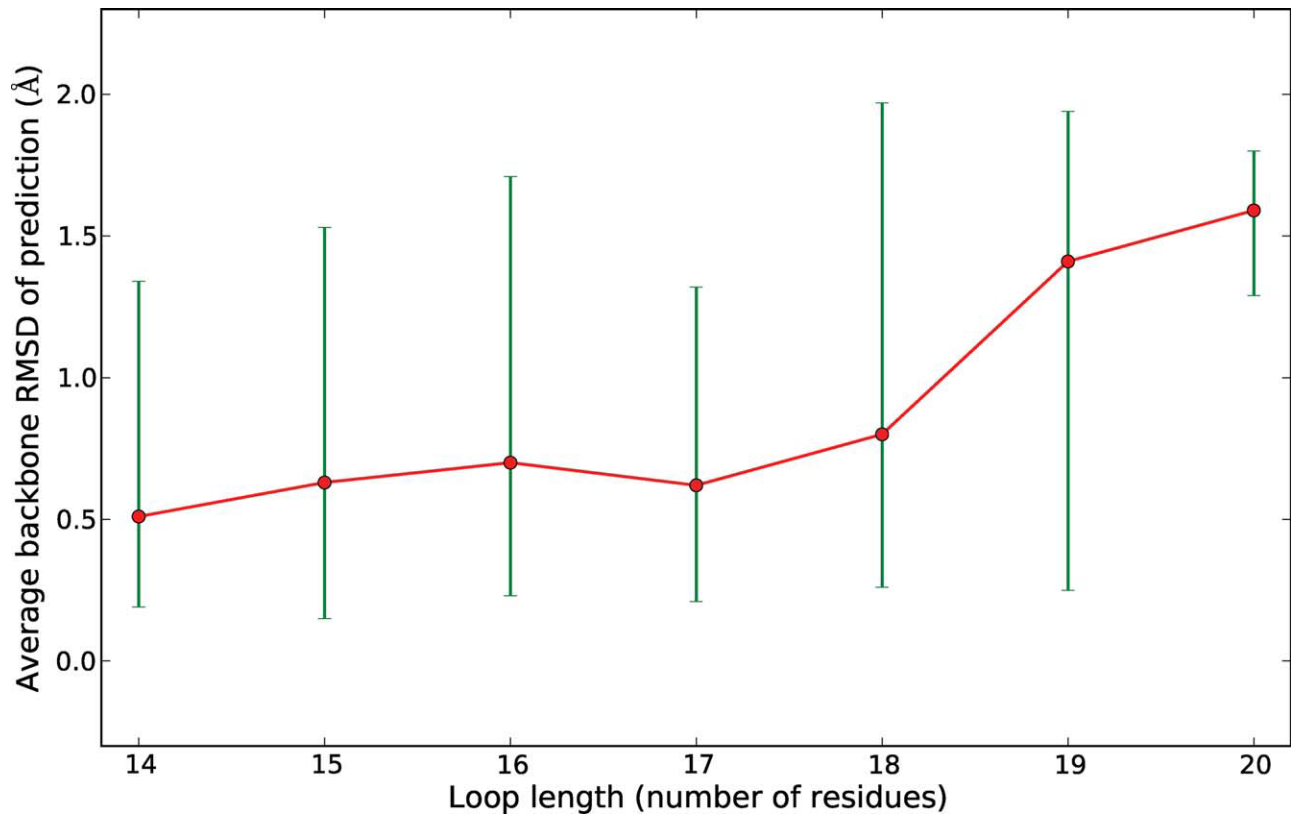

Figure 3
Accuracy of super long loop predictions. Red dots represent the average backbone RMSDs whereas green lines represent the ranges of the backbone RMSD. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

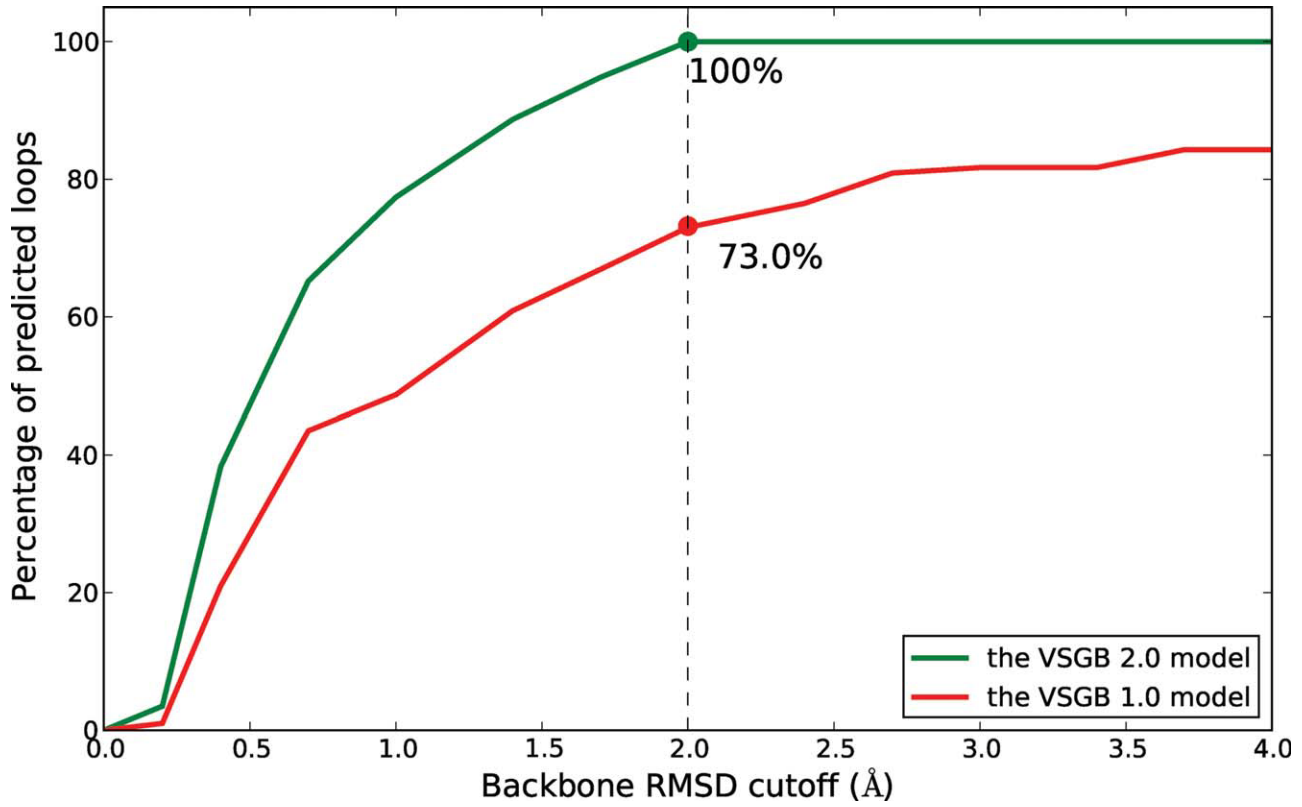

Figure 4

Comparison of loop predictions with the VSGB 2.0 model and the VSGB 1.0 model. The *y*-axis shows the percentage of predicted loops within the RMSD cutoff on the *x*-axis. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

loops with RMSD lower than a cutoff is plotted as a function of the given cutoff. The prediction difficulty often increases dramatically with loop length; however, our results with the VSGB 2.0 model reflect average backbone RMSDs in a narrow range from 0.51 to 0.80 Å for 14-18 residue loops. Even though it is extremely challenging to predict loops longer than 18 residues, the VSGB 2.0 model still displays high accuracy with average backbone RMSDs of only 1.41 Å and 1.59 Å for 19 and 20 residue loops. Apart from the backbone, the VSGB 2.0 model also significantly reduces the RMSDs for side chains in the loops. For example, for 18 residue loops, we obtained average backbone/side chain RMSDs of prediction are 0.80/1.78 Å with the VSGB 2.0 model compared to 4.18/5.59 Å with the VSGB 1.0 model.

For super long loop prediction, our hierarchical algorithm is able to enrich the native-like conformations along the increasing fixed stages. The analysis of the average RMSDs at each stage shows that the VSGB 2.0 model not only provides an accurate score for the final prediction, but also improves the accuracy for each prediction stage (see Figure 5), allowing higher percentage of native-like confirmations and faster convergence towards the global minimum. For example, both 14 and 15 residue loop predictions reach average RMSD below 2.0 Å at

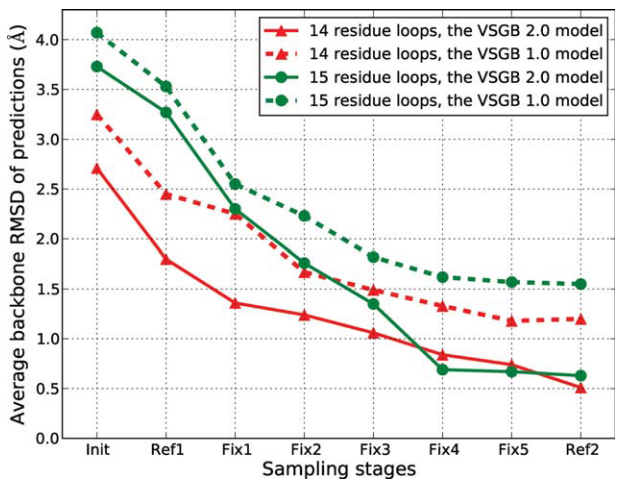

Figure 5

Average backbone RMSD of the predicted loops (14 and 15 residues) at each sampling stage. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Table VII
Results of Super Long Loop Prediction in Inexact Environments

				Exact environment			Inexact environment		
PDBID	Loop start	Loop length	Number of surrounding residues	Starting structure RMSD (Å)	Backbone RMSD (Å)	Side chain RMSD (Å)	Starting structure RMSD (Å)	Backbone RMSD (Å)	Side chain RMSD (Å)
1E6U	A274	14	28	0.00	0.27	0.85	3.37	0.31	1.07
1ZEQ	X53	14	61	0.00	0.21	2.10	3.52	0.28	2.08
2BWR	A269	14	69	0.00	0.31	1.70	3.35	0.58	2.64
3BY9	A205	14	59	0.00	0.32	0.81	3.14	0.41	0.83
3EHR	A95	14	83	0.00	0.51	3.40	4.09	1.30	4.08
1QAZ	A298	15	51	0.00	0.99	3.00	3.46	2.18	4.52
1RA0	A283	15	36	0.00	0.30	2.13	6.51	1.24	2.38
1RA0	A361	15	81	0.00	0.52	1.66	5.10	0.52	1.70
3CSS	A95	15	52	0.00	0.52	1.39	3.06	1.98	3.62
1WM3	A67	16	49	0.00	0.23	1.24	3.08	0.26	1.63
Average RMSD					0.42	1.83		0.91	2.47

The RMSD of a starting structure was calculated as the backbone RMSD of the target loop in the starting structure, but the prediction of the loop was actually built from scratch.

least one stage earlier with the VSGB 2.0 model. This also implies that fewer sampling stages are required with the VSGB 2.0 model, making the loop predictions more cost efficient.

Super long loop predictions in inexact environments

A subset of 10 cases ranging from 14 to 16 residues (all with reliable, well resolved electron densities for the target loop as well as surrounding side chains) were used to test loop predictions in inexact protein environments. An inexact environment was here created by replacing the target loop by a non-native loop conformation (RMSD $>$ 3.0 Å) and minimizing the new structure with surrounding side chains. The loop prediction in inexact environments was performed with method SLLP-SS, as was described earlier in Methods section. The test results are shown in Table VII.

Compared to the starting structures, all the 10 cases display improvements in RMSDs after the loop reconstruction. This confirms that the VSGB 2.0 model together with the augmented sampling method presented herein is able to improve the quality of models, suggesting a high potential to succeed in refining comparative models. Although starting from the inexact environment degrades the overall performance by 0.49 Å in the average backbone RMSD, many of the cases actually reach similar accuracy as starting from the exact environment, such as a 14-residue loop in 3BY9 (A205-218) and a 15-residue loop in 1RA0 (A361-375). Furthermore, the average error is still a highly satisfactory 0.91 Å, and the maximum RMSD remains less than 2.5 Å. Systematic exploration of the increased error induced by the inexact environment requires investigation of a much larger data set; we plan to pursue this in future work.

DISCUSSION

The single side chain prediction results represent a major advance as compared to the accuracy levels reported in Reference 15. A great deal of the improvement is due to the use of a better side chain data set (in which all of the atoms in the side chain are reliably located by from the electron density obtained from the crystallographic observations), as opposed to improvements in the energy model. These results demonstrate that data set quality is vital in both development and assessment of molecular models. Problems with the data set will invariably produce misleading (and in some cases highly misleading) conclusions with regard to the performance of the model.

The improvements in accuracy attributable to the new model are relatively small, but significant, particularly as they are clustered in the polar and charged side chains. Furthermore the vast majority of cases, where predictions failed, have quite small energy gaps, implying that the impact on loop prediction of such errors would be relatively minimal. However, we note that elimination of erroneous predictions is not the only benefit to adding a new, physically important term to the energy model. In many cases, the VSGB 1.0 model may have given a correct prediction for a particular side chain, but not properly evaluated the energy of the side chain conformation as compared to possible alternatives. As an example, consider the π–π packing term, which rewards stacking of aromatic residues. In many cases, these residues are in the hydrophobic core of the protein, and reprediction of the side chain conformation of the residue can have only one outcome due to steric considerations; the problem is like a jigsaw puzzle in which the "piece" can fit back into the puzzle in only one "conformation". However, the energy gained from forming the entire core in the specific fashion that enables many stacking interactions to

Table VIII
Cases that Require ICDA Preparation to Achieve Accurate Prediction

			VSGB 2.0, with ICDA		VSGB 1.0, with ICDA		VSGB 1.0, without ICDA	
PDBID	Loop start	Loop length	EGAP (kcal/mol)	Backbone RMSD (Å)	EGAP (kcal/mol)	Backbone RMSD (Å)	EGAP (kcal/mol)	Backbone RMSD (Å)
1RA0	A283	15	7	0.30	0	1.03	−39	2.78
2PKF	A26	15	0	0.65	10	0.47	−8	2.34
2BG1	A708	16	−3	0.40	5	0.59	−4	2.15
2PYW	A321	16	−1	0.99	−5	0.72	−37	2.57
2HDW	A131	17	10	0.60	−3	0.42	−126	2.22
3CUZ	A384	18	4	0.82	−4	0.70	−22	3.45
3EH1	A813	18	11	1.60	5	1.54	−22	2.84
3GGQ	A550	18	8	0.37	8	0.53	−26	5.12

EGAP: energy of the prediction–energy of the minimized native structure.

be formed may be underestimated by an energy model that does not reward stacking interactions. This would have consequences not only for loop prediction (where, as we show below, incorrect loop conformations often fail to make key stacking interactions present in the native conformation) but also for refining a homology model so that it has the optimal interlocking pattern of side chain interactions. Optimization of the π–π stacking and other terms to improve side chain prediction uses a small fraction of "sensitive" side chain cases to detect problems in the energy model and optimize the terms that improve these cases. The success of this strategy is manifest in the results reported above for long loop prediction, where very large improvements in average backbone RMSD and associated average side chain RMSD are seen uniformly for the 14–20 residue loop database. The VSGB 2.0 model fixes a number of cases, which previously had substantial energy errors leading to inaccurate loop RMSDs.

Impact of systematic application of protonation state assignment methodology

When making either single side chain predictions or loop predictions, it is extremely difficult to achieve accurate predictions if one or more ionizable side chains are represented in the incorrect protonation state. As a simple example, some Asp and Glu residues form carboxylate "dimers" with neighboring Asp or Glu residues (without any nearby metal ions); in the crystal structure, oxygen atoms from the pair of carboxylates belonging to each side chain are observed to be within hydrogen bonding distance (~3.0 Å). This type of structure implies that at least one of the carboxylates must be protonated. This costs free energy with regard to the standard protonation state of a carboxylate in solution, but is clearly necessary to avoid large repulsive interactions between charged oxygen atoms that would otherwise occur (vs. forming a strong hydrogen bond). If the unprotonated forms are used, the "dimer" structure will never be predicted as lowest in energy. Loop prediction is more subtle, but side chains in loops do form salt bridges and hydrogen bonds which are dependent upon protonation state. If these interactions cannot be formed due to failure to incorporate a nonstandard (but accessible) protonation state, the energy of the native loop conformation may fail to be competitive with incorrect alternatives.

All of the loop and side chain prediction results shown below have been generated by running an automated protonation state assignment program, the ICDA, based on methods described in Reference 26, on the entire protein of each member of the test set. Since the publication of Reference 26, we have put significant efforts into improving the ICDA by running a large number of test cases and examining them visually to make sure that obvious errors are eliminated. However, the data sets in the present article have been treated in an automated fashion. The effects on accuracy of failing to run the ICDA has been examined for a selected set of loops, those which yielded large energy errors in previous work which did not employ systematic ICDA preparation (but also used a different energy model). By rerunning these cases with ICDA preparation and the VSGB 1.0 model, it is possible to identify a subset of cases where ICDA preparation is essential to achieving accurate results. At least eight of 115 cases (7.0%) require ICDA preparation to achieve accurate structure prediction (see Table VIII).

In this study, we are able to assign protonation states with high accuracy due to having the native structure available. In the context of realistic homology model refinement, the native structure would not be known in advance. Achieving the correct protonation states would then require sampling differing protonation states on the fly during the simulations, and/or running a number of alternative protonation states as a part of the iterative refinement algorithm protocol. A number of publications in the literature describe successful on-the-fly protonation state sampling algorithms,[51,52] which could be adapted to our refinement algorithms. It is likely though that a significant effort will have to be put into making

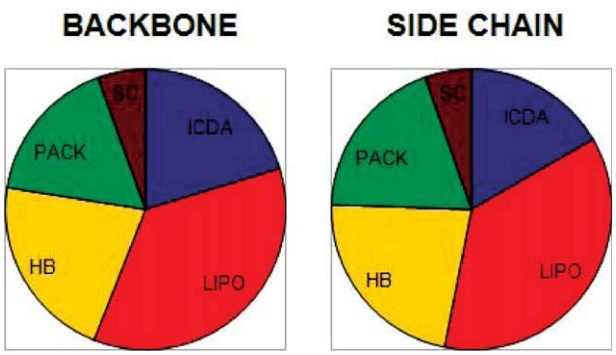

Figure 6

Relative importance of components in the VSGB 2.0 model. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

sure that these methods are sufficiently accurate to reliably achieve refinement objectives. The present work should be viewed as proof of concept, demonstrating that if accurate protonation states can be assigned, significant improvement in blind structural prediction efforts will result. Running tests without proper protonation states, in contrast, would make it impossible to distinguish intrinsic failures of the energy model from failures to assign the correct protonation state.

Importance of each component in the VSGB 2.0 model

To investigate the importance of each component in the VSGB 2.0 model, we ran a number of tests over the subset of 14-residue loops with one component removed from the model in each test. These tests were carried out as loop predictions with the same setting mentioned before, and the results show that all the incomplete models have decreased performance giving higher average backbone and side chain RMSDs. The average RMSD changes compared to the VSGB 2.0 model were normalized and projected to pie charts in Figure 6, so that we can get a sense about the relative importance of each component in our model. Summary of the test results are provided in Table S5 in the Supporting Information.

The pie charts in Figure 6 show that the most important component is the hydrophobic term (LIPO), which has a significant impact on both the backbone and side chain accuracy. Preparation of protonation states (ICDA), the hydrogen bonding correction (HB) and the π–π packing correction (PACK) display close importance, but the last two components have slightly bigger importance for the side chain than the backbone. This agrees to the stronger accumulation of hydrogen bonding and π–π interactions in side chains than in backbone. The self-contact correction (SC), only applicable to the side chains of few amino acid residues, has the smallest importance as expected.

A better description of protein energy landscape

As a mini folding problem, super long loop prediction at high resolution demands an energy model with highly precise physical description of the loop in question as well as of the environment. However, an inaccurate energy model with missing physics usually fails to discriminate the native conformations from the non-native ones, leading to wasted sampling effort in the false global minimum which could be far away from the true one. This explains why sometimes a more intensive sampling effort leads to a poorer prediction. Through our analysis of such mispredictions, the missing physics in our previous energy models appeared to be stemming from inac-

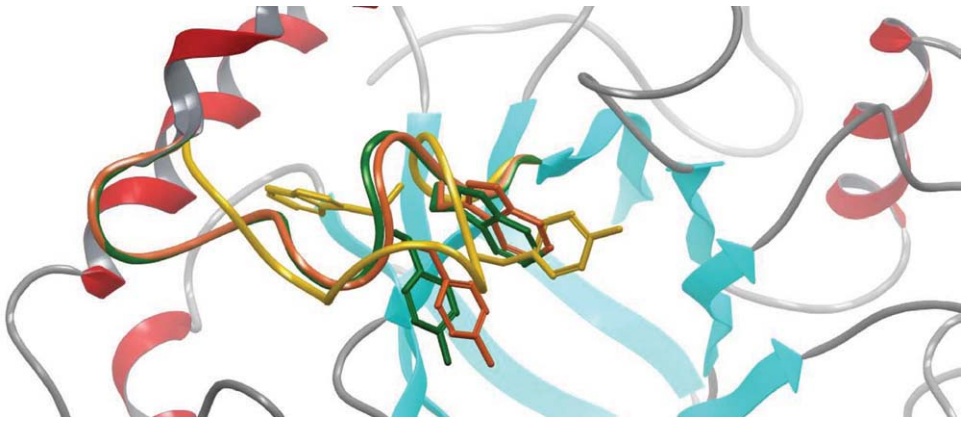

Figure 7

The π–π packing correction improves a 14 residue loop prediction (PDBID: 2C0H. Loop A40-53). Compared to the native (green), the prediction with the VSGB 2.0 model (orange, backbone RMSD = 0.55 Å) forms the correct T-stacked side chains with Trp and Tyr; the prediction with the VSGB 1.0 model (yellow, backbone RMSD = 5.96 Å) have these two side chains far away.

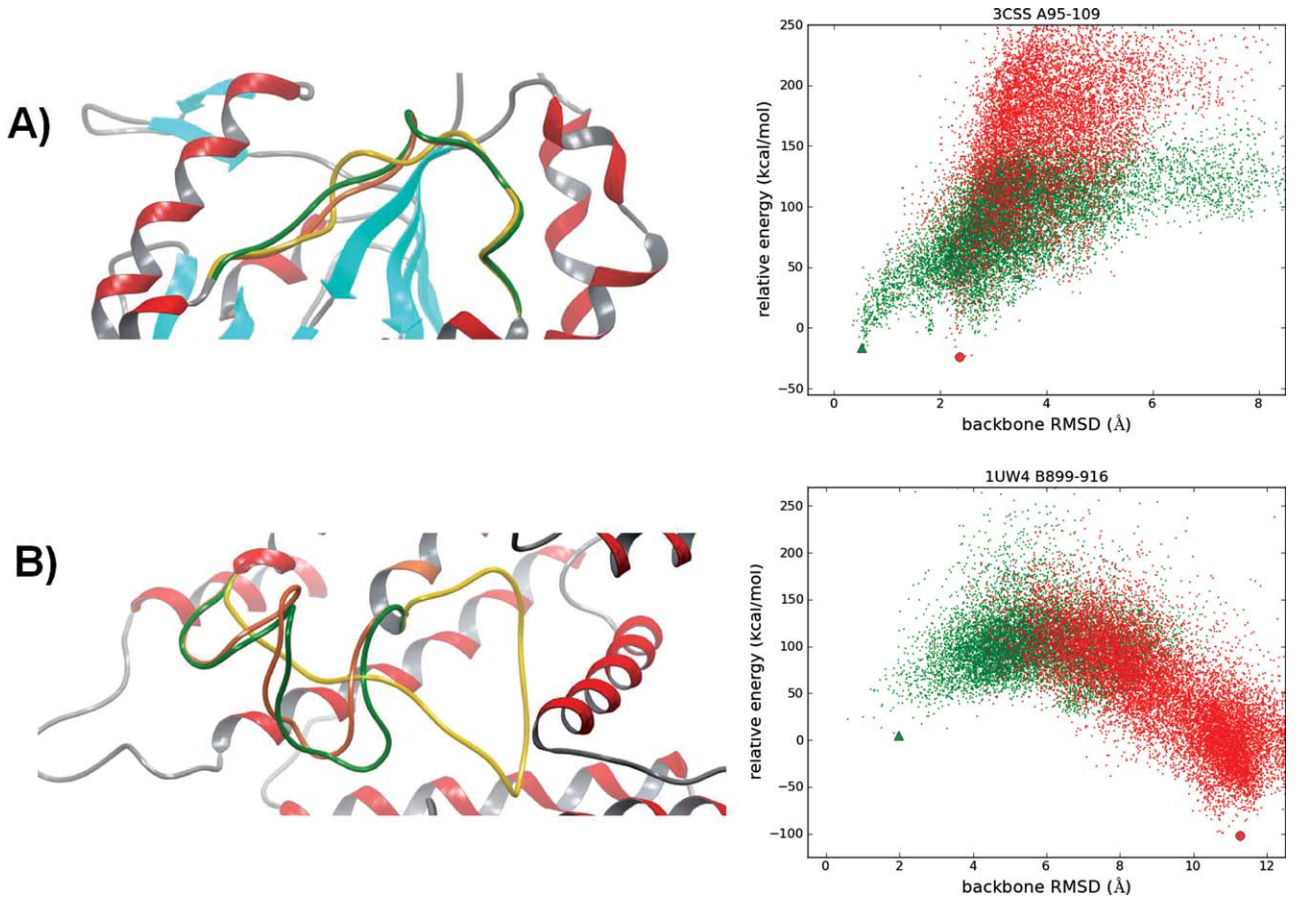

Figure 8

Loop prediction results of two cases with serious energy error in the VSGB 1.0 model. The protein structures are shown on the left: the native structures are green, the predicted structures by the VSGB 2.0 model are orange, and the predicted structures by the VSGB 1.0 model are yellow. The plots of relative energies against RMSDs are on the right. All the conformations generated in the loop prediction are included; the green dots represent the VSGB 2.0 model, and the red dots represent the VSG 1.0 model. The predictions are marked as a large green triangle and a large red dot. **A**: PDBID: 3CCS. Loop A95-109: backbone RMSD = 0.52 Å, EGAP = −17 kcal/mol (the VSGB 2.0 model); backbone RMSD = 2.36 Å, EGAP = −24 kcal/mol (the VSGB 1.0 model). **B**: PDBID: 1UW4. Loop B899–916: backbone RMSD = 1.97 Å, EGAP = 5 kcal/mol (the VSGB 2.0 model); backbone RMSD = 11.27 Å, EGAP = −102 kcal/mol (the VSGB 1.0 model).

curate descriptions of electrostatic, hydrogen-bonding, hydrophobic, and π–π interactions.

Using the optimized variable dielectric solvent treatment and the additional physics-based corrections, the VSGB 2.0 model is likely to provide a better, more complete physical description for protein high resolution modeling. The addition of corrections, especially to the hydrogen bonding, π–π interactions and hydrophobic interactions, compensates the effects that were incompletely described by the electrostatics, van der Waals, and nonpolar interactions in the VSGB 1.0 model (see Figure 7 as an example). As a result, the corresponding native or native-like conformations are stabilized, with respect to the competing non-native conformations, and thus are more likely to be the global minimum energy structure on the potential energy surface. In addition to the more accurate description of the global minimum energy structure, the VSGB 2.0 model also incorporates a stronger bias along the whole of the energy surface towards more native-like conformations, which could be due to the improved physical description of the overall energy surface. One indication of the greater bias of potential energy surface of the new energy function towards more native like structures is shown in Figure 5, where the VSGB 2.0 model gives lower average backbone RMSDs for each stage compared to the VSGB 1.0 model. This indication could be interpreted as preliminary evidence that the new energy function is better describing the "folding funnel" of the protein potential energy surface.[53,54] Another indication, more direct, is the consistent correlations of relative energies to the native and RMSDs for all the conformations that have been sampled during the loop prediction. Two examples are presented in Figure 8 to further demonstrate the better physical description in the VSGB 2.0 model. More plots of

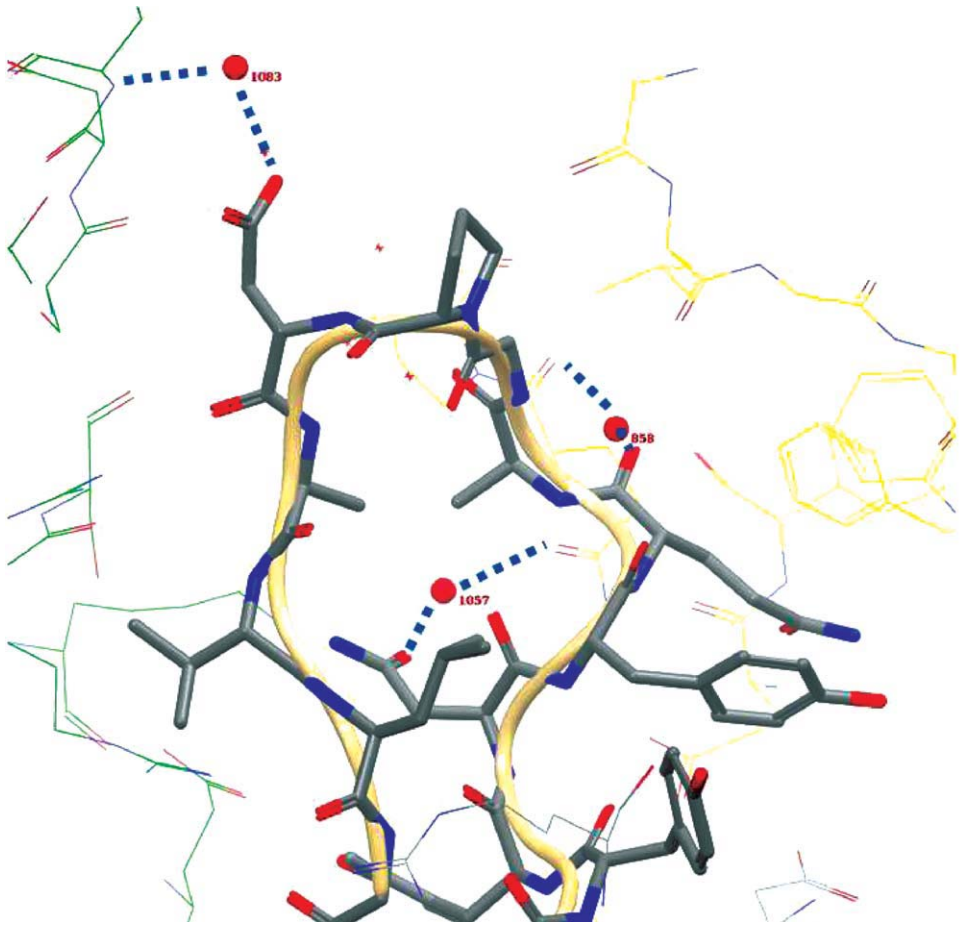

Figure 9

The 14 residue loop in a nucleotidase (PDBID: 1JP4. Loop A153-166) and bound crystal water molecules. Protein molecules from the crystal environment are shown in element color scheme with green or yellow carbons. Hydrogen bonds that connect the loop in question and the crystal environment are shown in blue dotted lines.

relative energies against RMSDs are shown in the Supporting Information.

Water between protein molecules in the crystal structure

The 14 residue loop (A153-166) in a nucleotidase (PDBID: 1JP4) is unique in our data set of super long loops: at the beginning, neither the VSGB 1.0 nor the VSGB 2.0 model alone gave reasonable predictions. (The VSGB 1.0 model: backbone RMSD = 7.26 Å, EGAP = −43 kcal/mol; the VSGB 2.0 model: backbone RMSD = 2.40 Å, EGAP = −30 kcal/mol). Considering our starting point of improving the physical description, what is the missing physics in this case? An analysis of the native structure has shown that the water molecules bridge the loop in question and the crystal environment, which could lead to the incorrect energy evaluation due to the limitations of the implicit solvent model.

The sequence of this 14 residue loop (PYYNYQAGPDAVLG) has a high fraction of non-charged residues, which in this case have strong hydrophobic interactions, unusually, predominantly with several other protein molecules in the crystal environment, as opposed to with the hydrophobic core of the protein molecule which the loop is a part of. Instead of forming the extended native conformation which contacts the neighboring proteins, all the mispredictions are packed into the protein body. However, we found that there are three bound water molecules (HOH 858, 1057, and 1083) which contribute significantly to the stability of the extended native conformation: all these water molecules connect the loop to the crystal environment through their hydrogen bonding network and consequently pin down the extended conformation (see Figure 9). Such a first-shell solvation effect is unlikely to be well represented by any implicit solvent model, and thus we could not create the correct physical environment unless these water molecules were included.

Table IX
Corrected Results of Super Long Loop Prediction with VSGB 2.0 Model

		VSGB 2.0				VSGB 2.0 with problematic cases corrected			
Loop length	Number of cases	Median backbone RMSD (Å)	Average backbone RMSD (Å)	Average side chain RMSD (Å)	Percentage of cases with RMSD <2 Å	Median backbone RMSD (Å)	Average backbone RMSD (Å)	Average side chain RMSD (Å)	Percentage of cases with RMSD <2 Å
14	36	0.38	0.51	1.67	100.0	0.38	0.61	1.89	91.7
15	30	0.54	0.63	1.85	100.0	0.56	0.71	1.89	96.7
16	14	0.43	0.70	1.85	100.0	0.43	0.86	2.03	92.3
17	9	0.57	0.62	1.84	100.0	0.57	0.62	1.84	100.0
18	16	0.60	0.80	1.78	100.0	0.60	1.06	2.18	87.5
19	7	1.60	1.41	3.46	100.0	1.60	1.67	3.66	85.7
20	3	1.68	1.59	2.88	100.0	1.68	1.59	2.88	100.0
All	115	0.52	0.69	1.91	100.0	0.53	0.82	2.08	94.8

RMSD calculation: A predicted structure is superimposed to the native protein structure excluding the target loop. Backbone RMSDs are calculated with N, C_a, and C atoms; side chain RMSDs are calculated with heavy atoms.
The problematic cases were identified by preliminary assessment of the effects of full "Fix 10" sampling.

In order to repredict this loop in the correct physical environment, we explicitly added these three water molecules and the ones that form hydrogen bonds with them (HOH 858, 911, 1057, 1083, 1108, 1145, and 1173) to our all-atom model. Hydrogen bonds between the explicit water molecules and the proteins were considered. With the presence of these water molecules and the hydrogen bonding correction applied to protein–water interactions, the prediction with the VSGB 2.0 model yields high accuracy (backbone RMSD = 0.31 Å, side chain RMSD = 0.67 Å, EGAP = −4 kcal/mol). This case study highlights the limitations of implicit solvent models, and suggests possible treatments of including or predicting crystal water positions in future works.

It should be noted that the explicit waters are required specifically in the interstitial region between protein molecules in the crystal structure. For a single protein molecule in solution, the loop in question would in fact almost certainly adopt a different conformation, since the hydrophobic region of the loop is buried in a hydrophobic region of the neighboring protein molecule in the crystal; in fact, it might well be found in the competing conformation selected without the crystal waters, in which the hydrophobic interactions of the loop are primarily intramolecular, as opposed to with the neighboring molecule in the crystal.

Possible sampling errors in the current data set, and their potential effect upon accuracy of the energy model

The simulation of 115 long loops carried out in this article requires a substantial amount of computation time. Furthermore, the lengthiest simulations (using 10 fixed stages, referred to in the text as "Fix 10") requires considerably more computational effort than the five fixed stage ("Fix 5") algorithm that was employed for most of the calculations. This is the reason that, in the data set presented above, we utilized the "Fix 10" algorithm only in cases where the "Fix 5" simulations displayed an RMSD >2.0 Å.

However, it must be recognized that such a protocol could conceivably lead to bias in the results, as we do not know whether running the full "Fix 10" protocol on all of the cases would produce additional errors. We have carried out a preliminary assessment of this problem using a variety of computational experiments, which has led to the conclusion that no more than nine of the 115 cases potentially would exhibit energy errors leading to RMSDs greater than 2.0 Å if run with the full "Fix 10" protocol. We carried out the full Fix10 calculations for these nine cases; all produced structures lower in energy than the native structure, and six yielded RMSDs >2.0 Å, with the range of errors observed falling between 2.0 and 4.5 Å (see Table S6 in Supporting Information). If these results are substituted for the previous "Fix 5" results and then used to recalculate the average and median backbone and side chain RMSDs for the entire data set, the results obtained are given below in Table IX. The changes in the statistics are modest and most visible for the loops longer than 18 residues, and it is possible that some of the problems are due to residual protonation state errors, an issue that requires further investigation. We intend to perform a full study of the entire data set at the "Fix 10" (or higher) level, using a uniform sampling protocol, and report the results in a future publication.

CONCLUSION

In this study, we have described a new energy model (VSGB 2.0) that contains an optimized solvent model and physics-based correction terms. The VSGB 2.0 model was fit to a large database of protein single side chain and loop (11–13 residues) prediction and validated by a large set of super long loop predictions. It was shown that the VSGB 2.0 model, combined with the systematic protonation state assignment, improved the accuracy of

super long loop predictions by 27.0% compared to our previous energy model (VSGB 1.0). A series of extensive analysis shows that the VSGB 2.0 model not only improved the results of single side chain and loop predictions, but also provides a better physical description for high resolution protein structure modeling. Further tests will include receptor-ligand docking, longer loop prediction, loop-helix-loop prediction, and applications on a variety of proteins, such as kinases, G protein-coupled receptors, and cytochrome P450s.

ACKNOWLEDGMENTS

The authors thank Dr. Tyler Day, Dr. Wolfgang Damm, Dr. Thomas Hughes, and Severin Schneebeli for their helpful discussions. R. Friesner has a significant financial STAKE in Schrödinger, Inc., is a consultant to Schrödinger, Inc., and is on the Scientific Advisory Board of Schrödinger, Inc.

REFERENCES

1. Patard L, Stoven V, Gharib B, Bontems F, Lallemand JY, DeReggi M. What function for human lithostathine? Structural investigations by three-dimensional structure modeling and high-resolution nmr spectroscopy. Protein Eng 1996;9:949–957.
2. Evers A, Klabunde T. Structure-based drug discovery using gpcr homology modeling: successful virtual screening for antagonists of the alpha1a adrenergic receptor. J Med Chem 2005;48:1088–1097.
3. Muegge I, Enyedy IJ. Virtual screening for kinase targets. Curr Med Chem 2004;11:693–707.
4. Sander C, Schneider R. Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins 1991;9:56–68.
5. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 1993;234:779–815.
6. Schwede T, Kopp J, Guex N, Peitsch MC. Swiss-model: an automated protein homology-modeling server. Nucleic Acids Res 2003;31:3381–3385.
7. Rockey WM, Elcock AH. Structure selection for protein kinase docking and virtual screening: homology models or crystal structures? Curr Protein Pept Sci 2006;7:437–457.
8. Sousa SF, Fernandes PA, Ramos MJ. Protein-ligand docking: current status and future challenges. Proteins 2006;65:15–26.
9. Tannor DJ, Marten B, Murphy R, Friesner RA, Sitkoff D, Nicholls A, Ringnalda M, Goddard WA, Honig B. Accurate first principles calculation of molecular charge distributions and solvation energies from ab initio quantum mechanics and continuum dielectric theory. J Am Chem Soc 1994;116:11875–11882.
10. Ghosh A, Rapp CS, Friesner RA. Generalized born model based on a surface integral formulation. J Phys Chem B 1998;102: 10983–10990.
11. Gallicchio E, Levy RM. Agbnp: an analytic implicit solvent model suitable for molecular dynamics simulations and high-resolution modeling. J Comput Chem 2004;25:479–499.
12. Jacobson MP, Kaminski GA, Friesner RA, Rapp CS. Force field validation using protein side chain prediction. J Phys Chem B 2002; 106:11673–11680.
13. Jacobson MP, Pincus DL, Rapp CS, Day TJF, Honig B, Shaw DE, Friesner RA. A hierarchical approach to all-atom protein loop prediction. Proteins 2004;55:351–367.
14. Zhu K, Pincus DL, Zhao SW, Friesner RA. Long loop prediction using the protein local optimization program. Proteins 2006;65:438–452.
15. Zhu K, Shirts MR, Friesner RA. Improved methods for side chain and loop predictions via the protein local optimization program: variable dielectric model for implicitly improving the treatment of polarization effects. J Chem Theory Comput 2007;3:2108–2119.
16. Felts AK, Gallicchio E, Chekmarev D, Paris KA, Friesner RA, Levy RM. Prediction of protein loop conformations using the agbnp implicit solvent model and torsion angle sampling. J Chem Theory Comput 2008;4:855–868.
17. Zhao SW, Zhu K, Li JN, Friesner RA. Progress in super long loop prediction. DOI: 10.1002/prot.23129.
18. Fiser A, Do RKG, Sali A. Modeling of loops in protein structures. Protein Sci 2000;9:1753–1773.
19. Galaktionov S, Nikiforovich GV, Marshall GR. Ab initio modeling of small, medium, and large loops in proteins. Pept Sci 2001;60:153–168.
20. Im WP, Lee MS, Brooks CL. Generalized born model with a simple smoothing function. J Comput Chem 2003;24:1691–1702.
21. Feig M, Onufriev A, Lee MS, Im W, Case DA, Brooks CL. Performance comparison of generalized born and poisson methods in the calculation of electrostatic solvation energies for protein structures. J Comput Chem 2004;25:265–284.
22. Fan H, Mark AE, Zhu J, Honig B. Comparative study of generalized born models: protein dynamics. Proc Natl Acad Sci USA 2005;102:6760–6764.
23. Geney R, Layten M, Gomperts R, Hornak V, Simmerling C. Investigation of salt bridge stability in a generalized born solvent model. J Chem Theory Comput 2006;2:115–127.
24. Zhou RH. Free energy landscape of protein folding in water: explicit vs. implicit solvent. Proteins 2003;53:148–161.
25. Zhang LY, Gallicchio E, Friesner RA, Levy RM. Solvent models for protein-ligand binding: comparison of implicit solvent poisson and surface generalized born models with explicit solvent simulations. J Comput Chem 2001;22:591–607.
26. Li X, Jacobson MP, Zhu K, Zhao SW, Friesner RA. Assignment of polar states for protein amino acid residues using an interaction cluster decomposition algorithm and its application to high resolution protein structure modeling. Proteins 2007;66:824–837.
27. Jacobson MP, Friesner RA, Xiang ZX, Honig B. On the role of the crystal environment in determining protein side-chain conformations. J Mol Biol 2002;320:597–608.
28. Xiang ZX, Honig B. Extending the accuracy limits of prediction for side-chain conformations. J Mol Biol 2001;311:421–430.
29. Sellers BD, Zhu K, Zhao S, Friesner RA, Jacobson MP. Toward better refinement of comparative models: predicting loops in inexact environments. Proteins 2008;72:959–971.
30. Hendsch ZS, Tidor B. Do salt bridges stabilize proteins? A continuum electrostatic analysis. Protein Sci 1994;3:211–226.
31. Yu ZY, Jacobson MP, Friesner RA. What role do surfaces play in gb models? A new-generation of surface-generalized born model based on a novel gaussian surface for biomolecules. J Comput Chem 2006;27:72–89.
32. Still WC, Tempczyk A, Hawley RC, Hendrickson T. Semianalytical treatment of solvation for molecular mechanics and dynamics. J Am Chem Soc 1990;112:6127–6129.
33. Qiu D, Shenkin PS, Hollinger FP, Still WC. The GB/SA continuum model for solvation. A fast analytical method for the calculation of approximate born radii. J Phys Chem A 1997;101:3005–3014.
34. Gallicchio E, Paris K, Levy RM. The agbnp2 implicit solvation model. J Chem Theory Comput 2009;5:2544–2564.
35. Wagoner JA, Baker NA. Assessing implicit models for nonpolar mean solvation forces: the importance of dispersion and volume terms. Proc Natl Acad Sci USA 2006;103:8331–8336.
36. Ponder JW, Wu C, Ren P, Pande VS, Chodera JD, Schnieders MJ, Haque I, Mobley DL, Lambrecht DS, DiStasio RA, Head-Gordon M, Clark GNI, Johnson ME, Head-Gordon T. Current status of the amoeba polarizable force field. J Phys Chem B 2010;114:2549–2564.
37. Morozov AV, Kortemme T, Tsemekhman K, Baker D. Close agreement between the orientation dependence of hydrogen bonds observed in protein structures and quantum mechanical calculations. Proc Natl Acad Sci USA 2004;101:6946–6951.

38. Burley SK, Petsko GA. Aromatic-aromatic interaction: a mechanism of protein structure stabilization. Science 1985;229:23–28.
39. Yakovchuk P, Protozanova E, Frank-Kamenetskii MD. Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. Nucleic Acids Res 2006;34:564–574.
40. Churchill CDM, Navarro-Whyte L, Rutledge LR, Wetmore SD. Effects of the biological backbone on DNA-protein stacking interactions. PCCP 2009;11:10657–10670.
41. Wang LJ, Sun N, Terzyan S, Zhang XJ, Benson DR. Histidine/tryptophan pi-stacking interaction stabilizes the heme-independent folding core of microsomal apocytochrome b(5) relative to that of mitochondrial apocytochrome b(5). Biochemistry 2006;45:13750–13759.
42. Gazit E. A possible role for pi-stacking in the self-assembly of amyloid fibrils. FASEB J 2002;16:77–83.
43. Kawakami J, Okabe S, Tanabe Y, Sugimoto N. Recognition of a flipped base in a hairpinloop DNA by a small peptide. Nucleosides Nucleotides Nucleic Acids 2008;27:292–308.
44. Chelli R, Gervasio FL, Procacci P, Schettino V. Stacking and T-shape competition in aromatic-aromatic amino acid interactions. J Am Chem Soc 2002;124:6133–6143.
45. Sinnokrot MO, Sherrill CD. Highly accurate coupled cluster potential energy curves for the benzene dimer: sandwich, T-shaped, and parallel-displaced configurations. J Phys Chem A 2004;108:10200–10207.
46. Jurecka P, Sponer J, Cerny J, Hobza P. Benchmark database of accurate (mp2 and ccsd(t) complete basis set limit) interaction energies of small model complexes, DNA base pairs, and amino acid pairs. PCCP 2006;8:1985–1993.
47. Paton RS, Goodman JM. Hydrogen bonding and pi-stacking: how reliable are force fields? A critical evaluation of force field descriptions of nonbonded interactions. J Chem Inf Model 2009;49:944–955.
48. Magalhaes A, Maigret B, Hoflack J, Gomes JNF, Scheraga HA. Contribution of unusual arginine-arginine short-range interactions to stabilization and recognition in proteins. J Protein Chem 1994;13:195–215.
49. Pal TK, Sankararamakrishnan R. Self-contacts in asx and glx residues of high-resolution protein structures: role of local environment and tertiary interactions. J Mol Graphics Model 2008;27:20–33.
50. Verdonk ML, Cole JC, Hartshorn MJ, Murray CW, Taylor RD. Improved protein-ligand docking using gold. Proteins 2003;52:609–623.
51. ten Brink T, Exner TE. Influence of protonation, tautomeric, and stereoisomeric states on protein-ligand docking results. J Chem Inf Model 2009;49:1535–1546.
52. Machuqueiro M, Baptista A. Constant-ph molecular dynamics with ionic strength effects: protonation-conformation coupling in decalysine. J Phys Chem B 2006;110:2927–2933.
53. Bryngelson JD, Onuchic JN, Socci ND, Wolynes PG. Funnels, pathways, and the energy landscape of protein folding: a synthesis. Proteins 1995;21:167–195.
54. Onuchic JN, LutheySchulten Z, Wolynes PG. Theory of protein folding: the energy landscape perspective. Annu Rev Phys Chem 1997;48:545–600.