Adapting Poisson-Boltzmann to the self-consistent mean field theory: Application to protein side-chain modeling

Patrice Koehl,[1,a),b)] Henri Orland,[2,c)] and Marc Delarue[3,d)]
[1]*Department of Biological Sciences, National University of Singapore, 117543 Singapore*
[2]*Institut de Physique Théorique, CEA-Saclay, 91191 Gif/Yvette Cedex, France*
[3]*Unité de Dynamique Structurale des Macromolécules, URA 2185 du CNRS, Institut Pasteur, 75015 Paris, France*

(Received 31 May 2011; accepted 11 July 2011; published online 4 August 2011)

We present an extension of the self-consistent mean field theory for protein side-chain modeling in which solvation effects are included based on the Poisson-Boltzmann (PB) theory. In this approach, the protein is represented with multiple copies of its side chains. Each copy is assigned a weight that is refined iteratively based on the mean field energy generated by the rest of the protein, until self-consistency is reached. At each cycle, the variational free energy of the multi-copy system is computed; this free energy includes the internal energy of the protein that accounts for vdW and electrostatics interactions and a solvation free energy term that is computed using the PB equation. The method converges in only a few cycles and takes only minutes of central processing unit time on a commodity personal computer. The predicted conformation of each residue is then set to be its copy with the highest weight after convergence. We have tested this method on a database of hundred highly refined NMR structures to circumvent the problems of crystal packing inherent to x-ray structures. The use of the PB-derived solvation free energy significantly improves prediction accuracy for surface side chains. For example, the prediction accuracies for χ_1 for surface cysteine, serine, and threonine residues improve from 68%, 35%, and 43% to 80%, 53%, and 57%, respectively. A comparison with other side-chain prediction algorithms demonstrates that our approach is consistently better in predicting the conformations of exposed side chains. © *2011 American Institute of Physics*. [doi:10.1063/1.3621831]

I. INTRODUCTION

A protein structure can be described by two sets of variables: the $\phi - \psi - \omega$ dihedral angles that define the conformation of its main-chain and the χ angles that define the conformations of its side chains. While these two sets are disjoint, the values of their variables are strongly related. Understanding this dependence is one of the major tasks that need to be solved if we want to break the protein structure prediction code; attempts to unravel these rules have focused on the problem of predicting side-chain conformations, either in the theoretical case in which the actual backbone structure is known, or in the more general case in which a model for the backbone structure is available. This problem is a difficult one, with two major issues.

Predicting the conformations of side chains is mostly a combinatorial problem. Even under the assumption that side chains can only adopt a discrete set of conformations, the so-called "rotamers,"[1] a complete search of all possible side-chain packings for even a small protein is out of reach of the fastest computers currently available. In fact, it was shown theoretically that this discrete approximation of the side-chain prediction problem is not only NP complete,[2] but also "inapproximable,"[3] i.e., it is unlikely that there exists a polynomial time method that guarantees a good solution for all instances of the problem. These theoretical hardness results for side-chain positioning may not always hold in practice and have not deterred scientists from searching ways to solve this problem. This is clearly demonstrated by the considerable progress made through the developments of both exhaustive and heuristic techniques. Successful methods in this field include applications of the dead-end elimination theorem,[4–7] simulated annealing,[8,9] Monte Carlo search,[10] and graph theoretical approaches,[11–13] among others (for recent reviews, see Refs. 14 and 15). SCWRL4,[16] developed by Dunbrack *et al.*, has become a method of choice for predictors at the critical assessment of structure prediction (CASP) bi-annual experiment. It uses a detailed backbone-dependent rotamer library and a graph-theoretical approach to select the best rotamer for each side chain. Our own approach is based on the self-consistent mean field (SCMF) theory.[17] The SCMF has the advantage of being very fast as its computing time requirement grows linearly with the number of residues in the protein; in addition, it provides an estimate of the conformational entropy of the side chains of a protein.[18]

The second important issue with side-chain modeling relates to the definition of the energy function. Side chains in the core of a protein form a tight jigsaw puzzle, mostly controlled by steric effects and charge attractions and repulsions; as such, most methods that use a potential based on a

a)Author to whom correspondence should be addressed. Electronic mail: koehl@cs.ucdavis.edu.
b)On leave from Department of Computer Science and Genome Center, University of California, Davis, California 95616, USA.
c)Electronic mail: henri.orland@cea.fr.
d)Electronic mail: delarue@pasteur.fr.

Lennard-Jones potential for van der Waals (vdW) effects and a Coulomb potential for electrostatics perform well on buried side chains.[14] It should be noted that the way these potentials are implemented matters: for example, Mendes *et al.* managed to obtain better performance with our SCMF method by simply scaling the 1-4 interactions.[19] The quality of the predictive power of side-chain prediction methods decreases significantly, however, for surface residues: this is usually understood as the inability of the energy function to account for the solvent and ion atmosphere around the protein.[20] Explicit representation of water is usually not practical for predicting side-chain conformations, at least for the bulk water as it would require long simulations for averaging the positions of the corresponding water molecules. Jiang *et al.* did introduce the concept of "solvated rotamer" to account for hydrogen-bond interactions between the side chains and ordered water molecules in their proximity; they showed, however, little improvement compared to using standard rotamers.[21] The most common approach still relies on implicit solvent models. Most of these models combine a non-polar term to account for the hydrophobic effect and a polar term to account for the electrostatics contribution of the solvent.

There has been much effort invested into accounting for the hydrophobic effect when predicting side-chain conformations. The solvent-accessible-surface-area (SASA) model[22] is widely used to measure hydrophobic effects in proteins as well as to incorporate this effect in modeling, most likely due to its simplicity: SASA computes the solvation energy as a linear function of surface areas that can be computed exactly[23] or using a pairwise approximation[24] if it is to be included as a pairwise potential. Wilson *et al.* were the first to introduce the SASA model in their approach for predicting side-chain conformations;[25] they showed significant improvement in the prediction of the conformations of surface side chains when compared to other contemporary methods; their results, however, are based on a small number of proteins, and the comparisons they mention are not based on the same proteins. Mendes *et al.*[26] introduced the SASA model in the SCMF procedure for side-chain prediction and computational protein design: their results are more mitigated, as only specific surface side chains showed improvement compared to prediction without the SASA model. There is currently no clear evidence that the SASA model can improve the prediction of surface side-chain conformations. It is worth mentioning other notable efforts, however, to include the hydrophobic effect. Liang and Grishin,[27] for example, introduced a desolvation potential computed from the buried surfaces of nonhydrogen-bonded polar atoms, while Xiang *et al.*[28] implemented both a new hydrogen bond term that depends on the SASA of the side chain considered and a phenomenological energy term that favors conformations found in frequently sampled regions, thereby approximating the conformational entropy of the side chain, the so-called "colony energy."[29] Both methods show significant improvements for the prediction of the conformations of surface side chains.

Much less efforts are put into including the polar contribution of water in side-chain modeling. The simplest approach for computing the electrostatics energy of a molecule in a continuum solvent is to introduce a relative dielectric permittivity for the medium, ϵ_w ($\approx$ 80 for water) into the Coulomb's law, so that the energy of interaction between two charges is effectively screened. This expression is valid for individual point charges in a continuum solvent, but is not applicable for a large solute with a complex interface with the solvent. One approach to circumvent this limitation is to use a "distance dependent" dielectric permittivity, in which ϵ_w is set proportional to the inter-charge distance. This model is computationally simple and very efficient, and used extensively for side-chain prediction; it is, however, purely empirical, and physically unsatisfying, as it is insensitive to the actual environment of the charges. The Poisson-Boltzmann (PB) theory provides a better framework for calculating the electrostatics solvation free energy of a solute in a dielectric continuum.[30,31] As such, it is widely used to probe molecular structures; however, it is not often incorporated as a tool to compute electrostatics in a simulation or in modeling, mostly because it is computationally costly. There have been some attempts to include PB into molecular dynamics simulations based on implicit solvent models (for review, see Ref. 32). Mayo and co-workers have developed approximate formulations of the finite difference Poisson-Boltzmann method that are pairwise decomposable by side chains for computational protein design.[33–35] In their approach, the solvation free energy of one conformation of a side chain (one rotamer) is computed approximately using reduced representations of the protein structure based on the backbone and the side chain considered (the one-body PB method), or based on the backbone, the side chain considered and one other side chain, treating each pair of side chains independently (the two-body PB method). This method effectively circumvents the combinatorial complexity inherent in side-chain modeling or computational sequence design experiments; it is, however, an approximation that is most likely too simplistic for modeling clusters of charged side chains that are often observed in protein active sites.[36] Harbury *et al.* expanded upon this approach and developed FDPB_MF, a method that exhaustively samples side-chain conformational space and rigorously calculates multi-body protein-solvent interactions.[37] This method has been used for computing the ionization states of titratable residues in proteins; its implementation, however, is very slow (days of computing time) and limited to predicting the pKas and conformations of only a few residues.

In this paper, we revisit the problem of incorporating the polar term of implicit solvent models for the prediction of protein side-chain conformations. We propose to combine the SCMF and PB theories and derive a new, exact formalism, SCMF-PB, that is fast enough to be used for predicting the conformations of all the side chains of even a large protein, with only minutes of computing time on a desktop computer.

The paper is organized as follows. In Sec. II, we provide very brief overviews of the SCMF and PB theories, and then we derive how they are combined into the SCMF-PB formalism. Section III describes our implementation of the SCMF-PB formalism, with an emphasis on efficiency. The Results section compares the predictive powers of SCMF-PB with four other leading softwares for side-chain prediction, OPUS-ROTA,[9] SCAP,[10] TREEPACK,[38] and SCWRL4,[16] as they are tested on the DRESS database.[39] We conclude with a

discussion on possible improvements and new applications of the SCMF-PB method.

II. CONTINUUM ELECTROSTATICS FOR SELF-CONSISTENT MEAN FIELD THEORY

A. The self-consistent mean field formalism

The SCMF method for modeling side chains is fully described in Refs. 17 and 18. Briefly, let us consider a protein with N residues whose coordinates are stored in $\mathbf{X}$. If the side chain of residue i can take up to K_i conformations (or rotamers), the total conformational space to explore is of size $\prod_{i=1}^{n} K_i$. To circumvent this combinatorial explosion, the SCMF method considers an effective multi-copy system $\mathbf{X}^0$ in which each side-chain i is represented by the list of its K_i possible rotamers. Assuming independence of the different multi-copy subsystems, the probability density function for this effective system is given by a Hartree product,

$$\rho(\mathbf{X}) = \prod_{i=0}^{N} \sum_{j=1}^{K_i} P(i,\, j)\delta\left(\mathbf{X}_i - \mathbf{X}_{ij}^0\right), \tag{1}$$

where $i = 0$ corresponds to the backbone of the protein. $P(i,\, j)$ are normalization or weight factors that satisfy $\sum_{j=1}^{K_i} P(i,\, j) = 1$ for all i.

Assuming that the multiple copies of any subsystem do not interact with each other, the variational free energy of the whole system is

$$\mathcal{F} = \; E_{eff}(\mathbf{X}^0) + k_B T \sum_{i=0}^{N} \sum_{j=1}^{K_i} P(i,\, j) \ln(P(i,\, j)), \tag{2}$$

where k_B is the Boltzmann constant and T is the temperature. The first term on the right side of this equation is the effective potential energy of the multi-copy system (see below) and the second term is the conformational entropy. Note that in this formalism, the positions of the different copies of the side chains are fixed while the probabilities P are variable. Setting $d\mathcal{F}/dP(i,\, j) = 0$, we find that these factors satisfy the equations (see Ref. 40),

$$P(i,\, j) = \frac{\exp\left(-\beta \frac{dE_{eff}}{dP(i,j)}\right)}{\sum_{k=1}^{K_i} \exp\left(-\beta \frac{dE_{eff}}{dP(i,k)}\right)}. \tag{3}$$

These are fixed point equations as E_{eff} is a function of the factors $P(i,\, j)$. Just like for any Hartree method, they are solved iteratively using a fixed point algorithm until self-consistency is reached.[18]

The energy E_{eff} is the key factor that defines the quality of side-chain prediction. In the current implementation of SCMF, it is defined as

$$E_{eff}(\mathbf{X}^0) = \sum_{i=1}^{N} \sum_{j=1}^{K_i} P(i,\, j) \times \left(U_p(i_j) + \frac{1}{2} \sum_{k \neq i} \sum_{l=1}^{K_k} P(k,\, l) U_p(i_j, k_l) \right), \tag{4}$$

where $U_p(i_j)$ and $U_p(i_j, k_l)$ are shortcuts for the internal energy of side-chain i in its configuration (rotamer) j and the energy of interaction between side-chain i in configuration j and side-chain k in configuration l, respectively. This definition follows a standard Hartree method, in which each "particle" (i.e., side chain) feels the mean field created by all other "particles."

The energy function U_p is pairwise additive; for a pair of atoms a and b, it is defined as the sum of the energies of their vdW and electrostatic interactions,

$$U_p(\mathbf{r_a}, \mathbf{r_b}) = \frac{A_{ab}}{r_{ab}^{12}} - \frac{B_{ab}}{r_{ab}^{6}} + \frac{1}{4\pi\epsilon_0\epsilon_p} \frac{q_a q_b}{r_{ab}}, \tag{5}$$

where q_a and $\mathbf{r_a}$ are the charge and position of atom a, r_{ab} is the inter-atomic distance, A_{ab} and B_{ab} are the parameters of the vdW energy based on the types of atoms a and b, ϵ_0 is the vacuum permittivity, and ϵ_p is the relative permittivity inside the protein.

The fixed-point equation (3) requires the derivatives of the effective energy with respect to the weights $P(i,\, j)$, which are given by

$$\frac{dE_{eff}}{dP(i,\, j)} = U_p(i_j) + \sum_{k \neq i} \sum_{l=1}^{K_k} P(k,\, l) U_p(i_j, k_l). \tag{6}$$

B. The Poisson-Boltzmann formalism

The Poisson-Boltzmann theory provides a rigorous theoretical framework for calculating the electrostatics energy of a solute surrounded by such a dielectric continuum; it is by far the most popular method in this category. The PB equation in its standard form is given by

$$\nabla[(\epsilon_0 + \gamma_{solv}(\mathbf{r})\epsilon_0\chi)\, \nabla\phi(\mathbf{r})] - \gamma_{ion}(\mathbf{r})\kappa^2 \sinh(\phi(\mathbf{r})) + 4\pi\beta e_C \rho_f(\mathbf{r}) = 0, \tag{7}$$

where ϵ_0 is the dielectric constant *in vacuo* and χ the dielectric susceptibility of water (considered to be constant). $\kappa^2 = 8\pi\beta e_C^2 I$, where e_c is the charge of the electron and I is the ionic strength of the bulk solution. The solute is described by a fixed charge density ρ_f and a solvent accessibility function $\gamma_{solv}(\mathbf{r})$ that is zero for points at positions $\mathbf{r}$ inside the envelope of the solute and one otherwise. This envelope can be taken as the molecular surface or the accessible surface of the solute. We consider here a 1-1 salt, in which case $I = c_s$, the bulk salt concentration. The indicator function $\gamma_{ion}(\mathbf{r})$ for the presence of ions at position $\mathbf{r}$, is usually set equal to $\gamma_{solv}(\mathbf{r})$ though sometimes an ion-excluded zone is defined in the neighborhood of the solvent, the so-called *Stern zone*. Note that in this formulation, ions are not represented explicitly. Instead, the ions are considered to be in thermal equilibrium with each other and relatively free to move.

The Poisson-Boltzmann equation can be expressed as the state function of a functional free energy, as introduced by Sharp and Honig,[41]

$$\mathcal{F}_{PB} = -\frac{1}{2} \int \epsilon(\mathbf{r})\, (\nabla\phi(\mathbf{r}))^2\, d\mathbf{r}^3 + 4\pi \int \rho_f(\mathbf{r})\phi(\mathbf{r}) d\mathbf{r}^3 - \kappa^2 \int \gamma_{ion}(\mathbf{r})(\cosh(\phi(\mathbf{r})) - 1) d\mathbf{r}^3, \tag{8}$$

where $\epsilon(\mathbf{r}) = \epsilon_0 + \gamma_{solv}(\mathbf{r})\epsilon_0\chi$ is the position-dependent dielectric constant, and the integrals are computed over all space.

The contribution of electrostatics to the solvation free energy of the solute considered is then given by

$$\mathcal{F}_{solv} = \mathcal{F}_{\rm PB}(\bar{\phi}_{solvent}) - \mathcal{F}_{\rm PB}(\bar{\phi}_{vacuo}), \tag{9}$$

where $\bar{\phi}_{solvent}$ and $\bar{\phi}_{vacuo}$ are the solutions of the PB equation solved in the presence and absence of solvent, respectively. Note that by taking the difference between the free energy in water and the free energy *in vacuo*, we get exactly the effect of the solvent on the solute. In particular, the grid effects inherent to solving the PB equation numerically, such as self-energies due to the discretization of the fixed charges,[42] are removed.

C. The SCMF-PB formalism

The SCMF-PB formalism is a modification of the SCMF method for predicting side-chain conformation that takes into account the electrostatic contribution of solvation, as computed by the PB equation. In practice, this means that the effective potential E_{eff} for the multi-copy system is modified as

$$E_{eff,\rm PB} = E_{eff} + \mathcal{F}_{\rm PB}(\bar{\phi}_{solvent}) - \mathcal{F}_{\rm PB}(\bar{\phi}_{vacuo}), \tag{10}$$

where $\bar{\phi}_{solvent}$ and $\bar{\phi}_{vacuo}$ are the solutions of the PB equation solved for the multi-copy system in the presence and absence of solvent, respectively. Note that by taking the difference between the free energy in water and the free energy *in vacuo*, we remove all interactions between multi-copies of the same side chain (in addition to removing the grid effects), a necessary condition of our SCMF approach.

Applications of the SCMF theory requires the derivatives of $\mathcal{F}_{\rm PB}(\bar{\phi})$ with respect to the weight factors $P(j,k)$ of the different side-chain copies, for $\bar{\phi}$ solution of the PB equation. In Eq. (8), the coefficients $\epsilon(\mathbf{r})$, $\gamma_{ion}(\mathbf{r})$ and the charge density $\rho_f(\mathbf{r})$ depend directly on the coefficients P, while $\bar{\phi}$ depends implicitly on P. The latter dependence, however, can be safely ignored. Indeed, if we use the chain rule

$$\begin{aligned}\frac{d\mathcal{F}_{\rm PB}}{dP(i,j)}(\bar{\phi}) &= \frac{\delta\mathcal{F}_{\rm PB}}{\delta P(i,j)}(\bar{\phi}) + \frac{\delta\mathcal{F}_{\rm PB}}{\delta\phi}(\bar{\phi})\frac{\delta\phi}{\delta P(i,j)}(\bar{\phi}) \\ &= \frac{\delta\mathcal{F}_{\rm PB}}{\delta P(i,j)}(\bar{\phi})\end{aligned} \tag{11}$$

as the derivatives $\delta\mathcal{F}_{\rm PB}/\delta\phi(\phi)$ are equal to 0 for $\phi = \bar{\phi}$. In the following, we explain how to set up the PB equation for a multi-copy system and how to compute these derivatives.

1. Solving the PB equation for a probabilistic solute

The PB equation can be solved numerically using finite difference methods. In these methods, it is discretized on a Cartesian 3D mesh, defined by its vertices $\mathbf{r}$. The discretization is complete once the four maps $\gamma_{solv}(\mathbf{r})$, $\epsilon(\mathbf{r})$, $\gamma_{ion}(\mathbf{r})$, and the charge density $\rho_f(\mathbf{r})$ are known.

The map $\gamma_{solv}(\mathbf{r})$ is usually set to 0 if the mesh point $\mathbf{r}$ is covered by the solute and 1 otherwise. Extension to the multi-copy system leads to a probabilistic function $\gamma_{solv}(\mathbf{r})$

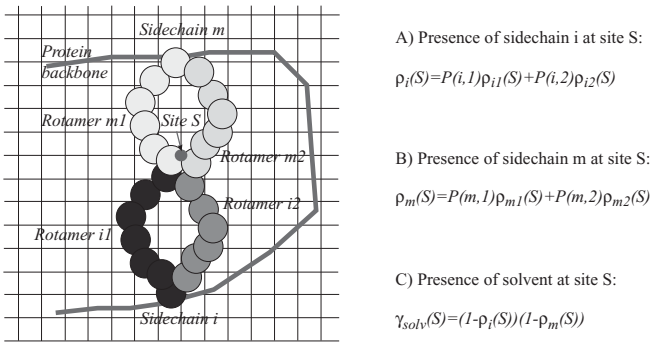

FIG. 1. Illustration of the definition of probabilist maps for Poisson-Boltzmann calculation. Let us consider a toy protein consisting of two amino acids, i and m, that can both adapt one of the two possible conformations, or rotamers. The two rotamers for amino acid i are assigned weights $P(i,1)$ and $P(i,2)$ such that $P(i,1) + P(i,2) = 1$. The corresponding multi-copy system is shown on the left, overlaid on a Cartesian grid. As the two conformations for amino acid i are independent realization of the side-chain position, the probability $\rho_i(S)$ that amino acid i covers a site S is simply the weighted sum of the probabilities $\rho_{i1}(S)$ and $\rho_{i2}(S)$ of each of its conformations covering S (A). The same applies for amino acid m (B). As i and m co-exist in the protein, the probability $\gamma_{solv}(S)$ that site S remains "uncovered," i.e., accessible to solvent, is given by a product rule (C).

as a mesh point can be covered by a mixture of solute and solvent (see Fig. 1 for illustration). Let us consider atom k in rotamer j of residue i. k is represented as a ball $\mathcal{B}_k$ with center $\mathbf{C}_{ijk}$ and radius R_{ijk} (set to its vdW radius). The probability $\rho_{ijk}(\mathbf{r})$ that atom k is present at $\mathbf{r}$ is 1 if $\mathbf{r}$ is inside the ball $\mathcal{B}_k$ and 0 otherwise. Several atoms from the same rotamer j may cover the mesh site $\mathbf{r}$. The probability of any atom of rotamer j present at $\mathbf{r}$ is then given by the product rule,

$$\rho_{ij}(\mathbf{r}) = 1 - \prod_{k=1}^{N_i}(1 - \rho_{ijk}(\mathbf{r})), \tag{12}$$

where N_i is the number of atoms in side-chain i. We define S_{ij} as the set of all grids points where $\rho_{ij}(\mathbf{r})$ is nonzero. The different rotamers j are independent realizations of side-chain i; the probability that i is present at grid point $\mathbf{r}$ is, therefore, given by the additive rule,

$$\rho_i(\mathbf{r}) = \sum_{j=1}^{N_i} P(i,j)\rho_{ij}(\mathbf{r}). \tag{13}$$

Several side chains may cover $\mathbf{r}$. Using the product rule, the probability of presence of the solvent at position $\mathbf{r}$ is

$$\gamma_{solv}(\mathbf{r}) = \prod_{i=1}^{N}(1 - \rho_i(\mathbf{r})) \tag{14}$$

and the probability of presence of the solute at the same position is simply $\gamma_{solute}(\mathbf{r}) = 1 - \gamma_{solv}(\mathbf{r})$.

In presence of salt, we define an "ion map," i.e., the probability $\gamma_{ion}(\mathbf{r})$ of the presence of ions at each site $\mathbf{r}$ in the lattice. As a first approximation, we equate the ion map to the solvent map, $\gamma_{ion}(\mathbf{r}) = \gamma_{solv}(\mathbf{r})$.

For the dielectric map, we note that the Born energy is a linear function of the water density.[37] This leads to

$$\epsilon(\mathbf{r}) = \left[\gamma_{solv}(\mathbf{r})\epsilon_w^{-1} + (1 - \gamma_{solv}(\mathbf{r}))\epsilon_p^{-1}\right]^{-1}, \tag{15}$$

where ϵ_p and ϵ_w are the dielectric constant inside the protein and in the solvent, respectively (a similar expression was originally derived by Davis and McCammon for smoothing the dielectric boundary around a solute to improve the convergence and accuracy of PB solvers[43]).

For the charge density map, let us define z_{ijk} as the charge of atom k in rotamer j of side-chain i. It is customary to position this charge at the atom center C_{ijk}; this center, however, most likely does not match with a vertex in the grid considered, especially if this grid is Cartesian. The standard approach to circumvent this problem is to project the charge on multiple vertices of the grid. We have implemented the sphere charging model,[44]

$$z_{ijk}(\mathbf{r}) = \frac{\rho_{ijk}(\mathbf{r}) z_{ijk}}{N_{ijk}}, \tag{16}$$

where N_{ijk} is the number of mesh points covered by $k (N_{ijk} = \sum_{\mathbf{r}} \rho_{ijk}(\mathbf{r}))$. The total charge contributed by rotamer j of side-chain i at $\mathbf{r}$ is, therefore,

$$z_{ij}(\mathbf{r}) = \sum_k z_{ijk}(\mathbf{r}) \tag{17}$$

and the total (fixed) charge at the same position is

$$z(\mathbf{r}) = \sum_{i=0}^{N} \sum_{j=1}^{N_i} P(i,j) z_{ij}(\mathbf{r}) \tag{18}$$

or as a density,

$$\rho_f(\mathbf{r}) = \frac{1}{V} \sum_{i=0}^{N} \sum_{j=1}^{N_i} P(i,j) z_{ij}(\mathbf{r}), \tag{19}$$

where V is the volume element associated with the grid point at position $\mathbf{r}$.

2. Derivatives of the PB free energy with respect to P(i, j)

From the definition of the PB free energy and as a consequence of Eqs. (11) and (15), we have

$$\begin{aligned} \frac{d\mathcal{F}_{\mathrm{PB}}}{dP(i,j)}(\bar{\phi}) = &-\frac{1}{2} \int \frac{d\gamma_{solv}(\mathbf{r})}{dP(i,j)} \epsilon(\mathbf{r})^2 \left(\frac{1}{\epsilon_p} - \frac{1}{\epsilon_w} \right) \\ &\times \left(\nabla \bar{\phi}(\mathbf{r}) \right)^2 d\mathbf{r}^3 + 4\pi \int \frac{d\rho_f(\mathbf{r})}{dP(i,j)} \bar{\phi}(\mathbf{r}) d\mathbf{r}^3 \\ &- \int \frac{d\gamma_{solv}(\mathbf{r})}{dP(i,j)} \kappa^2(\mathbf{r}) (\cosh(\bar{\phi}(\mathbf{r})) - 1) d\mathbf{r}^3. \end{aligned} \tag{20}$$

The derivatives of the solvent map and charge density map with respect to the weight factors $P(i,j)$ are computed analytically: from Eq. (14), we get

$$\frac{\delta \gamma_{solv}(\mathbf{r})}{\delta P(i,j)} = -\rho_{ij}(\mathbf{r}) \prod_{m \neq i} (1 - \rho_m(\mathbf{r})) \tag{21}$$

and from Eq. (18), we get

$$\frac{\delta \rho_f(\mathbf{r})}{\delta P(i,j)} = \frac{z_{ij}(\mathbf{r})}{V}. \tag{22}$$

III. IMPLEMENTATION

The SCMF-PBE algorithm was implemented based on the AQUASOL program.[45] It proceeds as follows.

ALGORITHM I. *The SCMF-PB method for predicting side-chain conformations*

```
Initialize side-chain weights P_0(i, j) = 1/K_i
for n = 1, ... until convergence do
    (1) setup the PB equation: compute γ_solv(r), ε(r)
        and ρ_f(r) for the current P(i, j),
    (2) solve iteratively the nonlinear PB equation
        twice, for φ̄_solvent and φ̄_vacuo,
    (3) compute the variational free energy
        (Eq. (2)) and its derivatives with respect
        to the weights P,
    (4) compute P_calc(i, j) using Eq. (3) and update
        weights:
          P_n(i, j) = λ P_calc(i, j) + (1 − λ) P_{n−1}(i, j),
    (5) check for convergence: if ‖P_n − P_{n−1}‖ < TOL,
        stop
end for
```

First, the multi-copy protein is generated, using a "chemistry-based" library of rotamers, with χ dihedral angles set systematically to their preferred $g+(+60°)$, $g-(-60°)$ and t (180°) conformations, except for the χ_2 of aromatic residues that are set to $\pm 90°$. In the initialization step, each copy of each side chain is assigned a weight P equal to the inverse of the number of copies for that side chain. As part of the initialization process, all side-chain interaction energies are computed and stored for use in the refinement cycles (see Ref. 18 for details). These energies include a Lennard-Jones term for the vdW interactions and a Coulomb term for internal electrostatics (see Eq. (5)). 1-4 interactions are scaled down with a factor 0.4, and the value for the potential U_{ab} between two atoms a and b is truncated to a maximum value of 10 kcal/mol to avoid problems related to using fixed rotamer conformations.[18,46] A distance-dependent geometric potential is also introduced to direct the formation of disulphide bridges as defined in the ECEPP (Empirical Conformational Energy Program for Peptides) energy function.[47] To speed up energy calculations between side chains, a 17 Å cutoff is used (residues whose C_α are at a distance larger than this cutoff have their interaction energies set to 0). The corresponding computing time is linear with respect to the total number of side chains in the multi-copy system.

In step (1), the current multi-copy system is projected on the Cartesian grid used to solve the PB equation numerically. The solvent, dielectric, and fixed charge density maps are computed over all vertices in that grid, following the methods implemented in AQUASOL.[45] In particular, we use the sphere charging model[44] to project charges in the grid. Note that these maps are probabilistic, as described above, and need to be computed at each cycle, as they depend on the weights $P(i,j)$. The computing time required to set up these maps is proportional to the total number of atoms in the multi-copy system.

Once the multi-copy system is set up on the grid, the Poisson-Boltzmann equation is solved twice in step (2), in

the presence and absence of solvent, respectively. We use the truncated Newton method with multigrid pre-conditioner implemented in AQUASOL; this method was originally introduced by Holst and Saied[48] for solving the nonlinear PB equation. To speed up convergence, the electrostatic map computed at cycle n is used as input for the PB solver at step $n+1$: as the weights $P(i, j)$ start converging, the different maps computed in step 1 do not change significantly between two cycles, and the previous electrostatic potential map becomes a good estimate of the current electrostatic map. The computing time required to solve the PB equation is linear in the total number of vertices of the 3D grid considered.[45]

Once the electrostatic potential maps are known, we compute in step (3) the corresponding solvation free energy that is subsequently added to the internal energy and the conformational entropy to get the full variational free energy (Eq. (2)). The derivatives of this free energy with respect to the different weights $P(i, j)$ are computed analytically, using Eq. (6) for the internal energy part, and Eqs. (11), (21), and (22) for the solvation free energy part. In theory, this step is the most time consuming: each derivative of the solvation energy with respect to a weight $P(i, j)$ requires a numerical integration over the whole grid used to solve the PB equation (see Eq. (20)). We note, however, that for a given rotamer j of a residue i, the derivatives of the solvent map and charge density maps with respect to the weight $P(i, j)$ (defined in Eqs. (21) and (22)) are nonzero only over S_{ij}, the set of vertices covered by the rotamer. The number of such vertices is small compared to the full size of the grid; it depends on the number of atoms in the side chain (smaller than 20), the vdW radii of all these atoms, and the grid size. This observation makes the calculation of all derivatives fast, basically proportional to the total number of side-chain copies in the whole multi-copy system.

A new set of weights $P_{calc}(i, j)$ is computed in step (4) using Eq. (3) and the weights $P(i, j)$ are subsequently updated with a damping factor λ, to avoid oscillation in the convergence.[18,49]

The procedure is iterated until convergence, i.e., until the weights and energy terms are self-consistent. This convergence can be monitored by measuring the changes in the matrix $\mathbf{P}$ whose generic element is $P(i, j)$; convergence is reached when the L_2 norm of the difference between two consecutive matrices is 0, or in practice when this norm falls below a cutoff TOL set to 0.001.

The predicted conformation of each side-chain is set to its copy with the highest weight in the converged matrix $\mathbf{P}$.

IV. RESULTS AND DISCUSSION

We evaluate our new method for predicting protein side-chain conformations on the DRESS database.[39] We show that adding the electrostatic solvation free energy to the variational free energy minimized by the SCMF method improves the accuracy with which the conformations of exposed polar side chains can be predicted. Next we compare our results to those of four successful side-chain prediction programs: OPUS-ROTA, SCAP, TREEPACK, and SCWRL4, the latter being the method of choice used by modellers at CASP. As the emphasis of the paper is on combining the SCMF and PB formalisms, we then show that the increased complexity of the combined free energy functional and the corresponding increase in the complexity of the implementation of SCMF-PB do come at a computational cost that is, however, very much manageable as the computing time required to predict the conformations of all side chains of a 370 residue protein remains below 2 min on a single core commodity personal computer.

A. Protein dataset and accuracy analysis

Test sets that are used to assess the quality of side-chain modeling programs usually include only high resolution x-ray structures, as those structures are considered to be of better quality than those obtained by other experimental techniques. They are not, however, devoid of biases. For example, it was shown that the crystal environment plays an important role in determining the conformations of polar side chains on the surfaces of proteins.[50] Clearly, addition of the crystal packing constraints is expected to lead to improved side-chain predictions.[28,50] This is not only difficult, as it would involve reconstructing explicitly the environment around a protein using the known space group and unit cell dimensions as well as taking into account all hetero-atoms, including the precipitating agents, but also somewhat artificial as it is not directly related to the problem of predicting the conformations of protein side chains in solution. In this study, we focus specifically on the effect of solvation on protein side-chain conformations. We therefore did not use a test set based on x-ray structures; instead, we used the DRESS database.[39] We chose this database as it contains 100 high resolution NMR structures that have been refined using molecular dynamics in explicit water: as such, the conformations of their exposed side chains are expected to be good approximations of their actual conformations in solution. These structures vary in size from 20 to 370 amino acids. The corresponding PDB (Protein Data Bank) files were processed to remove all information about the native conformations of the side chains.

In analyzing the results, a dihedral angle was considered "correctly predicted" if it was within 40° from its conformation in the native structure.[14,18] χ_1 indicates the percentage of side chains for which the first torsion angle was correctly predicted, while χ_{1+2} represents the percentage of side chains for which both the first and the second dihedral angles were correctly predicted. Statistics are computed for core (buried) and surface (exposed) residues separately. Residues were defined as buried when their accessibility is less than 20%, and exposed otherwise. Surface areas were calculated using the program ENVIRON.[51]

B. Effects of solvation on the accuracy of side-chain modeling

We evaluate the quality of prediction of side-chain conformations for both the standard SCMF procedure and our new SCMF-PB algorithm on the DRESS database. The energy function U_p is defined according to CHARMM19 parameters; ϵ_p was set to 4 for both the Coulomb term in U_p and to define the dielectric inside the protein in the PB equation. The

TABLE I. Improvement of SCMF-PB over SCMF.

	Core residues[a]			Surface residues		
AA	Nres[b]	χ_1, % SCMF (+PB)	χ_{1+2}, % SCMF (+PB)	Nres	χ_1, % SCMF (+PB)	χ_{1+2}, % SCMF (+PB)
VAL	302	79.47 (81.79)		236	72.46 (74.58)	
ILE	266	86.84 (87.97)	63.91 (64.66)	177	76.27 (76.27)	53.11 (51.98)
LEU	358	82.96 (83.24)	57.26 (57.82)	298	75.50 (73.49)	52.01 (52.01)
PRO	63	88.89 (87.30)		231	67.10 (68.40)	
MET	67	82.09 (82.09)	56.72 (56.72)	106	51.89 (51.89)	35.85 (34.91)
CYS	114	85.09 (87.72)		60	**68.33 (80.00)**	
SER	120	**42.50 (60.00)**[c]		418	**35.89 (53.11)**	
THR	131	**58.78 (67.18)**		321	**43.61 (57.01)**	
PHE	165	91.52 (92.73)	82.42 (83.64)	92	77.17 (77.17)	70.65 (71.74)
TYR	96	90.62 (89.58)	78.12 (78.12)	163	75.46 (73.62)	73.01 (72.39)
TRP	57	75.44 (70.18)	***59.65 (49.12)***	55	65.45 (65.45)	***27.27 (21.82)***
HIS	21	95.24 (90.48)	***33.33 (28.57)***	105	75.24 (71.43)	24.76 (23.81)
ASN	60	75.00 (80.00)	***51.67 (41.67)***	287	60.63 (61.67)	35.89 (33.80)
GLN	36	91.67 (94.44)	75.00 (69.44)	281	67.62 (68.33)	44.84 (44.13)
ASP	65	80.00 (72.31)	46.15 (49.23)	378	51.32 (48.94)	39.42 (43.39)
GLU	55	67.27 (72.73)	45.45 (47.27)	572	63.99 (64.51)	42.13 (44.23)
LYS	44	75.00 (75.00)	59.09 (59.09)	561	60.61 (63.28)	**36.72 (42.60)**
ARG	31	**61.29 (70.97)**	41.94 (41.94)	407	**48.16 (65.85)**	**30.96 (41.77)**
All	2051	79.18 (81.47)	61.85 (61.39)	4748	59.84 (64.11)	42.02 (44.57)

[a]Core and surface refer to residues whose accessibility is lower than or greater than 20%, respectively.
[b]Nres is the number of residues of the given type in the whole DRESS database.
[c]Changes greater than 10% are highlighted in bold and in bold-italic when the addition of solvation leads to improvement or decrease in quality, respectively.

dielectric constant for the solvent is set to 80. The PB calculations are carried out on a 65 cubic mesh with 1 Å spacing in each direction. A 0.05M monovalent ion atmosphere was added. Results concerning the accuracy of the predictions of χ_1 and χ_{1+2} between the "correct" structure in the database and the final model generated are given in Table I and in Figs. 2 and 3.

As shown in Table I and Fig. 2, adding solvation improves the prediction of most polar residues while having no significant effects on hydrophobic residues. The accuracy

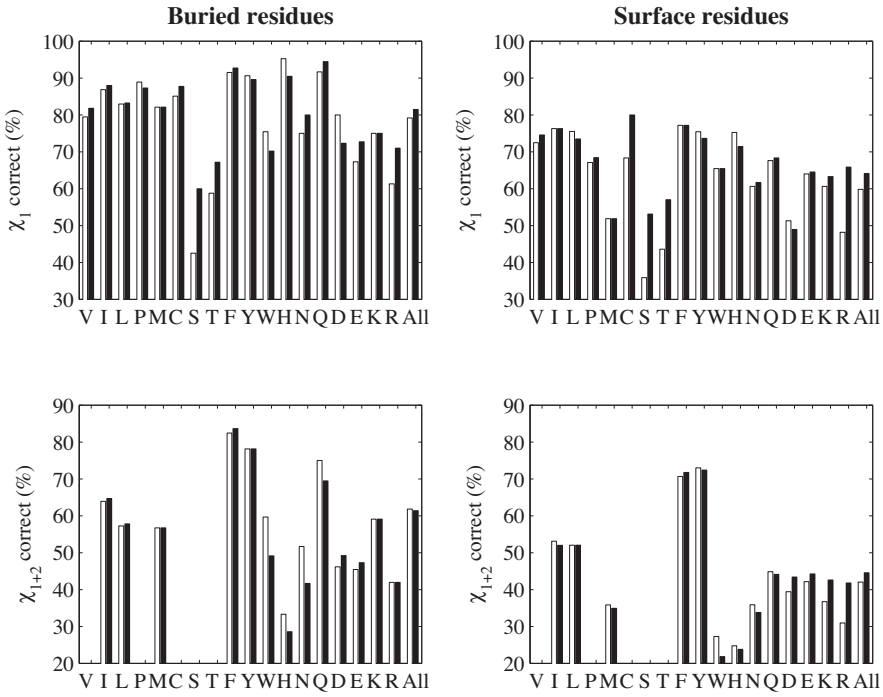

FIG. 2. Effects of adding the electrostatic free energy as computed by PB in the SCMF functional free energy on side-chain prediction accuracy. The dihedral angles χ_1 and χ_2 are considered to be correctly predicted if their values are within 40° of their reference values in the native structures.

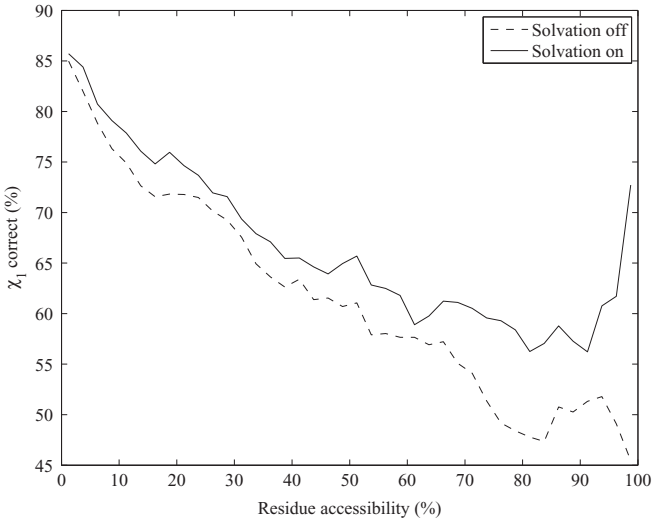

FIG. 3. The accuracy with which the dihedral angles χ_1 of all residues in the DRESS database are predicted is plotted as a function of the accessibility of the residue to solvent, for the SCMF calculation with (continuous line) and without (dashed line) the PB solvation free energy.

with which the χ_1 dihedral angle of side chains can be predicted increases from 60% to 64% over all surface residues and only from 79% to 81% for core residues. The largest improvements are observed for arginines, serines, cysteines, and threonines. Arginine, serine, and threonine are the three types of amino acids that are the least well predicted *in vacuo*; as these three amino acids are polar, it is expected that the addition of solvation would lead to significant improvements, which is indeed observed. The case of cysteine is a little different. Most buried cysteine residues are involved in disulphide bridges, which are handled adequately by the geometric term added to the effective energy U_p (we note that SCMF does not use any information on the pairing of cysteine in disulphide bridges; instead, this information can be deduced from the predicted conformations of the cysteine side chains). Cysteines at the surface, however, are mostly reduced; their side chains are small. It is interesting that even though these are neutral, non-polar residues, their orientations are better predicted in the presence of solvent. It is noteworthy that the addition of solvation does not always lead to improvement: we observe a decrease in the accuracies with which the χ_2 angles of buried histidine, tryptophan, asparagine, and to a lesser extent glutamine residues are predicted. The loss of performance for histidines is likely to be related to incorrect assignments of their ionization states, as they are chosen arbitrarily at the initiation step of the modeling procedure; such incorrect assignments are expected to be more significant when solvation effects are considered. It is interesting that SCMF-PB underperforms for the two amino acid types whose side chains contain an amide group; this is most likely related to the solvation term in the variational free energy disrupting the hydrogen bond patterns that these amide groups can form within the protein, though it is unclear why this would be the case. The failure to improve tryptophan highlights that there is still a need for better electrostatic models for large aromatic residues (see, for example, Ref. 52).

TABLE II. Accuracy of SCMF-PB compared to other methods.

	Core residues[a]			Surface residues		
Program	Nres[b]	χ_1 (%)	χ_{1+2} (%)	Nres	χ_1 (%)	χ_{1+2} (%)
OPUS	2051	82.69	68.66	4748	63.77	41.56
SCAP	2051	81.03	64.95	4748	63.98	38.77
TREE	2051	76.11	58.82	4748	61.46	37.31
SCWRL	2051	80.64	63.97	4748	61.18	38.14
SCMF-PB	2051	81.47	61.39	4748	64.11	44.57

[a]Core and surface refer to residues whose accessibility is lower than or greater than 20%, respectively.
[b]Nres is the number of residues of the given type in the whole DRESS database.

The effect of adding the solvation free energy on the prediction accuracy is evident from Fig. 3 in which we plot the percentage of correctly predicted χ_1 angles as a function of the accessibility of the residue, for all 6809 non-Gly, non-Ala residues of the proteins in the DRESS database. Clearly, the improvement resulting from adding a solvation term increases as the residue accessibility increases, as intuitively expected.

The solvation free energy considered in the SCMF-PB model is derived from the solution of the PB equation for the multi-copy system representing the protein under study. In all test cases considered above, we solved the PB equation over a $65 \times 65 \times 65$ cubic grid. We also tested finer meshes with up to 257^3 vertices and did not observe significant differences; smaller grids, however, with larger grid spacing resulted in loss of performance; we therefore used the 65^3 grid size in all subsequent computations. Similarly, all calculations presented above were performed in the presence of 0.05M salt; we tested higher salt concentrations, up to 1M and saw only small variations that were statistically nonsignificant.

C. Comparison with other programs

Table II and Fig. 4 allow for the comparison of the results obtained with the program SCMF-PB with those obtained with four other programs, OPUS-ROTA,[9] SCAP,[10] TREEPACK,[38] and SCWRL4 (Ref. 16) on the DRESS database.

Similar to what was observed in the comparison for SCMF-PB with SCMF, the effect of adding solvation in the energy function of SCMF-PB is evident from Fig. 4: clearly, SCMF-PB produces the most accurate results for exposed residues, especially as the residue accessibility becomes large. Of all the other methods tested here only OPUS-ROTA accounts explicitly for solvation: its energy function contains a surface-dependent term based on the Eisenberg and McLachlan model.[22] The surface-based solvation free energy term, however, only accounts for the non-polar contribution of solvation; a similar term has already been introduced within the SCMF model for side-chain prediction and protein design with only mitigated results.[26] OPUS-ROTA, TREEPACK and SCWRL4 use a backbone dependent rotamer library and include a statistical term in their energy functions that accounts for the rotamer frequency, as observed in native proteins. We expect this term to matter for buried side chains but not for exposed side chains in the DRESS database, as

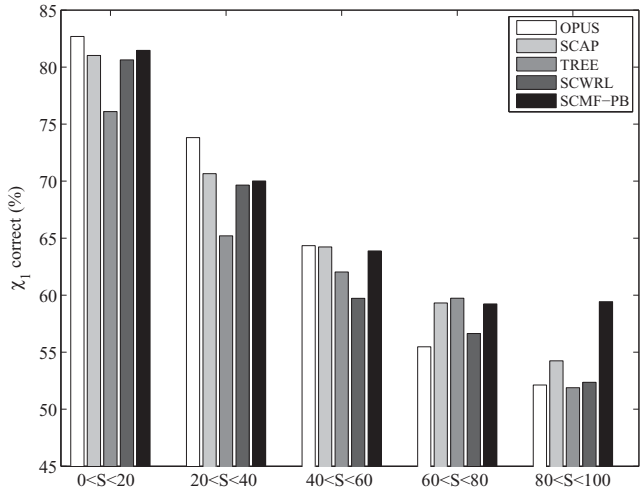

FIG. 4. The accuracy of SCMF-PB is compared to those of other side-chain prediction programs for a range of side-chain accessibility S (in %).

proteins in these databases were refined in solvent and not within crystals. Indeed, OPUS-ROTA showed the most accurate results for core residues. Although OPUS-ROTA, TREEPACK, and SCWRL4 use the same backbone-dependent rotamer library, OPUS-ROTA has a different rotamer frequency term; in addition, it includes an orientation-dependent side-chain packing potential[53] that may explain its improved prediction levels. In contrast, SCMF-PB uses a simple chemistry-based rotamer library with no statistical information from existing protein structure databases; it is encouraging that it still performs well for core residues. The version of SCAP that is publicly available is based on the original work from Xiang and Honig.[10] As such, its energy function only includes a vdW term and a torsion term. Surprisingly, it performs relatively well on exposed residues (see Fig. 4), being outperformed only by SCMF-PB for highly accessible residues. It is possible that the addition of the colony energy as described in a more recent work from the same authors[28] would significantly improve the results of SCAP.

D. Central processing unit (CPU) time requirements

It is important to consider the running time of an algorithm as it determines whether it is usable in practice. To our knowledge, there is so far only one report of the combination of SCMF with the PB equation to study the ionization state of titratable amino acids in proteins. The corresponding FDPB_MF algorithm was found to be slow, its running times varying between 2 and 6 days on a Pentium4, 2.4 GHz CPU.[37] Our implementation of a combined SCMF-PB algorithm is much faster, with computing times in the order of 2 min for predicting the conformations of all side chains of large proteins (370 residues) on a Intel Core I7 at 2.66 GHz with 8 GB of memory. We believe that the main difference between the two implementations comes from the fact that we use an analytical expression for the derivatives of the variational free energy of the SCMF method with respect to the weights $P(i, j)$, while FDPB_MF computes these derivatives numerically.[37]

On an Intel Core I7 at 2.66 GHz, the average CPU time required per protein in the DRESS database was 0.3, 0.3, 1.9, 2, 9, and 93 s for SCMF, TREEPACK, OPUS-ROTA, SCWRL4, SCAP, and SCMF-PB, respectively. Clearly, SCMF-PB is slower than the other methods; the time difference, however, is only of one order of magnitude, and SCMF-PB offers an improved level of prediction accuracy for exposed side

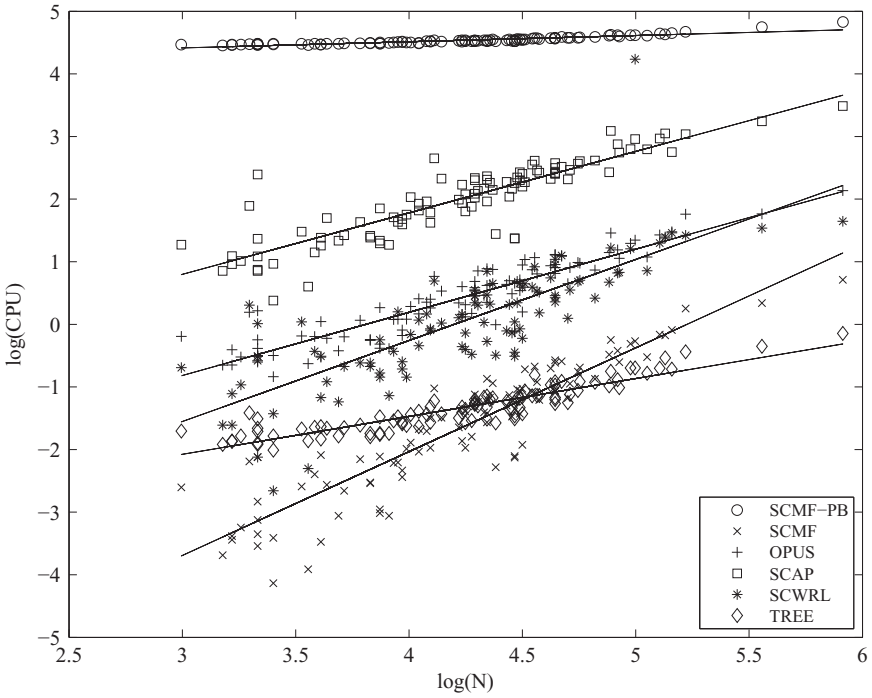

FIG. 5. CPU time required for predicting the conformations of all side chains of a protein versus the number of residues of the protein on a log-log scale, for different methods. The slope of the fitted lines are 0.1, 0.6, 0.98, 1.0, 1.3, and 1.6 for SCMF-PB, TREEPACK, SCAP, OPUS, SCWRL4, and SCMF, respectively.

chains. In addition, it is interesting to consider the dependence of the running time of the different algorithms with respect to the number N of residues of the test proteins (Fig. 5). As expected from their strategies for conformational search, OPUS-ROTA, SCAP, TREEPACK, and SCWRL4 have nearly linear behaviors with respect to N. The SCMF (without PB) has a nearly quadratic behavior: this is a consequence of the initialization step during which the matrix of side-chain–side-chain interactions is computed with a very large cutoff value (see implementation above). Interestingly, SCMF-PB has a very weak dependence on the number of residues in the protein: the bottleneck of this technique is the step that solves the PB equation twice, in presence and absence of solvent. This step is repeated at each cycle of the SCMF procedure and depends mostly on the size of the Cartesian grid considered for solving the PB equation numerically, and only marginally on the number of rotamers. This indicates that SCMF-PB could even be competitive with other techniques such as OPUS-ROTA for very large proteins.

V. CONCLUSION

Water plays a central role in biology as it helps define the structures and properties of biomolecules. Recent methods that incorporate solvent effects explicitly describe fine-scale structural characteristics of these molecules, albeit at a heavy computational costs. The formalism presented here combines SCMF theory for efficient sampling with the PB theory for modeling solvation with a physically sound approach. It is simple and its equations can be solved numerically with a reasonable computational cost. We have shown that its implementation for the problem of predicting the conformations of side chains of a protein is slower (one order of magnitude) than other competing programs such as SCWRL4 or OPUS-ROTA but not to the extend of been unusable in practice, as it predicts all side-chain conformations of a 370 residue proteins within 2 min CPU on a commoditiy PC. It does, however, improve the quality of prediction; the improvement is mostly observed for exposed residues, as intuitively expected.

We believe that the combination of SCMF theory and PB theory will prove useful in other modeling problems in biology. One important factor, for example, that has not been considered here involves the protonation states of the different side chains of the protein. For instance, the values for the χ_2 torsion angles of histidine residues are poorly predicted both in the core and at the surface of the proteins. The addition of the solvation free energy did not improve the prediction accuracy for histidine (in fact even lead to a decrease in accuracy, see Table I and Fig. 2), most likely because the protonation state of histidine was arbitrarily chosen in this work. We are working on extending SCMF-PB to predict simultaneously the conformations of all side chains of a protein and the protonation state of its titratable residues.

The approach presented here is not free of limitations. It is based on the Poisson-Boltzmann equation whose approximations have proven to be limited factors in many cases. Of particular relevance to this study, the PB model neglects any possible solvent structure, i.e., water is treated as a continuum with a uniform dielectric constant. However, close to the surface of a charged molecule the structure of water is perturbed as compared to bulk water, with an increased orientational ordering of its dipoles that leads to modified electrostatic interactions. The structuring of water molecules is usually coupled with an accumulation of counter-ions at the surface of the molecule. Both effects are strongly dependent on the sizes of the solvent molecules and ions because of excluded volume effects; these quantities, however, are ignored by the PB equation. In addition, the PB equation treats ions with a mean field model, ignoring possible correlations that are likely to be significant for non-monovalent ions at least. Modified PB equations have been derived to account for some of these effects (for reviews see Refs. 31 and 45). For example, several attempts have been made to define a structure for the solvent within the PB model by redefining it as an assembly of Langevin dipoles whose orientations and concentrations are defined self-consistently with respect to the local electrostatic potential.[54–57] Such a model has proved useful for characterizing dielectric permittivity profiles in the vicinity of charged surfaces[54,57] as well as for characterizing the hydration layer in the vicinity of proteins and nucleic acids.[45] We believe that the structuring of water molecules and ions in the neighborhood of the surface of a protein plays a role in defining the conformations of exposed side chains. We are, therefore, currently working on adapting the SCMF-PB formalism to incorporate these modified PB equations.

ACKNOWLEDGMENTS

P.K. acknowledges support from the National of Institutes of Health (NIH) under Contract No. GM080399. We are grateful to Antoine Koehl for reading this manuscript carefully.

[1] J. Ponder and F. Richards, J. Mol. Biol. **193**, 775 (1987).
[2] N. Pierce and E. Winfree, Prot. Eng. **15**, 779 (2002).
[3] B. Chazelle, C. Kinsfort, and M. Singh, INFORMS J. Comput. **16**, 380 (2004).
[4] J. Desmet, M. DeMaeyer, B. Hazes, and I. Lasters, Nature (London) **356**, 539 (1992).
[5] R. Goldstein, Biophys. J. **66**, 1335 (1994).
[6] D. Gordon and S. Mayo, J. Comput. Chem. **19**, 1505 (1998).
[7] L. Looger and H. Hellinga, J. Mol. Biol. **307**, 429 (2001).
[8] R. Peterson and P. Dutton, Protein Sci. **13**, 735 (2004).
[9] M. Lu, A. Dousis, and J. Ma, Protein Sci. **17**, 1576 (2008).
[10] Z. Xiang and B. Honig, J. Mol. Biol. **311**, 421 (2001).
[11] R. Samudrala and J. Moult, J. Mol. Biol. **279**, 287 (1998).
[12] A. Canutescu, A. Shelenkov, and R. Dunbrack, Protein Sci. **12**, 2001 (2003).
[13] K. Dukka-Bahadur, E. Tomita, J. Suzuki, and T. Akutsu, J. Bioinfo. Comput. Biol. **3**, 103 (2005).
[14] A. Marabotti, Curr. Chem. Biol. **2**, 200 (2008).
[15] P. Dada, R. Patel, and R. Doerksen, Curr. Top. Med. Chem. **10**, 84 (2010).
[16] G. Krivov, M. Shapovalov, and R. Dunbrack, Proteins: Struct., Funct., Bioinfo. **77**, 778 (2009).
[17] P. Koehl and M. Delarue, Curr. Opin. Struct. Biol. **6**, 222 (1996).
[18] P. Koehl and M. Delarue, J. Mol. Biol. **239**, 249 (1994).
[19] J. Mendes, C. Soares, and M. Carrondo, Biopolymers **50**, 111 (1999).
[20] R. Petrella, T. Lazaridis, and M. Karplus, Folding Des. **3**, 353 (1998).
[21] L. Jiang, B. Kuhlman, T. Kortemme, and D. Baker, Proteins: Struct., Funct., Bioinfo. **68**, 893 (2005).
[22] D. Eisenberg and A. D. McLachlan, Nature (London) **319**, 199 (1986).
[23] H. Edelsbrunner and P. Koehl, *Discrete and Computational Geometry* (MSRI Publications, Cambridge University Press, New York, 2005), Vol. 52, p. 243.
[24] A. G. Street and S. L. Mayo, Folding Des. **3**, 253 (1998).

[25]C. Wilson, L. Gregoret, and D. Agard, J. Mol. Biol. **229**, 996 (1993).
[26]J. Mendes, A. Baptista, M. Carrondo, and C. Soares, J. Comput.-Aided Mol. Des. **15**, 721 (2001).
[27]S. Liang and N. Grishin, Protein Sci. **11**, 322 (2002).
[28]Z. Ziang, P. Steinbach, M. Jacobson, R. Friesner, and B. Honig, Proteins: Struct., Funct., Bioinfo. **66**, 814 (2007).
[29]Z. Xiang, C. Soto, and B. Honig, Proc. Natl. Acad. Sci. USA **99**, 7432 (2002).
[30]N. Baker, Methods Enzymol. **383**, 94 (2004).
[31]P. Grochowski and J. Trylska, Biopolymers **89**, 93 (2008).
[32]P. Koehl, Curr. Opin. Struct. Biol. **16**, 142 (2006).
[33]S. Marshall, C. Vizcarra, and S. Mayo, Protein Sci. **14**, 1293 (2005).
[34]C. Vizcarra and S. Mayo, Curr. Opin. Chem. Biol. **9**, 622 (2005).
[35]C. Vizcarra, N. Zhang, S. Marshall, N. Wingreen, C. Zeng, and S. Mayo, J. Comput. Chem. **29**, 1153 (2008).
[36]A. Gutteridge and J. Thornton, Trends Biochem. Sci. **30**, 622 (2002).
[37]P. Barth, T. Alber, and P. Harbury, Proc. Natl. Acad. Sci. USA **104**, 4898 (2007).
[38]J. Xu and B. Berger, J. ACM **53**, 533 (2006).
[39]S. Nabuurs, A. Nederveen, W. Vranken, J. Doreleijers, A. Bonvin, G. Vuister, G. Vriend, and C. Spronk, Proteins: Struct., Funct., Bioinfo. **55**, 483 (2004).
[40]R. Kubo, *Statistical Mechanics. An Advanced Course with Problems and Solutions* (North-Holland, Amsterdam, 1965).
[41]K. Sharp and B. Honig, J. Phys. Chem. **94**, 7684 (1990).
[42]N. Baker, Methods Enzymol. **383**, 94 (2004).
[43]M. Davis and J. McCammon, J. Comput. Chem. **12**, 909 (1991).
[44]R. Bruccoleri, J. Comput. Chem. **14**, 1417 (1993).
[45]P. Koehl and M. Delarue, J. Chem. Phys. **132**, 064101 (2010).
[46]M. Levitt, J. Mol. Biol. **170**, 723 (1983).
[47]F. Momany, R. McGuire, A. Burgess, and H. Scheraga, J. Phys. Chem. **79**, 2361 (1975).
[48]M. Holst and F. Saied, J. Comput. Chem. **16**, 337 (1995).
[49]A. Finkelstein and B. Reva, Nature (London) **351**, 497 (1991).
[50]M. Jacobson, R. Friesner, Z. Xiang, and B. Honig, J. Mol. Biol. **320**, 597 (2002).
[51]P. Koehl and M. Delarue, Proteins: Struct., Funct., Genet. **20**, 264 (1994).
[52]O. Guvench and C. L. Brooks III, J. Am. Chem. Soc. **127**, 4668 (2005).
[53]M. Lu, A. Dousis, and J. Ma, J. Mol. Biol. **376**, 288 (2008).
[54]A. Abrashkin, D. Andelman, and H. Orland, Phys. Rev. Lett. **99**, 77801 (2007).
[55]C. Azuara, H. Orland, M. Bon, P. Koehl, and M. Delarue, Biophys. J. **95**, 5587 (2008).
[56]D. Mengistu, K. Bohinc, and S. May, Europhys. Lett. **88**, 14003 (2009).
[57]A. Iglic, E. Gongadze, and K. Bohinc, Bioelectrochemistry **79**, 223 (2010).