Computational Sidechain Placement and Protein Mutagenesis With Implicit Solvent Models

Anne Lopes,[1†] Alexey Alexandrov,[1†] Christine Bathelt,[1] Georgios Archontis,[2*] and Thomas Simonson[1*]

[1]*Laboratoire de Biochimie (UMR CNRS 7654), Department of Biology, Ecole Polytechnique, 91128, Palaiseau, France*

[2]*Department of Physics, University of Cyprus, Nicosia, Cyprus*

ABSTRACT **Structure prediction and computational protein design should benefit from accurate solvent models. We have applied implicit solvent models to two problems that are central to this area. First, we performed sidechain placement for 29 proteins, using a solvent model that combines a screened Coulomb term with an Accessible Surface Area term (CASA model). With optimized parameters, the prediction quality is comparable with earlier work that omitted electrostatics and solvation altogether. Second, we computed the stability changes associated with point mutations involving ionized sidechains. For over 1000 mutations, including many fully or partly buried positions, we compared CASA and two generalized Born models (GB) with a more accurate model, which solves the Poisson equation of continuum electrostatics numerically. CASA predicts the correct sign and order of magnitude of the stability change for 81% of the mutations, compared to 97% with the best GB. We also considered 140 mutations for which experimental data are available. Comparing to experiment requires additional assumptions about the unfolded protein structure, protein relaxation in response to the mutations, and contributions from the hydrophobic effect. With a simple, commonly-used unfolded state model, the mean unsigned error is 2.1 kcal/mol with both CASA and the best GB. Overall, the electrostatic model is not important for sidechain placement; CASA and GB are equivalent for surface mutations, while GB is far superior for fully or partly buried positions. Thus, for problems like protein design that involve all these aspects, the most recent GB models represent an important step forward. Along with the recent discovery of efficient, pairwise implementations of GB, this will open new possibilities for the computational engineering of proteins. Proteins 2007;67:853–867.** © 2007 Wiley-Liss, Inc.

Key words: structure prediction; solvation; mean field; Generalized Born; Poisson equation; protein design

INTRODUCTION

Homology-based structure prediction and computational protein design are areas whose importance is increasing with the development of structural genomics.[1–4] Both techniques usually rely on a simplified description of protein conformational space, taking into account one or a few fixed backbone conformations and a discrete set of sidechain rotamers.[5–8] The most stable structure within this discrete space can be found by exact or approximate search methods.[6,9–15] This paper focusses on another important ingredient: the energy or scoring function and, in particular, the treatment of aqueous solvent. We consider the performance of several implicit solvent models for two key problems that occur in computational protein design: sidechain placement and the calculation of stability changes due to point mutations. Both problems have been extensively studied, using a range of models.[13–20] However, they must be considered together if one is to parameterize and test solvent models for protein design; i.e., for searching sequence and conformational space simultaneously. For example, previous solutions of the sidechain placement problem that omit electrostatics[14,16,21] are not acceptable in this context. Such combined analyses are much less common.[15,19,20] Furthermore, with the rapid progress of implicit solvent models, especially generalized Born models, it is important to reconsider this problem.[22,23]

Indeed, aqueous solvent plays an important role in the structure and stability of proteins.[24] Structure prediction and protein design are done almost exclusively with "implicit" solvent models, for efficiency. The solvent degrees of freedom are not explicitly represented; rather, they are taken into account through their effect on the intraprotein interactions.[25] The energy function for the protein is referred to as a Potential of Mean Force (PMF), or "effective" energy function. To obtain correct energetics for the protein, the PMF for any given protein conformation should coincide with the Boltzmann

The Supplementary Material referred to in this article can be found at http://www.interscience.wiley.com/fpages/0887-3585/suppmat/

[†]These authors contributed equally.

*Correspondence to: Thomas Simonson, Laboratoire de Biochimie (UMR CNRS 7654), Department of Biology, Ecole Polytechnique, 91128, Palaiseau, France. E-mail: thomas.simonson@polytechnique.fr or Georgios Archontis, Department of Physics, University of Cyprus, Nicosia, Cyprus. E-mail: archonti@ucy.ac.cy

Received 4 August 2006; Revised 26 October 2006; Accepted 16 December 2006

Published online 8 March 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21379

average of the energy over the solvent configurations.[25] In practice, only approximate PMFs can be constructed.

An implicit solvent model that has a clear physical basis is the Poisson model, which treats the solvent as a dielectric continuum,[26–29] by numerically solving the Poisson equation (PE). The essential physical ingredients are (1) the strong, attractive interactions between charged protein groups and the surrounding, high-dielectric solvent, and (2) the large shielding of protein–protein electrostatic interactions by solvent. The solvent contribution to the PMF is obtained as the electrostatic free energy of a collection of point charges in a dielectric cavity.[25,29] The PE model provides good accuracy for many applications,[30] including small molecule solvation,[31,32] acid/base equilibria,[28,33] ligand binding,[34] protein–protein binding,[35,36] and protein dynamics.[37] Unfortunately, PE methods cannot be used routinely for computational protein design. Indeed, in continuum electrostatics, the effective interaction between two protein residues depends on the entire protein's shape and the complementary volume occupied by high-dielectric solvent. Therefore, continuum electrostatic energies are many-body quantities that cannot ordinarily be expressed as a sum over residue or atom pairs.[19,20,23,29,38] This is a prohibitive limitation for protein design.

A more efficient alternative is the generalized Born (GB) model.[22,30,39,40] GB is based on the same physical picture as PE, with a dielectric continuum solvent surrounding a protein cavity. But it makes additional approximations that allow an analytical expression of the PMF. It has become feasible to use GB in a protein design context, because residue–pairwise implementations were recently discovered.[19,23] Several GB variants and parameterizations exist.[41–46] GB has been used for many applications, including small molecule solvation,[42–44,47] protein solvation,[45,48] acid/base equilibria,[49,50] protein dynamics,[51,52] ligand binding,[53] protein folding,[54,55] loop structure prediction,[56] and scoring native folds vs. decoys.[57] The best variants have an accuracy that is not much inferior to explicit solvent models,[22,46,49] and further improvements can be expected.

A third, even simpler class of implicit solvent models are the so-called Accessible Surface Area models.[58–60] These models characterize different atom types by "atomic solvation parameters," which reflect their hydrophobicity or hydrophilicity, and consider the fraction of each atom's surface area that is accessible to solvent. Each atom contributes to the PMF through the product of its solvation parameter and its solvent accessible surface area. Usually, this accessible surface area contribution is supplemented by another electrostatic term, which attempts to capture the shielding of protein–protein electrostatic interactions by the high-dielectric solvent. The simplest approach is to add to the PMF a screened Coulomb energy, so that protein–protein electrostatic interactions are reduced by a constant factor ε. We refer to this as the Coulomb/Accessible Surface Area (CASA) model. This class of models has been used for protein molecular dynamics,[60–62] structure prediction,[14] and protein–ligand binding.[63] It is routinely used for computational protein design.[4,64–67]

With the ongoing development of GB models, and with the discovery of methods to implement them efficiently in computational protein design,[19,23] it is important to systematically compare the behavior of the PE, GB, and CASA models for protein structure prediction and design. In the context of protein design, structure prediction is usually limited to sidechain placement with a fixed protein backbone. Many authors have analyzed this problem and shown that good results are obtained with very simple models, without any electrostatic or solvent treatment.[14,16,21] In computational protein design, however, sidechain placement must be done repeatedly, following rounds of random mutagenesis. The mutations will often modify the protein charge, frequently introducing charged sidechains into buried positions. For these effects, an accurate electrostatic treatment is desirable. Therefore, an integrated treatment of sidechain placement and ionized mutations is needed. Previous solutions of the placement problem that omit electrostatics are not acceptable. Many authors have studied the accuracy of PE and GB models for protein electrostatics.[18,29,30,68] However, very few studies have considered sidechain placement and ionized mutations simultaneously. Even fewer (none that we are aware of) have compared surface area models (CASA) to GB in this context. While CASA is the most common solvent model in computational protein design, GB models have greatly progressed in the last few years, so that a reevaluation and comparison of models is needed. For example, one recent study considered a suboptimal GB/ACE parameterization.[20]

In this paper, we consider both sidechain placement and the effect of amino acid mutations on protein stability. For sidechain placement, we use the CASA model. In a related test, we use CASA, GB, and PE to estimate the stability of large libraries of protein conformations, where the sidechain positions have been randomized. We then consider over 1000 mutations involving ionized sidechains, and use CASA, GB, and PE to compute the corresponding stability changes. The data set includes 140 mutations in 12 proteins for which experimental measurements are available. The experimental data span only a limited free energy range, and they do not include any mutations that introduce or delete ionized sidechains in the protein core, since these are usually not experimentally tractable.[69] The other mutations are mostly in buried positions and correspond to much larger stability changes, but no experimental measurements are available. For computational reasons, most of these mutations are not "biochemically exact": they consist in charge modifications, as opposed to real mutations. For example, negative charges are introduced onto valine sidechain methyls, roughly mimicking a valine → aspartate mutation. Nevertheless, they contain the relevant physical effects and represent a valid test of the solvent models. We refer to these two data sets as the "experimental" and "artificial" mutations, respectively.

An important aspect of this work is the parameterization of the CASA and GB models. With CASA, for example, we consider three sets of atomic surface parameters, a wide range of dielectric constants, and a range of scaling factors for various model terms. We consider two GB models: GB/ACE,[42] in combination with the Charmm19 force field,[70] and GB/HCT,[71] in combination with the AMBER force field.[72] A limited parameter optimization is done for GB/ACE and a more extensive one for GB/HCT. An essential point is that the parameterization must be useful for both structure prediction and stability changes. The goal of the GB/HCT parameter optimization is to take into account very recent improvements in the set of atomic radii used for continuum electrostatics calculations with the AMBER atomic charges.[73]

The "experimental" mutations are not used in the parameterization, for several reasons. First, they correspond to very small stability changes. More importantly, they cannot be computed without assuming a specific model for the unfolded protein's structure. Thus, for a rigorous optimization, the unfolded state model and the energy parameters would have to be tested and varied simultaneously. We prefer to optimize the energy function separately; therefore, we use only the artifical mutations for parameter optimization. Since experimental measurements are not available, we take as reference values the PE results. Indeed, PE has been extensively used to study protein electrostatics.[29,68] Many recent studies have shown that it is a valid reference for the parameterization and testing of implicit solvent models.[17–19] With this strategy, the unfolded state treatment does not influence the parameterization, since the same treatment is used for CASA, GB, and PE.

Once the energy parameters are chosen, we can use the experimental mutations for a blind test of the CASA and GB models. We use a very simple, tripeptide representation of the unfolded state, with noninteracting amino acids, which is commonly used in protein design. With CASA, the mean unsigned deviation from experiment is 2.1 kcal/mol. This appears comparable to the accuracy reported by Serrano et al.[74] for a comparable set of mutations, although they did not describe their subset of charged mutations separately. Remarkably, it is similar to the accuracy of molecular dynamics free energy simulations using explicit solvent models,[49,75] which are much more difficult and expensive (but which give additional structural and dynamical information). With GB/ACE and GB/HCT, the mean unsigned deviations from experiment are 2.9 and 3.4 kcal/mol, respectively. If a surface term is included in the GB/HCT model, to help represent dispersion and hydrophobic contributions,[32] the GB/HCT mean unsigned error drops to 2.1 kcal/mol, the same level as CASA.

In summary, we find that (a) the choice of electrostatic model is not very important for the sidechain placement calculations (confirming several earlier studies[14,16,21]); (b) GB/HCT and CASA give the same accuracy for the experimental mutations; (c) GB/HCT yields an enormous improvement for the total solvation free energies and for the artificial mutations. Thus, for problems like protein design that involve all these aspects, the most recent GB models represent an important step forward. Along with the recent discovery of efficient implementations of GB in protein design,[19,23] this will open new possibilities for the computational engineering of new proteins.

MATERIALS AND METHODS

Effective Energy Function

We tested several effective energy functions, or PMFs, which have the form:

$$E = E_{\text{bonds}} + E_{\text{angles}} + E_{\text{dihe}} + E_{\text{impr}} + E_{\text{vdw}} + E_{\text{coul}} + E_{\text{solv}}. \tag{1}$$

The first six terms in Eq. (1) represent the protein internal energy. They are taken from either the CHARMM19 or the AMBER empirical energy function.[70,72] They represent a covalent bond energy, a bond angle energy, a torsion energy associated with sidechain dihedrals, a term maintaining the chirality or planarity of selected atomic centers, a van der Waals energy, and a Coulomb electrostatic energy,

$$E_{\text{coul}} = \sum_{i<j} \frac{q_i q_j}{r_{ij}}, \tag{2}$$

where the sum is over all pairs of protein atoms i, j, of charges q_i, q_j, and r_{ij} is the distance between a pair. The last term on the right of Eq. (1), E_{solv}, represents the contribution of solvent.[25] In this work, we compare three different solvent treatments, described below.

CASA Solvent Treatment

Our first solvent treatment is an accessible surface area treatment: the CASA model. E_{solv} includes two energy terms that describe protein–solvent electrostatic and hydrophobic interactions:[25]

$$E_{\text{solv}} \equiv E_{\text{CASA}} = E_{\text{screen}} + E_{\text{surf}} = \left(\frac{1}{\varepsilon} - 1\right) E_{\text{coul}} + \alpha \sum_i A_i \sigma_i. \tag{3}$$

E_{screen} is a screened Coulomb energy; ε is the dielectric constant of the medium (relative to vacuum). Notice that $E_{\text{coul}} + E_{\text{screen}} = E_{\text{coul}}/\varepsilon$. E_{surf} is related to the atomic solvent-accessible surface areas. The sum is over all protein atoms i; A_i is the solvent-accessible area of atom i; σ_i is an atomic solvation parameter (measured in kcal/mol/Å^2); and α is an overall weight for the surface energy term. The coefficients σ_i reflect the preference of particular atom types to be exposed or hidden from solvent, and incorporate both electrostatic and nonelectrostatic effects.[25] Surface areas were computed by the Lee and Richards algorithm,[76] implemented in XPLOR,[77] using a

1.4 Å probe radius. Three different sets of atomic solvation parameters were tested,[59–61] along with different values of the dielectric constant ε and the weight α.

GB Solvent Treatment

Our second solvent treatment is a GB model[39,41,42]:

$$E_{\mathrm{solv}} \equiv E_{\mathrm{GB}} = \sum_i \Delta G_i^{\mathrm{self}} + \sum_{i<j} \Delta G_{ij}^{\mathrm{screen}}$$
$$= \tau \sum_i \frac{q_i^2}{2b_i} + \tau \sum_{i<j} \frac{q_i q_j}{\sqrt{r_{ij}^2 + b_i b_j exp[-r_{ij}^2/(4b_i b_j)]}} \qquad (4)$$

where $\tau = 1/\varepsilon_w - 1/\varepsilon_p$, r_{ij} is the distance between charges q_i and q_j, b_i is the effective Born radius of atom i, ε_w is the dielectric constant of water (set to 80), and ε_p is the protein dielectric constant (set to 1 unless otherwise mentioned). The first term is a sum of atomic self-energies $\Delta G_i{}^{\mathrm{self}} = \tau q_i{}^2/2b_i$, corresponding to the interaction of each atomic charge q_i with its own reaction field in the environment of the solvated biomolecule. The second term models the interaction of a charge q_i with the reaction field produced by a different charge q_j, and accounts for the screening of electrostatic interactions by the high-dielectric solvent.[39]

The self-energy term in Eq. (4) requires the calculation of an electrostatic energy density integral over the solute volume. We use both the GB/ACE and the GB/HCT models, which assume a Coulomb functional form for the electric field inside the solute and partition the volume into atomic contributions.[42,71] The GB/HCT parameters are described further on. The GB/ACE atomic volumes corresponded to the Voronoi database V01,[78] scaled by a factor of 0.8 as described in Ref. 52. The smoothing parameter that controls the width of atomic volumes[42] in GB/ACE was set to 1.3. The partial charges corresponded to the CHARMM19 energy function.[70] The calculations were performed with the XPLOR program.[77,79]

Poisson Equation Solvent Treatment

Our third solvent treatment is a continuum electrostatic model with numerical solution of the Poisson equation. We refer to it as the PE model. The solvation energy has the form

$$E_{\mathrm{solv}} \equiv E_{\mathrm{PE}} = \frac{1}{2} \sum_i q_i V_i^{\mathrm{reac}} \qquad (5)$$

where q_i is the charge of atom i and V_i^{reac} is the electrostatic potential on atom i due to the polarization charge at the protein–solvent interface, induced by every atomic charge. For a given protein structure, we computed V_i^{reac} by solving the Poisson equation numerically for the protein in solution and in the gas phase and taking the difference. The finite-difference program UHBD was used.[80] The protein–solvent dielectric boundary was defined by the molecular surface of the protein. The solution used a two-step focussing procedure and a cubic grid with spacings of 0.8 and 0.4 Å. The molecular surface was constructed with 2000 points per atom, using a probe sphere of radius 2 Å and the boundary smoothing method in UHBD.[80] Small voids in the interior of the rotameric structures were filled by dummy atoms, to prevent the occurence of artificial, high-dielectric internal cavities (sometimes overlooked in PE applications). Two PE parameterizations are used. The first uses Charmm19 atomic radii and charges,[70] except for hydrogen radii, which were set to 1 Å. The second uses atomic charges from the AMBER, all-atom force field,[72] along with atomic radii specifically and carefully optimized for PE with AMBER charges.[73]

Protein Set and Rotamer Library

Twenty-nine proteins were used for sidechain placement calculations; they are listed in Table III. Their sizes varied from 36 to 212 amino acids, with a mean of 110. Their structures were all determined using X-ray crystallography with a resolution of 1.8 Å or better. The rotamer construction and energy calculations were performed with the XPLOR program.[77] The backbone and C_β atoms of the X-ray structure were fixed during the construction. The sidechain atoms were geometrically constructed from the position of the N, C_α, C, and C_β atoms, using standard bond lengths and bond angles from the CHARMM19 force field. Sidechain dihedral angles were taken from the Tuffery rotamer library.[6]

Additional, solvation energy calculations were performed for the proteins BPTI (4PTI), lysozyme (1LZ1), thioredoxin (2TRX), and ubiquitin (1UBQ). We created a set of 750–1000 random structures for each one by holding the backbone in its native conformation and randomizing the orientation of all sidechains except Pro, Ala, and Cys.

Mean Field Optimization

We did sidechain placement using a mean field approximation.[14] This method calculates iteratively the Boltzmann probability $P(i,k)$ of each rotamer k of each residue i, which is related to the mean energy $E(i,k)$ of sidechain i:

$$E(i,k) = -RT \ln P(i,k). \qquad (6)$$

$E(i,k)$ is the Boltzmann average of the interaction energy between sidechain i and its environment; R is the ideal gas constant, and T is the temperature. Since the protein backbone is fixed, we can write

$$E(i,k) = E_{BB}(i,k) + \sum_{j \neq i} \sum_l E(ik, jl) P(j,l) \qquad (7)$$

where E_{BB} is the interaction energy with the backbone, the first sum is over protein sidechains j, the second sum is over the rotamers l of sidechain j, and $E(ik,jl)$ is the interaction energy between sidechains i and j when they occupy rotamers k and l. We assume the optimal sidechain positions correspond to the most probable rotamers.

The probabilities were computed iteratively. The current estimate $P(i,k)^{(n)}$ was updated according to

$$P(i,k)^{(n+1)} = \lambda P(i,k)^{(n)} + (1-\lambda)P(i,k)^{(n-1)}. \tag{8}$$

Different λ values were tested. The best results (reported below) were obtained with $\lambda = 0.35$ and uniform starting rotamer probabilities. About 20 cycles were typically needed for convergence.

An Approximate, Pairwise Surface Area Calculation

In the sidechain placement, the solvent accessible surfaces are calculated by an approximate, but very efficient procedure. The buried surface of a sidechain is computed by summing over the neighboring sidechain and backbone groups. For each neighboring group, the contact area with the sidechain of interest is computed, independently of other surrounding groups. The contact areas are then summed. This approach assumes the contact areas between sidechains are independent, and the total area of a sidechain is a pairwise sum over its neighbors. In fact, surface area buried by one sidechain may also be buried by another. Mayo et al.[81] showed that this approach overestimated the surface areas of buried sidechains, but had little effect on solvent exposed sidechains. Therefore, we performed most of the calculations with a scaling factor applied to the contact areas involving buried sidechains. To determine the optimal scaling factor, following Mayo et al.,[81] we considered the total surface area that is buried within the protein, A_{buried}. A_{buried} is defined as the difference between the total area of all residues, taken separately, and the total surface area of the protein structure; see Ref. 81 for details. With the pairwise approximation, it is computed as

$$\begin{aligned} A_{\text{buried}} &\approx A_{\text{buried}}^{\text{pairwise}} \\ &= \sum_i A^0_{i_r t3} - \left(\sum_i A_{i_r t} + \frac{1}{2}\sum_{ij} s_i (A_{i_r j_s t} - A_{i_r t} - A_{j_s t}) \right). \end{aligned} \tag{9}$$

The first sum on the right is over all residues i and represents the total area of all residues, taken separately. $A^0_{i_r t3}$ is the exposed area of sidechain i when it occupies the rotamer r, in the presence of just the backbone of residues $i-1$, i, $i+1$ (a tripeptide indicated by the subscript $t3$). The sums in parentheses represent the approximate total protein surface area. $A_{i_r}t$ is the exposed area of the sidechain and its backbone, in the presence of the entire protein backbone. $A_{i_r} j_s t$ is the exposed area of the sidechain pair ij in the presence of the entire protein backbone. Finally, s_i is the scaling parameter, which is set to 1 if residue i is exposed to solvent and to a smaller value $s < 1$ if it is buried. To validate this approximation and determine the optimal s, we performed surface area calculations on the 29 test proteins above, and compared the exact and approximate buried surface areas. Using a scaling factor of $s_i = 0.5$ for buried amino acids [Eq. (9)], we observed a correlation of 0.999 between the approximate and the true surfaces, with a slope of 0.993 and an offset of 73.3 Å^2. The RMSD between the approximate and true surfaces was just 199 Å^2, 1% of the mean surface. Therefore, a value of $s_i = 0.5$ was used for all the sidechain placements, unless otherwise mentioned.

Validation Methods

Two criteria were used to assess the accuracy of the sidechain placements. First, we calculated the percentage of sidechain dihedral angles within 40° of the value in the crystal structure. Second, we computed the RMSD between the sidechain atom positions in the model and the X-ray structure, excluding the C_β. The RMSD calculation took into account the rotational symmetry axis of Asp, Phe, Glu, and Tyr residues.

Later in this article, we compare protocols that use several different effective energy functions, corresponding to different solvent treatments. To determine the significance of the differences between protocols, statistical tests were done. Analysis of variance (ANOVA) was performed to evaluate groups of protocols. The null hypothesis is that for an observable of interest (e.g., RMSD), all the means observed for the various protocols are identical. Rejection of the null hypothesis means that there is at least one pair of means that are different from each other. To identify this pair, we then performed Student's t-tests on all pairs of means.

Charge Mutations

To test the effective energy functions in a situation that mimics protein design calculations, we performed two sets of charge mutations. In the first set, the atomic charges of selected sidechains were modified artificially. These mutations can be classified into three types: (1) charge deletions, which remove the net charge on Arg, Lys, Asp, Glu, and doubly-protonated His; (2) charge insertions, which add a ±1 charge to Ala, Ile, Leu, Val, Met, Pro, Thr, or Tyr; (3) polarity changes, which either make Asn, Gln, or singly-protonated His apolar, or introduce a dipole onto the Cys sidechain. The details of the charge modifications are given in Supplementary Material. While these charge transformations do not correspond to real amino acid mutations, they pose the same physical problems and represent a realistic test of the methodology. They were performed with the CASA, GB, and PE solvent models. Experimental data are not available for this set, so we take PE as a reference, since it is commonly judged to be the most realistic of the three solvent models.[25,29,40] The PE model uses a protein dielectric constant of one, because we wish to model solvent relaxation, not relaxation of the protein. The CASA model employed the Fraternali atomic solvation parameters:[61] a positive value of 0.0119 kcal/mol/Å^2 for carbon atoms and a negative value of -0.0598 kcal/mol/Å^2 for nitrogen and oxygen atoms. To optimize further the model for charge mutations, we explored several values of the CASA dielectric constant, and we introduced a new solvation parameter for charged atoms, by the following scheme. In charge deletions, the atoms involved have a large charge in their native state and a small one in the mutated state. These atoms were assigned a native-state surface coefficient σ_C (to be determined), and a mutant-state coefficient of 0.0119 kcal/mol/Å^2, reflecting their hydrophobic character after the mutation. The optimum σ_C value was determined by comparison with free energy differences evaluated by Poisson calculations (PE solvent model). In charge insertions, most of the atoms involved are carbons, with a zero initial charge and a large (positive or negative) final charge. For these mutations, the coefficients of all the relevant atoms were set to their chemical-type value at the native state, and to a value of σ_C in the mutant state. In polar-to-neutral mutations, the atoms involved were assigned a native-state coefficient according to their chemical type, and a mutant-state coefficient of -0.0598 kcal/mol/Å^2 (Cys) or 0.0119 kcal/mol/Å^2 (Asn, Gln, His). In all cases, hydrogen atoms were assigned a zero atomic surface coefficient in both states. A list of atom types associated with the new coefficient σ_C is given in Supplementary Material.

The second data set included 140 mutations in 12 proteins for which the experimental stability changes are known (listed in Supplementary Material). In each case, an ionized sidechain is introduced or removed. Each native protein structure is first energy minimized with its backbone fixed. The new sidechain is then positioned successively in each of its rotamers, minimized with the surrounding protein structure fixed, and the best rotamer is retained. To obtain the difference in stability between the mutant and the native protein, the charge modification must also be introduced in the unfolded protein. By subtracting the wildtype/mutant energy differences in the folded and unfolded states, the stability of the mutant relative to the native protein is obtained (see Fig. 1). In our calculations, the "unfolded" state corresponded to the mutated residue, isolated from the rest of the sequence (and with the same coordinates as in the native structure). This simple unfolded model is commonly used in protein design.[65–67] Mutations were performed using the CASA, GB/ACE, GB/HCT, and PE solvent treaments.

Optimization of the GB/HCT Parameters

With the GB/HCT model, each atom is assigned a volume and a scaling factor.[48] For the proteins considered here, there are 29 different atom types in the AMBER protein force field,[72] giving 58 atomic parameters. These were adjusted by an iterative, least-squares procedure to reproduce the stability changes computed with PE for artificial mutations, similar to those described above. 1020 of the mutations were chosen randomly to be optimized; the other 7130 were used for a cross-validation test of the optimization. The PE model employed the AMBER atomic charges and atomic radii optimized very recently by McCammon and coworkers[73] to reproduce a large body of experimental and simulation data. The optimized parameters are given in Supplementary Material.

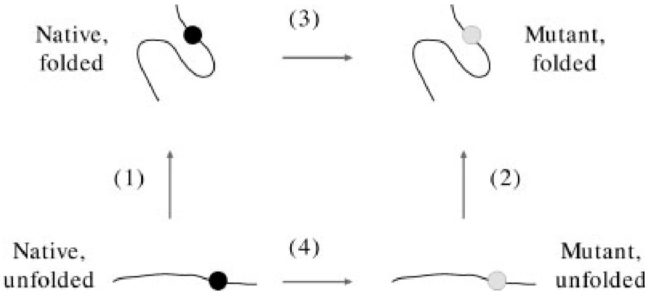

Fig. 1. Thermodynamic cycle used to calculate the change in protein stability due to charge mutations.

RESULTS

Sidechain Placement

We first describe sidechain placement with the CASA solvent model [Eq. (3)]. Twenty-nine proteins were used as a test set (Table III). Several parameterizations were compared in order to optimize the model.

Choice of the dielectric constant

We first considered the effect of the dielectric constant ε and the Coulomb energy term [Eq. (3)], using the CHARMM19 force field and three different sets of atomic solvation parameters, referred to as the Wesson, Ooi, and Fraternali sets, respectively.[59–61] The quality of the predicted structures was assessed by the agreement of sidechain dihedrals with the crystal structure and by the comparison of RMSD of the sidechain atoms with the crystal structure. Results for $\varepsilon = 4$ and $\varepsilon = 20$ are shown in Table I.

The dihedral prediction and sidechain RMSD values are comparable to those of Koehl and Delarue and Yang et al.[14,82] We compared the models with the two ε values and the three solvation parameter sets using the ANOVA statistical test. Differences of χ_1 predictions among the six models are statistically significant at the 1% significance level. The same is true for the predictions of both χ_1 and χ_2 (χ_{1+2}) and for the sidechain RMSD.

Student's test was then done for pairs of models with different ε values. For the Fraternali parameters, differences of the χ_1, χ_{1+2}, and RMSD predictions with the two ε values are not significant at the 1% significance level. The same is true for the Ooi and Wesson parameter sets. Thus, the difference detected by ANOVA is not

TABLE I. CASA Sidechain Predictions with Different Dielectric Constants ε

	Atomic parameters	χ_1 Mean (sd)	χ_{1+2} Mean (sd)	RMSD (Å) Mean (sd)	RMSD, no C_β (Å) Mean (sd)
$\varepsilon = 20$	Fraternali	0.76 (7)	0.61 (6)	1.85 (32)	1.62 (28)
	Wesson	0.71 (7)	0.56 (6)	2.05 (31)	1.77 (26)
	Ooi	0.70 (7)	0.54 (7)	2.12 (37)	1.85 (33)
$\varepsilon = 4$	Fraternali	0.77 (6)	0.63 (6)	1.85 (31)	1.61 (28)
	Wesson	0.72 (6)	0.58 (7)	1.97 (32)	1.73 (29)
	Ooi	0.69 (6)	0.55 (7)	2.13 (32)	1.86 (29)

Fraction of successful dihedral predictions and sidechain RMSD relative to crystal structure, averaged over 29 test proteins. Standard deviation in parentheses (in units of last digit). Rightmost column omits the C_β from RMSD.

TABLE II. CASA Sidechain Predictions with Different Surface Weights α

	Atomic parameters	χ_1 Mean (sd)	χ_{1+2} Mean (sd)	RMSD (Å) Mean (sd)	RMSD, no C_β (Å) Mean (sd)
$\alpha = 1/2$	Fraternali	0.79 (5)	0.66 (6)	1.79 (30)	1.57 (26)
	Wesson	0.76 (6)	0.62 (7)	1.86 (30)	1.63 (27)
	Ooi	0.74 (6)	0.62 (8)	1.94 (34)	1.70 (29)
$\alpha = 1/3$	Fraternali	0.80 (6)	0.67 (6)	1.75 (33)	1.53 (30)
	Wesson	0.78 (6)	0.65 (6)	1.82 (30)	1.59 (27)
	Ooi	0.76 (6)	0.64 (8)	1.88 (30)	1.65 (27)
	Fraternali[a]	0.80 (6)	0.66 (6)	1.74 (33)	1.52 (30)
$\alpha = 0$	none	0.84 (5)	0.68 (6)	1.68 (29)	1.46 (24)

Fraction of successful dihedral predictions and sidechain RMSD relative to crystal structure, averaged over 29 test proteins. Standard deviation in parentheses (in units of last digit). Rightmost column omits the C_β from RMSD. Dielectric constant is $\varepsilon = 4$.

[a] The surface areas of buried sidechains are not downscaled [$s_i = 1$ in Eq. (9)].

related to the choice of ε, which does not affect sidechain prediction. This is consistent with several earlier studies, in which good results were obtained without any electrostatic term.[14,16,21] As pointed out above, however, a model without electrostatics is not appropriate here, since our goal is to do protein design. All the following calculations include the Coulomb energy term with $\varepsilon = 4$.

Choice of the atomic solvation parameters and the weight α

Next, statistical tests were done to determine whether the three solvation parameter sets were significantly different (using $\varepsilon = 4$). Results are summarized in Table II. ANOVA indicates that the null hypothesis is rejected at the 1% significance level for the three quality criteria considered (χ_1, χ_{1+2}, RMSD). Thus, there is at least one prediction protocol that differs. Results with the Ooi and Wesson parameters are equivalent according to the Student test at the 5% significance level for all three quality criteria. However, differences in the χ_1 predictions and the RMSD calculations between Fraternali and the two other parameter sets are statistically significant at the 5% significance level, and differences for χ_{1+2} are significant at the 1% level. Thus, the Fraternali parameters perform better than the other two sets. The Fraternali protocol gave an average success rate for χ_1 and χ_{1+2} predictions of 77% and 63%, respectively.

We next considered a model with $\varepsilon = 4$ but no surface area term [$\alpha = 0$ in Eq. (3)]. The results are actually improved (see Table II). Indeed, the average of the χ_1 and χ_{1+2} predictions are about 84% and 68%, somewhat better than those of Koehl and Delarue and Yang et al.[14,82] It appears that the fully-weighted surface area term is too large, compared to the other terms in the energy function. Nevertheless, further calculations were done with a nonzero α. Indeed, we show below that $\alpha = 1/2$ gives much better results for experimental mutations and their associated stability changes. For protein design, with sidechain reconstruction and mutagenesis occurring simultaneously, we need a consensus model that performs well at both tasks.

Further calculations were therefore done with $\alpha = 1/2$ and $\alpha = 1/3$ (Table II). Results improve with decreasing α. Successful χ_1 prediction with the Fraternali parameters increases to 80% with $\alpha = 1/3$, while χ_{1+2} increases to 67%. These percentages approach those obtained with $\alpha = 0$. The Ooi and Wesson parameters give somewhat poorer results.

We performed ANOVA for each parameter set to compare the different α values (1/3, 1/2, 1). For the Fraternali set, for example, the three protocols were not equivalent according to the χ_1 and χ_{1+2} criteria at the 1% significance level. The $\alpha = 1/2$ and $\alpha = 1/3$ Fraternali models are not significantly different according to the Student's t-test; both are superior to the $\alpha = 1$ model.

Overall, the different Student's tests show that Fraternali with $\alpha = 0$ to 1/2 and Wesson with $\alpha = 1/3$ are equivalent and provide the best results. Table III compares Fraternali, $\alpha = 1/3$ with the work of Koehl and Delarue and Yang et al.[14,82]

TABLE III. CASA Sidechain Predictions Compared to Selected Earlier Work

		CASA, this work (Fraternali, $\varepsilon = 4$, $\alpha = 1/3$)				Koehl and Delarue			Yang et al.		
PDB code	Chain length	χ_1	χ_{1+2}	RMSD[a]	RMSD[b]	χ_1	χ_{1+2}	RMSD[a]	χ_1	χ_{1+2}	RMSD[b]
1PPT	36	0.73	0.64	1.59	1.40	0.73	0.67	1.57	–	–	–
1CRN	46	0.89	0.76	0.88	0.75	0.78	0.68	1.42	0.95	0.84	0.84
2OVO	56	0.90	0.79	1.30	1.12	0.69	0.60	2.35	–	–	–
4PTI	58	0.78	0.70	2.05	1.81	0.87	0.74	1.85	0.80	0.65	1.26
1IGD	61	0.84	0.72	1.22	1.06	–	–	–	0.76	0.74	1.28
1ISU	62	0.95	0.79	1.36	1.19	–	–	–	0.84	0.74	1.12
2SN3	65	0.81	0.68	2.31	2.03	0.59	0.45	2.76	–	–	–
1PTX	68	0.82	0.74	2.16	1.92	–	–	–	0.74	0.61	1.73
1UBQ	76	0.75	0.57	2.19	1.90	0.74	0.57	2.06	–	–	–
1PLC	99	0.77	0.66	1.40	1.21	0.76	0.67	1.55	0.82	0.70	1.24
1LN4	104	0.73	0.62	1.90	1.64	–	–	–	–	–	–
1AAC	105	0.81	0.68	1.81	1.58	–	–	–	0.92	0.69	1.05
256B	106	0.80	0.67	2.01	1.75	–	–	–	0.73	0.53	1.53
2TRX	108	0.82	0.61	1.88	1.65	–	–	–	–	–	–
5CPV	109	0.78	0.61	1.69	1.48	0.68	0.54	1.99	–	–	–
1CCR	112	0.75	0.64	1.86	1.62	–	–	–	0.84	0.60	1.21
1THX	115	0.76	0.65	1.53	1.32	–	–	–	–	–	–
1WHI	122	0.77	0.67	1.92	1.66	–	–	–	0.73	0.65	1.52
1PMY	123	0.80	0.63	1.95	1.70	–	–	–	0.83	0.63	1.23
3RN3	124	0.68	0.60	2.03	1.76	0.66	0.58	2.24	–	–	–
1LZ1	130	0.85	0.70	1.66	1.47	0.80	0.73	1.55	0.82	0.66	1.16
2END	137	0.79	0.63	1.89	1.66	–	–	–	0.79	0.64	1.37
2FOX	138	0.79	0.69	1.46	1.27	–	–	–	0.74	0.56	1.35
2HBG	147	0.77	0.61	1.79	1.57	–	–	–	0.76	0.61	1.34
2RN2	155	0.81	0.67	2.01	1.77	–	–	–	–	–	–
2CPL	165	0.86	0.71	1.49	1.31	–	–	–	–	–	–
1KOE	172	0.73	0.63	1.85	1.61	–	–	–	–	–	–
1XNB	185	0.81	0.70	1.58	1.39	–	–	–	0.78	0.66	1.89
1ES9	212	0.78	0.61	1.92	1.68	–	–	–	–	–	–
Mean	110	0.80	0.67	1.74	1.53	0.72	0.62	1.89	0.80	0.66	1.36

Fraction of successful dihedral predictions and sidechain RMSD (Å) relative to crystal structure.

[a] C_β excluded.

[b] C_β included.

To test for a possible force-field dependency of these results, calculations with the AMBER force field[72] were performed for six of the proteins. The accuracy of sidechain prediction is equivalent to that observed with the CHARMM19 force field, with the $\varepsilon = 4$, $\alpha = 1/3$, Fraternali protocol (data not shown). For example, the average χ_1 prediction rate is 75% with both AMBER and CHARMM19 for these six proteins.

The rate of successful prediction was analyzed as a function of amino acid type. In agreement with earlier work, it is easier to predict the rotamers of hydrophobic residues. Their rate of successful χ_1 prediction is 92% with the best parameterization, compared to 71% for hydrophilic residues. This can be explained because hydrophobic residues are mainly located in the protein core and are constrained by van der Waals packing interactions. There is little difference in accuracy between large and small residues. A low χ_{1+2} accuracy is observed for Asn, Gln, and His, since these residues' χ_2 angle can often be flipped by 180° with only a small change in energy.

Solvation Energies: Comparing Solvent Treatments

To directly compare the three solvent models, CASA, GB, and PE, we computed Coulomb and solvation energies for four proteins: trypsin inhibitor (4PTI), thioredoxin (2TRX), lysozyme (1LZ1), and ubiquitin (1UBQ). For each protein, we generated 750–1000 structures by randomizing sidechain rotamers. In the CASA energies, the Coulomb term is divided by a dielectric constant ε [Eq. (3)]. We initially searched a range of values, $\varepsilon = 1$–20, to find the optimum for each protein.

The RMSDs between PE and CASA are listed in Table IV. The optimal dielectric constants are fairly uniform among proteins: $\varepsilon = 1.5$–2.5. With these ε values, the PE/CASA RMSD ranges from 33 kcal/mol (4PTI) to 58 kcal/mol (2TRX). Ultimately, we require a consensus CASA model that performs well for sidechain placement, solvation energies, and charge mutations, and for all proteins. Sidechain placement (above) is insensitive to ε, while charge mutations (below) work best with a large value, $\varepsilon = 16$–20. Table IV also reports the CASA solva-

TABLE IV. Protein Solvation: RMS Deviation Between the CASA and GB Energies and the PE Reference Values

	CASA dielectric constant ε								
Protein	1	1.5	2.0	2.5	3.0	4.0	20.0	GB/ACE	GB/HCT
4PTI	46.9	**32.5**	32.8	35.7	38.5	42.9	55.6	71.7	20.7
2TRX	87.7	64.8	59.1	**58.3**	59.0	61.1	71.0	99.6	30.8
1LZ1	80.6	59.1	**56.2**	57.7	60.0	64.1	78.3	163.5	44.7
1UBQ	70.1	52.0	**49.8**	51.3	53.4	57.1	69.6	102.1	26.4

All values in kcal/mol. For each protein, the deviations are evaluated over 750–1000 rotameric structures. Best CASA results in boldface.

TABLE V. Comparison Between the CASA and Poisson Models for Charge Mutations

Protein	No of mutations	Optimal σ_C (kcal/mol/Å^2)	Optimal ε	PE-CASA RMSD[a] (kcal/mol)	Success[b] rate (%)
AspRS	518	−0.15 (−0.20)	16	17.5 (17.9)	80.1 (80.1)
3RN3	106	−0.20	16	13.1	79.0
4PTI	51	−0.20	16	11.0	74.4
5CYT	84	−0.20	16	13.5	79.8
2TRX	94	−0.20	16	12.3	80.8
1LZ1	108	−0.20	16	12.2	83.4
1UBQ	67	−0.20	16	10.7	86.5

Proteins are indicated by their PDB code, except AspRS, which is *Escherichia coli* aspartyl-tRNA synthetase. The surface coefficient σ_C is associated with atoms on charged sidechains (see text).

[a] RMS deviation between the surface area (CASA) and Poisson (PE) energy differences.

[b] Percentage of mutations that are predicted to have a positive or negative stability change by both CASA and PE.

tion results with $\varepsilon = 20$. The RMS deviations from PE increase to 56—71 kcal/mol. Thus, the loss in performance is significant when one takes the charge mutation dielectric as a consensus value for all proteins.

The CASA, GB/ACE, and GB/HCT total solution energies are compared with PE in Figure 2. Note that CASA and GB/ACE use the Charmm19 atomic charges, and so they are compared with PE performed with these charges. GB/HCT uses the AMBER charges, and is compared with PE performed with the AMBER charges. A CASA ε of 20 is used. There is a considerable energy dependency on the rotameric structure, with values between about −300 and +300 kcal/mol. The trend of the PE energies is reproduced approximately by both GB/ACE and CASA. The CASA surface term is almost constant for all structures (not shown), so that the CASA behavior is governed by the Coulomb term [Eq. (3)]. The CASA/PE correlation indicates, therefore, that the PE solvation energies correlate quite well with the Coulomb energy. The best correlation by far is obtained with GB/HCT. The RMS deviation between GB/HCT and PE is also quite small (21–45 kcal/mol; Table IV), considerably lower than that of CASA with the consensus dielectric constant.

Charge Modifications: Comparing Solvent Treatments

The third and most important test of the solvent models is the calculation of stability changes due to mutations that introduce or remove charged groups. In protein design, mutations are introduced randomly, and the largest stability changes will usually be associated with changes in the net protein charge. This is especially true for mutations that affect charges in fully or partly buried positions.[69] We performed two sets of mutations. The first included over 1000 mutations in seven proteins for which experimental data are not available, described in this section. The second included 140 mutations in 12 proteins for which experimental data are available, described in the next section.

Artificial mutations: CASA vs. PE

We used the Fraternali atomic surface parameters, along with a new atomic parameter σ_C for certain atoms belonging to ionized groups. The new parameter allows us to optimize the accuracy of the CASA model for charge deletions and insertions. The atom types associated with this new coefficient are listed in Supplementary Material. Values of the atomic parameter σ_C between −0.20 kcal/mol/Å^2 and the original value, −0.0598 kcal/mol/Å^2, were tried, in combination with dielectric constants in the range ε = 2–20. The CASA and PE energy changes for each mutation were compared. The combination of σ_C and ε that maximizes the agreement between CASA and PE is listed in Table V for the seven proteins. The optimum coefficient for atoms on charged sidechains (in their charged state) is between −0.15 and −0.20 kcal/mol/Å^2, and the optimum dielectric constant is 16–20. With these values, the RMSD between the CASA and PE stability changes ranges from 17.5 kcal/mol for the largest protein (AspRS) to 10.0–11.0 kcal/mol for the smallest proteins (ubiquitin and BPTI).

As explained above, we would like to obtain a consensus parameterization that works well for all proteins. The best σ_C and ε values are the same for all but one protein: a slightly smaller σ_C of −0.15 kcal/mol/Å^2 is

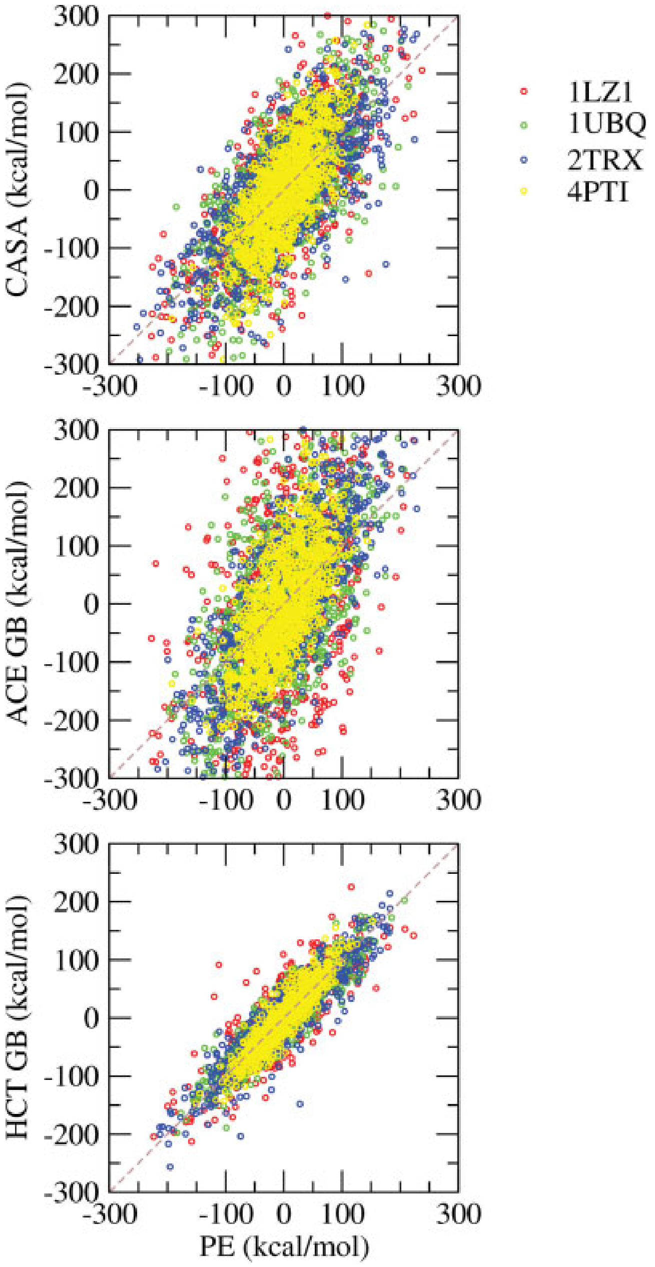

Fig. 2. Total solution energies of protein rotamer structures with the GB/ACE, surface area (CASA), and GB/HCT models, versus the corresponding Poisson values (PE). Points are colored according to the protein, listed by their PDB codes: 1LZ1, lysozyme; 1UBQ, ubiquitin; 2TRX, thioredoxin; 4PTI, bovine pancreatic trypsin inhibitor.

best for AspRS. Using the consensus σ_C value for AspRS gives almost the same result (Table V). The same σ_C and ε values work well for the experimental mutations, explained later. We conclude that they are likely to work well for most mutations in most proteins.

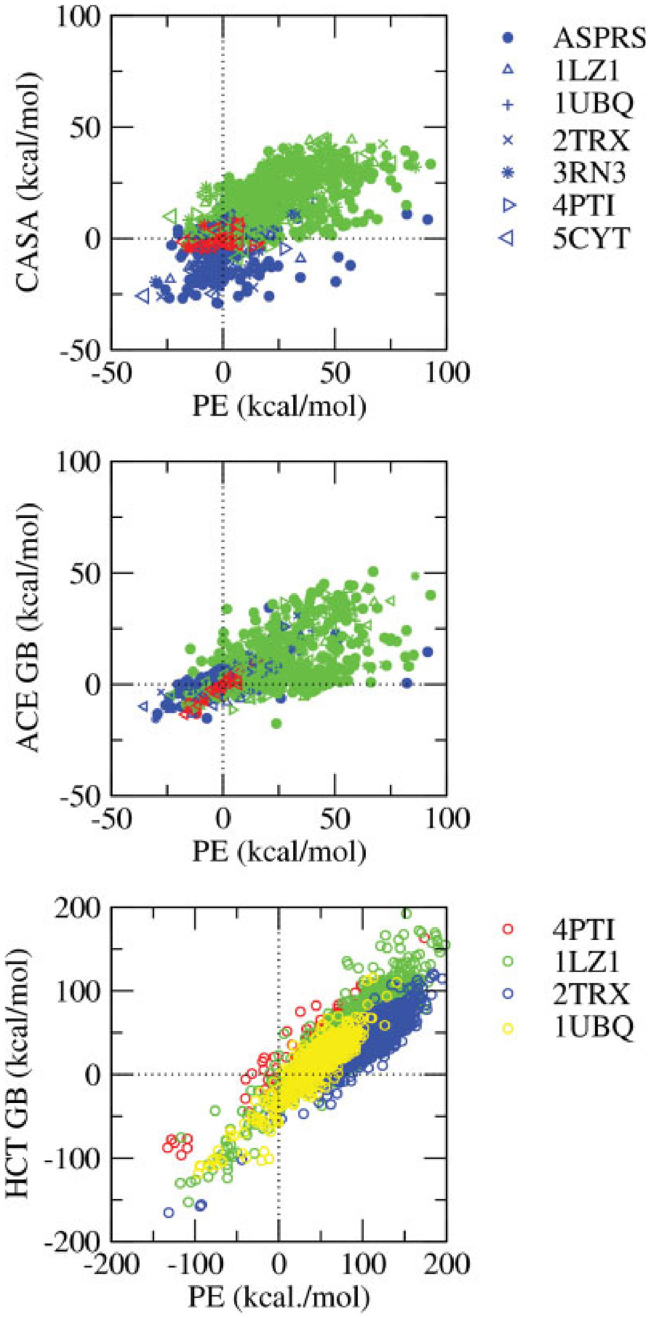

Fig. 3. Stability changes due to charge mutations with CASA, GB/ACE, and GB/HCT, vs. the corresponding changes in the PE model. For CASA and GB/ACE, green, blue and red points correspond, respectively, to charge insertions, charge deletions and polar-to-neutral conversions; different symbols correspond to different proteins, listed by their PDB codes. For GB/HCT, points are colored by protein.

The CASA stability changes are plotted against the PE ones in Figure 3. Most charge insertions decrease the protein stability, both in the CASA and PE models (i.e., they give a positive double free energy difference, $\Delta G_{\mathrm{mut}} - \Delta G_{\mathrm{nat}}$, for the thermodynamic cycle in Fig. 1). On the other hand, most charge deletions increase the protein stability. Most charged residues are at the protein surface and their interactions with the rest of the protein are screened by the nearby solvent. Eliminating the charge on such residues destabilizes the folded state, mainly due to the loss of interactions between the charge and the solvent. In the unfolded state, the same charges

TABLE VI. Comparison Between the GB and Poisson Models for Charge Mutations

		GB/ACE, Approx 1		GB/ACE, Approx 2		GB/HCT	
Protein	No of mutations	RMSD[a]	% success[b]	RMSD[a]	% success[b]	RMSD[a]	% success[b]
AspRS	518	40.7	47.9	18.2	82.2		
3RN3	106	21.6	66.9	14.0	84.0		
4PTI	51	13.7	74.5	7.0	70.6	8.3	99.2
5CYT	84	15.2	65.5	12.6	73.8		
2TRX	94	26.3	64.9	17.4	91.5	12.1	98.5
1LZ1	108	22.8	59.2	12.3	80.6	16.5	98.2
1UBQ	67	17.1	56.8	11.6	76.1	10.8	92.3

[a] RMS deviation between the GB and Poisson (PE) total energy differences (kcal/mol).

[b] Percentage of mutations predicted to have a positive or negative stability change by both GB and PE. Approximations 1 and 2 are explained in the text. Proteins are indicated by their PDB code, except AspRS, which is *Escherichia coli* aspartyl-tRNA synthetase.

are even more exposed to solvent. Thus, the charge elimination destabilizes the unfolded state even more, yielding a negative (favorable) double free energy difference. CASA and PE produce the same stability order (i.e., sign of the stability change) for 81% of mutations in all proteins. The CASA energy differences vary in a range of ≈80 kcal/mol, whereas the corresponding PE range is 150 kcal/mol; nevertheless, the two distributions are reasonably well-correlated.

Calculations were also done with a "distance-dependent" dielectric constant, $\varepsilon(r) = \varepsilon_0 r$, where r is the separation between a pair of charges and ε_0 is a constant. The success rate increased to about 86% if one uses $\varepsilon_0 = 2$. However, the performance of this model variant is much poorer for the experimental mutations, below. Increasing ε_0 improves the experimental mutations but deteriorates the artificial ones (not shown).

Artificial mutations: GB/ACE and GB/HCT vs. PE

The mutations are of three types: (1) charge insertions (usually, introduction of charge onto hydrophobic atom types); (2) charge deletions (elimination of charge on hydrophilic atom types), and (3) polar-to-neutral conversions. Thus, the mutated atoms have a combination of charge and van der Waals radius that is different from the optimum value for GB calculations. For this reason, it was necessary to modify some of the GB/ACE parameters. Table VI contains the results of two approximations. The first approximation corresponds to the usual GB/ACE parameterization (i.e., V01 volumes scaled by 0.8).[52] In the second approximation, the hydrogen types HC are assigned a Voronoi volume of 4.0, the V01 volumes of atom types CH1E, CH2E, CH3E (carbons of hydrophobic sidechains, carrying a significant partial charge in the charge-insertion mutations) were scaled by a factor 0.4, and nitrogen types NH3, NC2 were set to the original V01 volumes; for all other atom types, the V01 volumes were scaled by 0.8. The same parameterization is used for all proteins. As seen from Table VI, approximation 2 yields an RMSD between GB/ACE and PE of 7–18 kcal/mol. The fraction of mutations predicted with a positive or negative stability change by models is 80%. Thus, with this set of parameters, the GB/ACE model is comparable to CASA. Another recent study reported poorer results with GB/ACE for a similar test.[20] This was presumably due to the use of an early and non-optimal parameterization of GB/ACE; see Calimet et al.[52] for a critical discussion of the early parameterization. The GB/ACE stability changes with approximation 2 are plotted against the corresponding PE results in Figure 3. Most charge insertions destabilize the protein, as with CASA. On the other hand, most charge deletions increase the stability.

The GB/HCT results are also given in Table VI and Figure 3. The GB/HCT parameters were extensively fitted to the PE model in this work (see Methods section). However, all the GB/HCT—PE comparisons correspond to cross validation tests, using PE data not employed during the fits. The set of proteins and mutations is slightly different from GB/ACE, with fewer (four) proteins, but a larger number of mutations per protein. The RMSD between GB/HCT and PE is 8—16 kcal/mol, smaller than for CASA and GB/ACE. Correlation between GB/HCT and PE is excellent, far superior to CASA and GB/ACE, with 97% of the stability changes predicted to have the correct sign.

Charge Modifications: Comparing to Experiment

We considered 140 mutations in 12 proteins (Table VII), with experimental stability changes taken either from the ProTherm database[83] or Ref. 74. All mutations introduced or removed an ionized residue. The stability changes computed with CASA, GB/ACE, GB/HCT, and PE are summarized in Table VII and detailed in Supplementary Material. We report the mean unsigned error (MUE) for each protein, with selected outliers left out. The outliers were identified by a large van der Waals contribution to the stability change (10 kcal/mol or more). Such a large contribution reflects steric conflict in the mutant structure that results from our simple sidechain construction method. Overall, we identify 7 outliers with GB/HCT, 8 with CASA, and 21 with GB/ACE. Calculations with GB are slower than with CASA; e.g., with GB/HCT, a typical

TABLE VII. Comparison Between Models and Experiment for Charge Mutations

			Mean unsigned error (kcal/mol)			
Protein	PDB code	Total number of mutants	CASA $\varepsilon = 16$	GB/ACE $\varepsilon_P = 4$	GB/HCT $\varepsilon_P = 4$	PE $\varepsilon_P = 4$
Protein G, B_1 domain	1EM7	2	0.31 (2)	3.5 (2)	2.7 (2)	1.6 (2)
Chymotrypsin inhibitor	1YPC	15	3.3 (15)	4.3 (13)	4.4 (15)	4.6 (15)
Lysozyme	2LZM	14	2.6 (13)	2.6 (12)	4.2 (14)	4.4 (14)
Ribonuclease	2RN2	25	2.5 (24)	3.4 (23)	3.6 (24)	4.8 (24)
Src SH3 domain	1SHG	5	1.8 (5)	4.6 (5)	3.6 (5)	3.2 (5)
Staphylococcal nuclease	1STN	45	1.9 (40)	2.2 (33)	3.8 (42)	3.9 (42)
Thioredoxin	2TRX	3	5.1 (3)	6.8 (2)	1.4 (2)	1.8 (2)
Trypsine	1BPI	11	1.4 (10)	2.4 (10)	1.7 (9)	6.7 (9)
Ubiquitin	1UBQ	8	2.1 (8)	1.5 (8)	1.6 (8)	1.4 (8)
pepT1[c]	–	4	1.6 (4)	2.4 (4)	2.8 (4)	1.9 (4)
K2AE2[c]	–	4	1.9 (4)	2.3 (4)	1.8 (4)	1.4 (4)
KEAKE[c]	–	4	1.8 (4)	2.0 (4)	1.6 (4)	1.2 (4)
Total		140	2.1 (132)[a]	2.9 (120)	3.4 (133)	3.9 (133)
					2.1[b] ($\varepsilon_P = 8$)	2.6[b] ($\varepsilon_P = 8$)

[a] In parentheses: the number of mutations after discarding those with a van der Waals contribution of 10 kcal/mol or more.

[b] GB/HCT or PE supplemented by a surface area term and using a protein dielectric of 8 (see text).

[c] Experimental data from Ref. 88. Natives structures are not known experimentally; they were therefore modelled using the Swiss PDB Viewer, which constructs an ideal α-helix and places sidechains in favorable rotamers.[89]

mutation takes 2–4 times more CPU time than with CASA.

Excluding the outliers, the MUE is 2.1 kcal/mol with CASA. Poorer results are obtained if the constant dielectric in the Coulomb electrostatic term is replaced by a distance-dependent dielectric, $\varepsilon(r) = \varepsilon_0 r$: with $\varepsilon_0 = 2$, for example, the error increases to 7.9 kcal/mol. Similarly, while sidechain reconstruction (above) works well without any surface term [$\alpha = 0$ in Eq. (3)], results here are much poorer: the MUE increases to 4.6 kcal/mol if the surface term is left out. These results illustrate further that the best model for protein design should be a consensus model, not necessarily optimal for each prediction task but reasonably competent for all of them.

In the case of GB/ACE, GB/HCT, and PE, the protein dielectric constant ε_P is adjusted empirically to minimize the MUE. Indeed, our simple sidechain modeling does not allow dielectric relaxation of the protein structure when an ionized sidechain is introduced or removed. For all three models, GB/ACE, GB/HCT, and PE, the best results are obtained with a protein dielectric of 3–5. Note that the use of GB models with a protein dielectric greater than 1 [Eq. (4)] is not very common, but is straightforward.[42,84] A consensus value of four for the protein dielectric works well with all three models, giving MUEs of 3.4 kcal/mol with GB/HCT and 3.9 kcal/mol with PE. Results with a dielectric of one are very poor (e.g., the MUE is 8.7 kcal/mol with PE). With GB/ACE, mutations involving Asp and Glu sidechains were found to be poorly described. Therefore, another adjustable parameter was introduced, which empirically increases by 6.3 kcal/mol the contribution of Asp or Glu sidechains to the stability of the unfolded state. This leads to a MUE of 2.9 kcal/mol with GB/ACE (using $\varepsilon_P = 4$, and with 21 outliers omitted). With GB/HCT and PE, the prediction quality did not depend noticeably on the amino acid type.

Mutation of a charged sidechain could, in some cases, alter the protonation state of surrounding residues. It is not practical to model this effect in detail, because there are too many possible protonation states to consider. It is implicitly incorporated into the model through the choice of a protein dielectric constant greater than 1. Furthermore, for each protein, we systematically did CASA calculations with the histidine sidechains either all ionized or all neutral. Results for the mutations were reasonably similar (not shown).

The performance of the three models, GB/ACE, GB/HCT, and PE, could be improved by adding an additional, surface area term to describe hydrophobic solvation and dispersion interactions with the surrounding solvent, as is commonly done in protein modeling, see e.g. Refs. 32,85. Using a surface coefficient of $\sigma = -0.04$ kcal/mol/Å^2 for all atoms and a somewhat larger dielectric of $\varepsilon = 8$, the mean unsigned error for GB/HCT drops to 2.1 kcal/mol, identical to the CASA value. For PE, the mue drops to 2.6 kcal/mol, using a surface coefficient of $\sigma = -0.05$ kcal/mol/Å^2 for all atoms and a dielectric of $\varepsilon = 8$. We did not try this procedure with GB/ACE, because of its poorer overall performance.

The experimental agreement is slightly worse with PE than with CASA and GB/HCT. Nevertheless, we took PE to be the reference model for the artifical mutations, above. Indeed, it is the model with the clearest physical basis and the best performance in general for protein electrostatics.[29,68] CASA, in contrast, is not expected to give quantitative accuracy for mutations of buried sidechains, since the surface energy cannot distinguish between positions that are deeply buried and positions that are closer to the protein surface. GB/HCT has a MUE close to PE, largely because its parameters have been optimized to reproduce PE. Thus, even though the PE mue is slightly larger, our results do not contradict many earlier studies of protein electrostatics showing that PE is a valid reference model.

CONCLUSIONS

We have examined three problems that are important ingredients of computational protein design: sidechain placement, protein solvation, and mutagenesis involving charged sidechains. We have tested the behavior of four implicit solvent models: the Poisson model, considered to be the standard of accuracy, the GB/ACE and GB/HCT generalized Born models, and the CASA model. The CASA model is commonly used for protein design; GB/ACE and GB/HCT are more recent and sophisticated solvent models. Several methods have been proposed that allow GB models to be used for protein design.[19,23,38] The most recent one achieves a residue–pairwise GB implementation without any loss in accuracy.[23]

Using a standard mean field method for sidechain placement along with the CASA solvent model, we obtain results of similar quality to earlier workers.[14,82] The results are weakly sensitive to the details of the CASA model: solvent dielectric, overall weight of the surface term, force field. This is consistent with the good results obtained by some earlier workers without any electrostatic term or solvent model.[14,16,21]

In contrast, the mutagenesis results are much more sensitive to the solvent treatment. This is expected, since almost all the mutations involved insertions or deletions of a net charge. Many of the insertions were on buried sidechains. All four solvent models were compared. Different kinds of parameter optimization were performed. In the CASA model, we introduced a new atomic surface parameter for oxygen and nitrogen atoms in charged sidechains and we adjusted the dielectric constant. In the GB/ACE model, we adjusted the atomic volumes of selected atom types belonging to charged sidechains. In the GB/HCT model, we optimized the atomic volumes and scaling factors (58 parameters in all). All these optimizations used a large data set of artifical mutations, and took the Poisson model as the reference. The Poisson model employed a dielectric constant of one for the protein (and 80 for solvent), because we are optimizing the implicit solvent treatments, which try to reproduce the relaxation of solvent, not protein, in response to the mutations. Similarly, the GB models employed a protein dielectric of one and a solvent dielectric of 80. CASA uses a single dielectric constant, which tries to capture the average effect that solvent relaxation exerts within the protein interior. A value of about 20 worked well—intermediate between the solvent value of 80 and the value of one that is appropriate for the protein interior in these calculations. With these parameter adjustments and dielectric constants, the quality of CASA and GB/ACE were very similar, as compared to the Poisson reference. In 80–81% of the charged mutations, they correctly captured the sign and order of magnitude of the protein stability change. With GB/HCT, the sign was correct for 97% of the mutations, a dramatic improvement with this more recent GB variant.[48] Protein solvation energies were also strikingly better with GB/HCT than with CASA or GB/ACE.

Finally, in a separate test, we compared the models to experiment for a set of 140 point mutations. Comparison with experiment introduces several major difficulties. First, the stability changes are very small, so that a simple null model will usually give better agreement than any current implicit solvent model. Second, we need a model for the unfolded protein, whose structure is not known. Third, we must describe the structural relaxation *of the protein*, and not just that of the solvent. Fourth, we must describe hydrophobic contributions to stability, which are notoriously hard to capture with simple models.[25,86] To describe the unfolded state, we adopted the simplest possible, commonly-used, tripeptide model. To describe the protein relaxation, we increased the protein dielectric in the GB and PE models, exploring values between 2 and 32. With CASA, we kept a dielectric of 20, and verified that increasing it had only a small effect on the results. With GB/ACE, we also added an empirical correction for Asp/Glu sidechains in the unfolded state. This correction presumably reflects a problem with the GB/ACE parameterization of carboxylate groups. Finally, we incorporated hydrophobic contributions, along with solvent dispersion interactions into the GB/HCT and PE models by adding a surface term. All the solvent models gave fair agreement with experiment, with mean unsigned errors of 2.1 kcal/mol for CASA, 2.9 kcal/mol for GB/ACE, and 2.1 kcal/mol for GB/HCT supplemented by the surface term. GB/HCT also gave excellent results for the artifical mutations and the solvation energies.

In summary, (a) we confirm earlier observations that the choice of electrostatic model is not very important for sidechain placement; (b) GB/HCT and CASA give the same accuracy for surface mutations; (c) GB/HCT yields an enormous improvement for total solvation free energies and for mutations in fully or partly buried positions. Thus, for problems like protein design that involve all these aspects, the most recent GB models, and their best pairwise implementations,[87] represent an important step forward.

ACKNOWLEDGMENTS

We thank A. Jaramillo and M. Schmidt am Busch for useful discussions. Support was provided by the Egide program ZENON between Cyprus and France (to GA and TS), by the ACI IMPBio Program of the French Ministry of Research (to TS), by the HPC-EUROPA project RII3-CT-2003-506079 (to GA and TS), and by a Cyprus "Research Reinforcement of Young Cypriot Researchers" (PENEK) grant (to GA).

REFERENCES

1. Hellinga H. Metalloprotein design. Curr Opin Biotech 1996;7: 437–441.
2. Al-Lazikani B, Jung J, Xiang Z, Honig B. Protein structure prediction. Curr Opin Chem Biol 2001;5:51–56.
3. Marti-Renom M, Stuart A, Fiser A, Sanchez R, Melo F, Sali A. Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomol Struct 2000;29:291–325.

4. Bolon D, Mayo S. Enzyme-like proteins by computational design. Proc Natl Acad Sci USA 2001;98:14274–14279.
5. Ponder J, Richards FM. Tertiary templates for proteins: use of packing criteria in the enumeration of allowed sequences for different structural classes. J Mol Biol 1988;193:775–791.
6. Tuffery P, Etchebest C, Hazout S, Lavery R. A new approach to the rapid determination of protein side chain conformations. J Biomol Struct Dyn 1991;8:1267.
7. Dunbrack R, Karplus M. Backbone-dependent rotamer library for proteins. Application to sidechain prediction. J Mol Biol 1993; 230:543–574.
8. Dunbrack R, Cohen F. Bayesian statistical analysis of protein sidechain rotamer preferences. Prot Sci 1997;6:1661–1681.
9. Desmet J, De Mayaer M, Hazes B, Lasters I. The dead-end elimination theorem and its use in protein sidechain positioning. Nature 1992;356:539–542.
10. Goldstein R. Efficient rotamer elimination applied to protein sidechains and related spin glasses. Biophys J 1994;66:1335–1340.
11. Looger L, Hellinga H. Generalized dead-end elimination algorithms make large-scale protein sidechain structure prediction tractable: implications for protein design and structural genomics. J Mol Biol 2001;307:429–445.
12. Holm L, Sander C. Database algorithm for generating protein backbone and sidechain coordinate from a C_α trace: application to model building and detection of coordinates errors. J Mol Biol 1991;218:183–194.
13. Lee C, Subbiah S. Prediction of protein sidechain conformation by packing optimization. J Mol Biol 1991;217:373–388.
14. Koehl P, Delarue M. Application of a self-consistent mean field theory to predict protein sidechain conformations and estimate their conformational entropy. J Mol Biol 1994;239:249–275.
15. Mendes J, Baptista A, Carrondo M, Soares C. Implicit solvation in the self-consistent mean field theory method: sidechain modelling and prediction of folding free energies of protein mutants. J Computer Aided Mol Des 2001;15:721–740.
16. Liang S, Grishin N. Sidechain modeling with an optimized scoring function. Prot Sci 2002;11:322–331.
17. Onufriev A, Bashford D, Case D. Exploring protein native states and large-scale conformational changes with a modified generalized Born model. Proteins 2004;55:383–394.
18. Feig M, Onufriev A, Lee M, Im W, Case D, Brooks CL, III. 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.
19. Pokola N, Handel T. Energy functions for protein design I: efficient and accurate continuum electrostatics and solvation. Prot Sci 2004;13:925–936.
20. Jaramillo A, Wodak S. Computational protein design is a challenge for implicit solvation models. Biophys J 2005;88:156–171.
21. Canutescu AA, Shelenkov AA, Dunbrack R. A graph-theory algorithm for rapid protein side-chain prediction. Prot Sci 2003; 12:2001–2014.
22. Feig M, Brooks CL, III. Recent advances in the development and application of implicit solvent models in biomolecule simulations. Curr Opin Struct Biol 2004;14:217–224.
23. Archontis G, Simonson T. A residue-pairwise Generalized Born scheme suitable for protein design calculations. J Phys Chem B 2005;109:22667–22673.
24. Tanford C. The hydrophobic effect. New York: John Wiley; 1980.
25. Roux B, Simonson T. Implicit solvent models. Biophys Chem 1999;78:1–20.
26. Warwicker J, Watson H. Calculation of the electrostatic potential in the active site cleft due to α helix dipoles. J Mol Biol 1982;157:671–679.
27. Honig B, Nicholls A. Classical electrostatics in biology and chemistry. Science 1995;268:1144–1149.
28. Schaefer M, Vlijmen Hv, Karplus M. Electrostatic contributions to molecular free energies in solution. Adv Prot Chem 1998;51:1–57.
29. Simonson T. Electrostatics and dynamics of proteins. Rep Prog Phys 2003;66:737–787.
30. Simonson T. Macromolecular electrostatics: continuum models and their growing pains. Curr Opin Struct Biol 2001;11:243–252.
31. Sitkoff D, Sharp K, Honig B. Accurate calculation of hydration free energies using macroscopic solvent models. J Phys Chem 1994;98:1978–1988.
32. Simonson T, Brünger AT. Solvation free energies estimated from macroscopic continuum theory: an accuracy assessment. J Phys Chem 1994;98:4683–4694.
33. Bashford D, Karplus M. The pK_a's of ionizable groups in proteins: atomic detail from a continuum electrostatic model. Biochemistry 1990;29:10219–10225.
34. Archontis G, Simonson T, Karplus M. Binding free energies and free energy components from molecular dynamics and Poisson-Boltzmann calculations. Application to amino acid recognition by aspartyl-tRNA synthetase. J Mol Biol 2001;306:307–327.
35. Hendsch Z, Tidor B. Electrostatic interactions in the GCN4 leucine zipper: substantial contributions arise from intramolecular interactions enhanced on binding. Prot Sci 1999;8:1381–1392.
36. Gohlke H, Kiel C, Case D. Insight into protein-protein binding by binding free energy calculation and free energy decomposition for the Ras-Raf and Ras-RalGDS complexes. J Mol Biol 2003;330:891–913.
37. David L, Luo R, Gilson M. Comparison of Generalized Born and Poisson models: energetics and dynamics of HIV protease. J Comput Chem 2000;21:295–309.
38. Wisz M, Hellinga H. An empirical model for electrostatic interactions in proteins incorporating multiple geometry-dependent dielectric constants. Proteins 2003;51:360–377.
39. Still WC, Tempczyk A, Hawley R, Hendrickson T. Semianalytical treatment of solvation for molecular mechanics and dynamics. J Am Chem Soc 1990;112:6127–6129.
40. Bashford D, Case D. Generalized Born models of macromolecular solvation effects. Ann Rev Phys Chem 2000;51:129–152.
41. Hawkins G, Cramer C, Truhlar D. Pairwise descreening of solute charges from a dielectric medium. Chem Phys Lett 1995;246:122–129.
42. Schaefer M, Karplus M. A comprehensive analytical treatment of continuum electrostatics. J Phys Chem 1996;100:1578–1599.
43. Qiu D, Shenkin P, Hollinger F, Still W. 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.
44. Ghosh A, Rapp C, Friesner RA. Generalized Born model based on a surface area formulation. J Phys Chem B 1998;102:10983–10990.
45. Dominy B, Brooks CL, III. Development of a Generalized Born model parameterization for proteins and nucleic acids. J Phys Chem B 1999;103:3765–3773.
46. Lee M, Salsbury Jr, F, Brooks CL, III. Novel generalized Born methods. J Chem Phys 2002;116:10606–10614.
47. Wagner F, Simonson T. Implicit solvent models: combining an analytical formulation of continuum electrostatics with simple models of the hydrophobic effect. J Comput Chem 1999;20:322–335.
48. Onufriev A, Bashford D, Case D. Modification of the generalized Born model suitable for macromolecules. J Phys Chem B 2000; 104:3712–3720.
49. Simonson T, Carlsson J, Case DA. Proton binding to proteins: pK_a calculations with explicit and implicit solvent models. J Am Chem Soc 2004;126:4167–4180.
50. Lee M, Salsbury F, Jr., Brooks C, III. Constant pH molecular dynamics using continuous titration coordinates. Proteins 2004;56: 738–752.
51. Cornell W, Abseher R, Nilges M, Case D. Continuum solvent molecular dynamics study of flexibility in interleukin-8. J Mol Graph Model 2001;19:136–145.
52. Calimet N, Schaefer M, Simonson T. Protein molecular dynamics with the generalized Born/ACE solvent model. Proteins 2001; 45:144–158.
53. Majeux N, Scarsi M, Apostolakis J, Ehrhardt C, Caflisch A. Exhaustive docking of molecular fragments with electrostatic solvation. Proteins 1999;37:88–105.
54. Bursulaya B, Brooks C, III. Comparative study of the folding free energy landscape of a three-stranded β-sheet protein with explicit and implicit solvent models. J Phys Chem B 2001;104: 12378–12383.
55. Simmerling C, Strockbine B, Roitberg A. All-atom structure prediction and folding simulations of a stable protein. J Am Chem Soc 2002;124:11258–11259.
56. Rapp C, Friesner R. Prediction of loop geometry using a generalized Born model of solvation effects. Proteins 1999;35:173–183.
57. Felts A, Gallicchio E, Wallqvist A, Levy R. Distinguishing native conformations of proteins from decoys with an effective free energy estimator based on the OPLS all-atom force field and the surface generalized Born solvent model. Proteins 2002;48:404–422.

58. Eisenberg D, McClachlan A. Solvation energy in protein folding and binding. Nature 1986;319:199–203.
59. Ooi T, Oobatake M, Nemethy G, Scheraga H. Accessible surface areas as a measure of the thermodynamic hydration parameters of peptides. Proc Natl Acad Sci USA 1987;84:3086–3090.
60. Wesson L, Eisenberg D. Atomic solvation parameters applied to molecular dynamics of proteins in solution. Prot Sci 1992;1:227–235.
61. Fraternali F, van Gunsteren W. An efficient mean solvation force model for use in molecular dynamics simulations of proteins in aqueous solution. J Mol Biol 1996;256:939–948.
62. Ferrara P, Apostolakis J, Caflisch A. Evaluation of a fast implicit solvent model for molecular dynamics simulations. Proteins 2002;46:24–33.
63. Pei J, Wang Q, Zhou J, Lai L. Estimating protein-ligand binding free energy: atomic solvation parameters for partition coefficient and solvation free energy calculation. Proteins 2004; 57:661–664.
64. Dahiyat B, Mayo S. Protein design automation. Prot Sci 1996;5: 895–903.
65. Koehl P, Levitt M. Protein topology and stability define the space of allowed sequences. Proc Natl Acad Sci USA 2002;99: 1280–1285.
66. Ogata K, Jaramillo A, Cohen W, Briand J, Conan F, Wodak S. Automatic sequence design of MHC class-I binding peptides impairing CD8+ T cell recognition. J Biol Chem 2003;278:1281.
67. Liang S, Grishin N. Effective scoring function for protein sequence design. Proteins 2004;54:271–281.
68. Baker N. Poisson-Boltzmann methods for biomolecular electrostatics. Methods Enzym. 2004;383:94.
69. Dwyer J, Gittis A, Karp D, Lattman E, Spencer D, Stites W, Garcia-Moreno B. High apparent dielectric constants in the interior of a protein reflect water penetration. Biophys J 2000;79: 1610–1620.
70. Brooks B, Bruccoleri R, Olafson B, States D, Swaminathan S, Karplus M. Charmm: a program for macromolecular energy, minimization, and molecular dynamics calculations. J Comput Chem 1983;4:187–217.
71. Hawkins G, Cramer C, Truhlar D. J Phys Chem 1996;100:19824.
72. Cornell W, Cieplak P, Bayly C, Gould I, Merz K, Ferguson D, Spellmeyer D, Fox T, Caldwell J, Kollman P. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. J Am Chem Soc 1995;117:5179–5197.
73. Swanson J, Adcock S, McCammon J. Optimized radii for Poisson-Boltzmann calculations with the AMBER force field. J Chem Theory Comput 2005;1:484–493.
74. Guérois R, Nielsen J, Serrano L. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. J Mol Biol 2002;320:369–387.
75. Simonson T, Archontis G, Karplus M. Free energy simulations come of age: the protein–ligand recognition problem. Acc Chem Res 2002;35:430–437.
76. Lee B, Richards F. The interpretation of protein structures: estimation of static accessibility. J Mol Biol 1971;55:379–400.
77. Brünger AT. X-plor version 3.1, A system for X-ray crystallography and NMR. New Haven: Yale University Press; 1992.
78. Schaefer M, Bartels C, Leclerc F, Karplus M. Effective atom volumes for implicit solvent models: comparison between Voronoi volumes and minimum fluctuation volumes. J Comp Chem 2001; 22:1857–1879.
79. Moulinier L, Case D, Simonson T. X-ray structure refinement of proteins with the generalized Born solvent model. Acta Cryst D 2003;59:2094–2103.
80. Madura J, Briggs J, Wade R, Davis M, Luty B, Ilin A, Antosiewicz J, Gilson M, Baheri B, Scott L, McCammon J. Electrostatics and diffusion of molecules in solution: simulations with the University of Houston Brownian Dynamics program. Comput Phys Commun 1995;91:57–95.
81. Street A, Mayo S. Pairwise calculation of protein solvent-accessible surface areas. Folding Des 1998;3:253–258.
82. Yang J, Tsai C, Hwang M, Tsai H, Hwang J, Kao C. GEM: a Gaussian evolutionary method for predicting protein sidechain conformations. Prot Sci 2002;11:1897–1907.
83. Kumar M, Bava K, Gromiha M, Parabakaran P, Kitajima K, Uedaira H, Sarai A. ProTherm and ProNIT: thermodynamic databases for proteins and protein–nucleic acid interactions. Nucl Acids Res 2006;34:D204–206.
84. Sigalov G, Scheffel P, Onufriev A. Incorporating variable dielectric environments into the generalized born model. J Chem Phys 2005; 122:094511.
85. Simonson T. Free energy calculations: approximate methods for biological macromolecules. In Chipot C, Pohorille A, editors. Free energy calculations: theory and applications in chemistry and biology. New York: Springer Verlag; 2006, Ch. 12.
86. Gallicchio E, Kubo M, Levy R. Enthalpy-entropy and cavity decomposition of alkane hydration free energies: numerical results and implications for theories of hydrophobic hydration. J Phys Chem B 2000;104:6271–6285.
87. Archontis G, Simonson T. Proton binding to proteins: a free energy component analysis using a dielectric continuum model. Biophys J 2005;88:3888–3904.
88. Pace C, Scholtz J. A helix propensity scale based on experimental studies of peptides and proteins. Biophys J 1998;75:422–427.
89. Guex N, Peitsch M. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis 1997;18:2714–2723.