PROTEINS: Structure, Function, and Bioinformatics 67:606–616 (2007)

A New Implicit Solvent Model for Protein–Ligand Docking

Antonio Morreale, Rubén Gil-Redondo, and Ángel R. Ortiz*
Bioinformatics Unit, Centro de Biología Molecular "Severo Ochoa" (CSIC-UAM), Universidad Autónoma de Madrid, Cantoblanco, Madrid 28049, Spain

ABSTRACT **A new implicit solvent model for computing the electrostatics binding free energy in protein–ligand docking is proposed. The new method is based on an adaptation of the screening coulombic potentials proposed originally by Hassan et al. (J Phys Chem B 2000;104:6490–6498). In essence, it relies on two basic assumptions; (i) solvent screening can be accounted for by means of radially dependent sigmoidal dielectric functions and; (ii) the effective atom Born radii can be expressed only as a function of the exposed atom surface. Parameters of the model other than radii and charges are generic. These were optimized for a dataset of 826 protein–ligand complexes, comprising both X-ray complexes for 23 receptors as well as decoys generated by docking computations. We show that the new model provides satisfactory results when benchmarked against reference values based on the numerical solution of the Poisson equation, with a root mean square error of 4.2 kcal/mol over a range of ~40 kcal/mol in electrostatics binding free energies, a cross-validated r^2 of 0.81, a slope of 0.97, and an intercept of 1.06 kcal/mol. We show that the model is appropriate for ligands of different sizes, polarities, overall charge, and chemical composition. Furthermore, not only the total value of the electrostatic contribution to the binding free energy, but also its components (coulombic term, receptor desolvation, and ligand desolvation) are reasonably well reproduced. Computation times of ~0.030 s per pose are obtained on a single processor desktop workstation. Proteins 2007;67:606–616.**

Key words: electrostatics; force fields; solvation; binding free energies; virtual screening; docking

INTRODUCTION

An adequate treatment of solvation is yet an unsolved problem in protein–ligand docking.[1] Solvation (or hydration, for simplicity throughout this manuscript we will interchange both terms) plays an important role in the energetics of ligand–protein association,[2–5] and when using molecular mechanics energy functions, its physical model influences ranking in virtual screening,[6–10] and to a more limited extent, docking geometry.[11,12] Introduction of explicit solvent would possibly be the most rigorous means of incorporating the solvent effect, but this is impractical in docking computations. On the other hand, there is ample consensus nowadays that implicit solvation methods, while introducing various approximations to hydration effects,[13] provide an adequate balance between computational efficiency and physical soundness.

The implicit hydration free energy is usually divided into several components: cavity formation, short-range solvent–solute interactions, and electrostatic solvation. In this paper we will only consider the later. Implicit electrostatic solvation is achieved by presuming that the solvent is a continuum high-dielectric-constant medium that responds to the partial charges of a low-dielectric-constant solute. Two major continuum models have been applied in protein–ligand docking, one that uses the numerical solutions of the Poisson equation (PE),[14] and another that applies the generalized-Born (GB) approximation.[15,16] While the use of the PE method for docking has been explored, it is the GB formulation, due to its improved computational efficiency, the method most widely investigated to account for electrostatic hydration effects in protein–ligand docking.[17–20] Nevertheless, the computational cost of an straightforward implementation of these models is still significant, and in practice most docking protocols employ PE or GB models only as a second rescoring step, while the docking scoring functions employ a crude treatment of electrostatics and solvation.[6,10,19,21]

It would be beneficial to develop new or improved approaches to the calculation of the electrostatics binding free energy with a more appropriate balance of accuracy and speed for protein–ligand docking. Adaptations of

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

Grant sponsor: MEC; Grant numbers: BIO2001-3745, BIO2005-0576 and GEN2003-206420-C09-08; Grant sponsor: Comunidad de Madrid; Grant numbers: GR/SAL/0306/2004 and 200520M157; Grant sponsor: CSIC, intramural program (PIF 2005, project CAR); Grant number: PIF2005; Grant sponsor: Fundación Ramón Areces.

*Correspondence to: Ángel R. Ortiz Bioinformatics Unit, Centro de Biología Molecular "Severo Ochoa" (CSIC-UAM), Universidad Autónoma de Madrid, Cantoblanco, 28049 Madrid, Spain.
E-mail: aro@cbm.uam.es

Received 7 March 2006; Revised 10 August 2006; Accepted 15 September 2006

Published online 28 February 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21269

PE and GB methods have been proposed earlier to make them more efficient in docking. These are based on the realization that when two molecules are brought together, the main contribution to electrostatic desolvation originates from the displacement of the first shell of water molecules occupying the interacting surface, the so-called first shell approximation. Based on this idea, Arora and Bashford[22] presented the solvation energy density occlusion (SEDO) method as an approximation PE electrostatic desolvation. With SEDO, a solvation energy density is stored on a grid for each molecule in isolation. Desolvation is then calculated by integrating the solvation energy density of one molecule that is occluded by the other when the pair associates. Similarly, Caflisch and coworkers have used the first shell approximation to modify the GB model by using the Coulomb approximation of the electric displacement via volume-integration.[23] The modification allows estimating the energy in solution of ~300 protein-fragment binding modes per second on a 550 MHz Pentium III. For two different proteins binding to a set of small molecule fragments a r^2 of ~0.72 with respect to PE energies and slopes close to the unit were obtained in each case. However, no data are available on the overall performance with a large dataset of structurally different proteins. A limitation in both approaches is the requirement of a rigid protein structure, in order to efficiently use the grid technology.

Herein, we propose a new implicit solvent model (ISM) for computing the electrostatics binding free energy in protein–ligand docking, based on an adaptation to protein–ligand binding of the screening coulombic potentials (SCP) proposed originally by Hassan et al. to treat electrostatics interactions in proteins.[24–26] The model (SCP-ISM) shares similarities with GB approaches, but also differs from them in a number of important aspects. With SCP-ISM the system is described as immersed in a continuum that permeates all space and is completely characterized by the screening function. In this way the model departs from GB approaches, eliminating the need to define an internal dielectric constant and a (discontinuous) boundary between protein and solvent. This is advantageous in the computation of protein–ligand docking interactions, which mostly take place in the protein–solvent interface, where the precise location of the boundary is not well defined. A second feature of the model is the use of the first shell approximation, that is, the effective Born radii for each atom is expressed only as a function of the atom exposed surface accessible area. This allows an improved trade off between speed and accuracy. A third advantage is that the effective Born radii only appear in the self-energy terms, playing no role in the calculation of the interaction contributions, allowing storing the protein electrostatic potential in a grid for a fast computation of the pairwise electrostatic contribution to the electrostatic binding free energy, the so-called test charge approximation. Finally, the model parameters, other than radii and charges, are generic, allowing the model to be easily parametrized and therefore it is adaptable to large scale docking computations.

METHODS

Implicit Solvent Model Based on Screened Coulomb Potential

We review here for completeness of the most salient features of the SCP-ISM theory developed by Hassan et al. More details can be obtained from the original publications.[25,26] The model starts from the Lorentz–Debye–Sack theory of polar liquids, which establishes that the screening effect due to the solvent shows a sigmoidal-distance dependent dielectric function of the form:

$$D(r) = \frac{\varepsilon + 1}{1 + k\ \exp[-\lambda(\varepsilon + 1)r]} - 1 \tag{1}$$

where ε is the solvent dielectric constant, $k = (\varepsilon - 1)/2$ and λ is a parameter controlling the rate of change of $D(r)$. A similar screening function has also been previously introduced in the docking program Autodock.[27] A second key aspect of the model is the assumption that the main contribution to electrostatic desolvation of an atom originates from the displacement of the first shell of water molecules surrounding the atom and occupying the atomic surface. The model defines parameters $R_{i,\mathrm{B_S}}$ and $R_{i,\mathrm{B_V}}$ as the effective Born radii for the processes of transferring an atom from the vacuum into a protein interior, surrounded by solvent or vacuum, respectively. According to the first shell approximation, these radii are calculated using linear relationships of the form:

$$R_{i,\mathrm{B_S}} = R_{i,w}\xi_i + R_{i,p}(1 - \xi_i) \tag{2}$$

and,

$$R_{i,\mathrm{B_V}} = R_{i,\mathrm{V}}\xi_i + R_{i,p}(1 - \xi_i) \tag{3}$$

where ξ_i is the fraction of SASA (A_i) of the i atom: $\xi_i = A_i/4\pi(r_{\mathrm{vdw},i} + r_{\mathrm{probe}})^2$. With $R_{i,W} = R_{i,\mathrm{COV}} + h_{(+,-)}$, $R_{i,P} = R_{i,\mathrm{COV}} + g$, and $R_{i,\mathrm{V}} = R_{\mathrm{COV}}$. R_{COV} is the covalent radius, and $h_{(+,-)}$ and g are positive quantities that account for the enlargement of the cavity due to charge effects. In particular, $h_{(+,-)}$ depends on the atomic charge (see below). Applying this function to the solvation process (a detailed derivation of the model can be found in the original papers by Hassan et al.[25,26]) the following equation is obtained:

$$\Delta G_{\mathrm{elec}} = \sum_{i<j} \frac{q_i q_j}{r_{ij}} \left[\frac{1}{D_{\mathrm{S}}(r_{ij})} - \frac{1}{D_{\mathrm{V}}(r_{ij})}\right] + \frac{1}{2}\sum_{i=1}^{N} q_i^2 \left\{\frac{1}{R_{i,\mathrm{B_S}}}\left[\frac{1}{D_{\mathrm{S}}(R_{i,\mathrm{B_S}})} - 1\right] - \frac{1}{R_{i,\mathrm{B_V}}}\left[\frac{1}{D_{\mathrm{V}}(R_{i,\mathrm{B_V}})} - 1\right]\right\} \tag{4}$$

where the first term represents Coulombic interactions between charged particles screened by dielectric sigmoidal function depicted in Eq. (1), s stands for the solvent and v for vacuum.

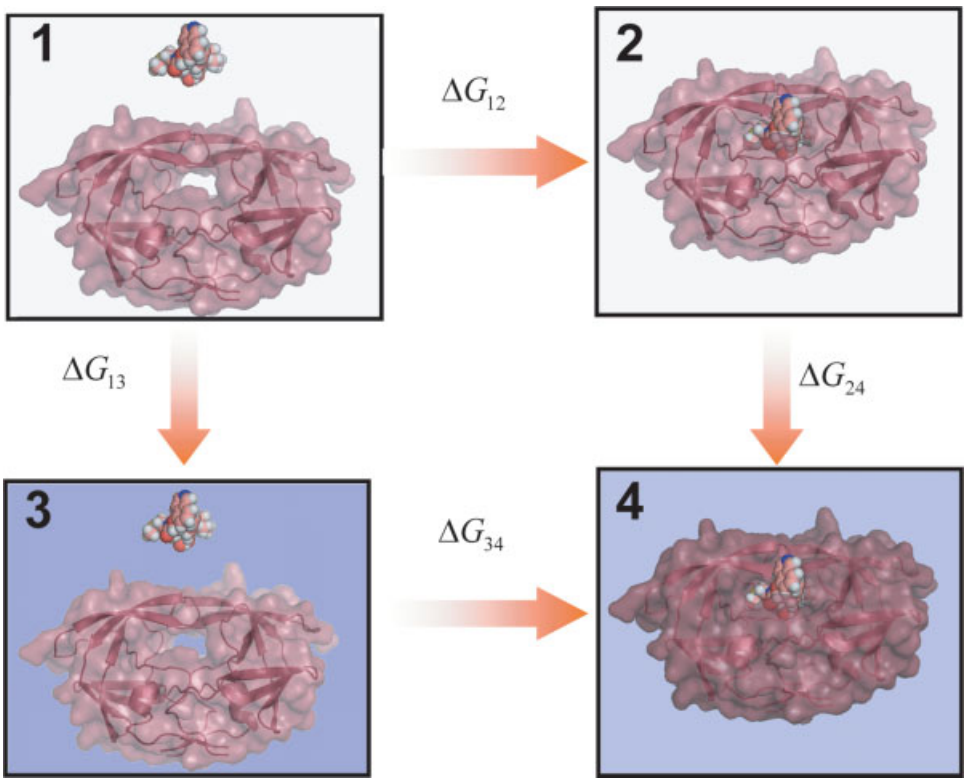

Fig. 1. Thermodynamic cycle employed for the calculation of the electrostatics binding free energy. The process of binding in solution (ΔG_{34}), can be alternatively computed as the process of desolvating the unbound species ($-\Delta G_{13}$), letting them interact in vacuum (ΔG_{12}), and solvating the resulting complex (ΔG_{24}). See text for additional details.

Extending SCP–ISM to the Ligand-Receptor Docking Problem

To extend Eq. (4) to the problem of ligand–receptor interactions, the thermodynamic cycle depicted in Figure 1 was built. Boxes represent vacuum (1, 2) and solvent (3, 4) states, in both unbound (1, 3) and bound (2, 4) forms. ΔG_{elec} is then calculated from Eq. (5):

$$\Delta G_{\text{elec}} = \Delta G_{34} = \Delta G_{12} + (\Delta G_{24} - \Delta G_{13}) \tag{5}$$

ΔG_{13}, ΔG_{24}, and ΔG_{12} that can be obtained directly from Eq. (4), and assuming a rigid ligand–protein binding, after reorganizing terms, Eq. (6) is obtained:

$$\Delta G_{\text{elec}} = \sum_{i=1}^{N_L}\sum_{j=1}^{N_R}\frac{q_i q_j}{D_s(r_{ij})r_{ij}} + \frac{1}{2}\sum_{i=1}^{N_L+N_R} q_i^2 \times \left[\left(\frac{1}{D_S(R_S^C)R_S^C} - \frac{1}{D_S(R_S^U)R_S^U}\right) + \left(\frac{1}{R_S^U} - \frac{1}{R_S^C}\right)\right] \tag{6}$$

N_L and N_R are the number of atoms in ligand and receptor, respectively, R_S^C and R_S^U are the effective Born radii for complexed and uncomplexed forms of both, ligand and receptor. In analogy with Eq. (4), the first term represents charge–charge interactions between the ligand and the receptor, and the second contains desolvation penalties to be paid for removing atoms from the solvent to form the complex. Eq. (6) is the main result of applying SCP–ISM theory to the protein–ligand binding problem.

Surface Model

SASA values required in Eqs. (2) and (3) were obtained with the LCPO approximation[28] with a solvent-probe radius (r_{probe}) of 1.4 Å. SASAs were computed using Eq. (7):

$$A_i = P_1 S_1 + P_2 \sum_{j\in N(i)} A_{ij} + P_3 \sum_{\substack{j,k\in N(i)\\ k\in N(j)\\ k\neq j}} A_{jk} + P_4 \sum_{j\in N(i)} A_{ij} \left(\sum_{\substack{j,k\in N(i)\\ k\in N(j)\\ k\neq j}} A_{jk} \right) \tag{7}$$

In this equation, S_1 in the first term is the surface area of the isolated sphere corresponding to atom i. A_{ij} is the area of sphere i buried inside sphere j, while $N(i)$

stands for the neighbor list of i, that is, the list of spheres that overlap with sphere i. Thus, the second term involves the sum of pairwise overlaps of sphere i with its neighbors. The third term is the sum of overlaps of neighbors of i with each other. The fourth and last term is a further correction for multiple overlaps. See Weiser et al.[28] for additional details. New parameters P_1 to P_4 were derived using our own training set of proteins. See the supporting information.

Hydrogen Bond Correction

A correction term was introduced into Eq. (6) to account for the effect of hydrogen bond interactions on the ligand desolvation energies. The origin for this correction term is analyzed in the discussion section. A regression analysis was performed between the difference in desolvation energies calculated with Poisson and SCP–ISM methods and the number and type of hydrogen bonds. The number of hydrogen bonds for each decoy was deduced using a donor–acceptor cut off distance of 3.4 Å and an angle of 120° between donor–acceptor–acceptor antecedents. Definitions for donor and acceptor atoms types were taken from HBPLUS[29] in the case of protein atoms. For the ligands, all the oxygen and nitrogen atoms were visually inspected and assigned. Hydrogen bonds were then classified by the type of interactions in charged–charged (cc), neutral–charged and charged–neutral (nc), and neutral–neutral (nn). We observed that for those cases involving protonated amino groups, an additional binary variable describing the presence or absence of this group was required (npn). This is easily understood bearing in mind the way in which LCPO surface parameters were obtained, by linear fitting over the surface accessible area of the residues in the native structures (see supplementary material). Therefore, parameters for protonated nitrogen atoms where derived from lysine, whose distribution of exposed surface areas is skewed towards high values, and consequently resulting in underrepresented counts for buried protonated amino groups. All in all, the correction term is then of the form:

$$\Delta G_{\mathrm{corr}} = a + b \cdot hb_{\cdot\,\mathrm{cc}} + chb_{\cdot\,\mathrm{nc}} + dhb_{\cdot\,\mathrm{nn}} + enpn \quad (8)$$

where hb_{xx} is the number of hydrogen bonds of type xx. The final electrostatics binding free energy computed with the SCP–ISM approximation is therefore:

$$\Delta G_{\mathrm{ISM}} = \Delta G_{\mathrm{elec}} + \Delta G_{\mathrm{corr}} \quad (9)$$

Comparison With Electrostatics Binding Free Energies Obtained With Numerical Solutions of the Poisson Equation

Because the usefulness of the PE for calculating electrostatics effects in intermolecular interactions is well established,[14] the tests of the SCP–ISM approximation presented in this article are mainly intended to see how well it approximates the electrostatics binding free energy calculated using the full PE approach. The total electrostatics binding free energy (ΔG_{ele}) was calculated from the total electrostatic energy of the system obtained when solving the PE by running three consecutive calculations on the same grid: one for all the atoms in the complex ($G_{\mathrm{ele}}^{\mathrm{LR}}$), one for the ligand atoms alone ($G_{\mathrm{ele}}^{\mathrm{L}}$), and a third one for the receptor atoms alone ($G_{\mathrm{ele}}^{\mathrm{R}}$). Since the grid definition is the same in the three calculations, the artifactual grid energy cancels out when the electrostatic contribution to the binding free energy is expressed as the difference in energy between the product and the reactants:

$$\Delta G_{\mathrm{ele}} = G_{\mathrm{ele}}^{\mathrm{LR}} - (G_{\mathrm{ele}}^{\mathrm{L}} + G_{\mathrm{ele}}^{\mathrm{R}}) \quad (10)$$

Alternatively, we also considered a different description of the binding process, consisting of first desolvating the opposing surfaces of both ligand and receptor and then letting the charges of the two molecules interact, in order to isolate the three different contributions to the electrostatics binding free energy: the ligand–receptor interaction energy in the presence of the surrounding solvent ($E_{\mathrm{ele}}^{\mathrm{LR}}$), the change in solvation energy of the ligand upon binding ($\Delta G_{\mathrm{desolv}}^{\mathrm{L}}$), and the change in solvation energy of the receptor upon binding ($\Delta G_{\mathrm{desolv}}^{\mathrm{R}}$):

$$\Delta G_{\mathrm{ele}} = E_{\mathrm{ele}}^{\mathrm{LR}} + (\Delta G_{\mathrm{desolv}}^{\mathrm{L}}) + \Delta G_{\mathrm{desolv}}^{\mathrm{R}}) \quad (11)$$

The first term, the Coulombic contribution to ΔG_{elec}, was obtained by computing the product of ligand charges and the electrostatic potential generated by the protein on the ligand charge centres. On the other hand, receptor and ligand electrostatic desolvation energies were calculated in two successive steps: a first one, where a calculation is performed for receptor and ligand alone; and the second one, for the ligand, with uncharged receptor, and the receptor, with uncharged ligand.

All these calculations were performed by numerically solving the linear PE using the finite difference method as implemented in DelPhi.[30] For all calculations with DelPhi PARSE atomic radii[31] were used while charges were assigned either based on the AMBER force field[32] (protein case), or computed with MOPAC[33] (ligand case). Each complex was immersed in a cubic box occupying 65% of the total volume, with a grid spacing of 0.5 Å. Solute dielectric constant was set to 4, while the solvent and the dielectric medium was set to 80. We note that although a bulk dielectric constant of 1 or 2 is commonly employed to model the solute interior in continuum calculations, the microscopic value is environment dependent.[34] Higher values can be employed when energetics of processes involving polar sites are of interest, such as in pKa calculations or when modelling ligand–DNA interactions.[35,36] Although a matter of debate, and since ligand binding sites are moderately polar, we[5] and

TABLE I. Summary of the Decoy Dataset Used in this Paper

PDB ID	Description	No.of atoms (heavy)	No. of decoys	RMSD min–max (Å)	Total charge	No. of HBD	No. of HBA
1HVI	HIV-1 protease	116 (58)	43	0.3–15.1	0	6	8
1HVJ	HIV-1 protease	115 (57)	65	0.2–14.1	0	5	7
1HVK	HIV-1 protease	116 (58)	50	0.3–15.8	0	6	8
1HIH	HIV-1 protease	91 (41)	57	0.4–15.8	0	5	5
1HPX	HIV-1 protease	87 (46)	85	0.1–9.6	0	4	6
1MCJ	Immunoglobulin	60 (32)	44	0.3–11.3	0	4	6
1RBP	Retinol binding protein	51 (21)	40	0.2–3.4	0	1	1
2UPJ	HIV-1 protease	81 (41)	48	0.2–10.1	0	3	4
1ABE	L-arabinose binding protein	20 (10)	60	0.1–3.8	0	4	4
1AJX	HIV-1 protease	74 (40)	64	0.1–3.7	0	3	3
5ABP	L-arabinose binding protein	22 (12)	34	0.1–2.5	0	5	5
1DBB	Immunoglobulin	53 (23)	3	1.4–6.9	0	0	2
1FKG	FK506 binding protein	68 (33)	20	0.4–8.5	0	0	3
1FKH	FK506 binding protein	74 (33)	20	0.4–8.0	0	0	3
1MRK	α-trichosanthin	32 (19)	9	0.5–4.5	0	6	8
1STP	Biotin binding protein	31 (16)	78	0.1–8.5	−1	3	7
1B9V	Influenza virus neuraminidase	50 (25)	20	0.4–5.0	−1	0	6
1DBM	Immunoglobulin	64 (31)	20	0.5–3.3	−1	0	5
1TNG	Trypsin	24 (8)	3	0.7–1.8	+1	3	0
1TNI	Trypsin	27 (11)	20	0.6–3.2	+1	3	0
1TNK	Trypsin	24 (10)	20	0.6–2.9	+1	3	0
1TNL	Trypsin	22 (10)	3	1.6–2.1	+1	3	0
1BMA	Trypsin	73 (37)	20	0.6–4.2	+1	0	3

See Methods for additional details.

others[2,37,38] consider more appropriate a value of four. The dielectric boundary was calculated using a solvent probe radius of 1.4 Å. A minimum separation of 11 Å was allowed between any solute atom and the box walls. The potentials at the grid points delimiting the box were calculated analytically by treating each charge atom as a Debye–Hückel sphere.

Training Set of Complexes Used in the Development of the SCP–ISM Model

A training set was formed with 23 different proteins (see Table I) summing up a total of 826 decoys. Some of the decoys (those for 1HVI, 1HVJ, 1HVK, 1HIH, 1HPX, 1MCJ, 1RBP, 1UPJ, 1ABE, 1AJX, 5ABP, and 1STP) were taken directly from LPDB database.[39] Atom types and hydrogen atoms definitions where translated from the CHARMm force field into their equivalents in AMBER. Hydrogen atoms were removed, and added back with protonate program from the AMBER 8.0 package, were also charge and radii for all the atoms in the proteins assigned. For the ligands, AMBER radii and semiempirical charges fitted to electrostatic potentials with MOPAC[33] were used (keywords 1SCF, MNDO, ESP, and DIPOLE). The rest of the decoys (those for 1DBB, 1FKG, 1FKH, 1B9V, 1DBM, 1TNG, 1TNI, 1TNK, 1TNL, and 1BMA) were generated by us using our in-house docking program.[11,40] They were prepared in exactly the same way as stated before for LPDB data set.

Parametrization and Validation of the SCP–ISM Model

Description of the SCP–ISM parameters and their optimized values are listed in Table II, and in Table I of the supplementary material. We arrived at these parameters by performing exhaustive searches in a subset of the parameter space, using the quadratic error between the SCP–ISM and PE values as fitness function. The following parameters were systematically modified: scale factor for atomic radii (from 0.3 to 1.3 Å in 0.1 Å intervals; enlargement factor $h_{(+,-)}$, from 0.35 to 0.85 Å in 0.1 Å intervals; enlargement factor g, from 0.0 to 1.0 Å in 0.1 Å intervals; λ, from 0.001 to 0.020 in 0.001 intervals; fixed values were used for ε (78.39) and the solvent-probe radius (1.4 Å). To test for the robustness of the fitted parameters, a Leave One Out (LOO) procedure was applied. One by one, all decoys from a single target were removed from the complete set, the model was rebuilt (internal test) and the excluded decoys blindly predicted (external validation). At each step the RMSD and regression coefficients between the values for the external set and their Poisson counterparts were compared. Calculations were carried out with the R package (http://www.r-project.org/).

RESULTS

A key aspect of any method development involving parameter fitting is the training set of examples employed in the fitting process. Here, 826 different decoys have

TABLE II. Relevant Parameters Used in our SCP-ISM Model

Parameter	Value
Atomic Radii[a]	
Scale	0.6
$h_{(+,-)}$	0.85 (0.35)
g	0.5
Solvent[b]	
$\lambda_{(+,-)}$	0.013 (0.007)
ε	78.39
r_{probe}	1.4
Hydrogen bond[c]	
r	3.4
α	120
a	−0.29
b	1.02
c	−0.25
d	0.02
e	10.52
PE/SCP-ISM[d]	
A	5.3
B	0.09
C	1.06
D	0.97

[a]Initial AMBER radii for all the atoms are scaled down by the scale parameter. $h_{(+,-)}$ and g account for the enlargement of the radii when immersed in solvent. $h_{(+,-)}$ depends on the type of charge (0.85 for positive and 0.35 for negative), g is independent of the charge and it is always equals to 0.5, both in Å.
[b]Solvent related parameters are the slope of sigmoidal dielectric function ($\lambda_{(+,-)}$) with two values: 0.013 for all of the atoms except for those with a formal positive charge, and 0.007 for these last ones. ε is the dielectric constant of the bulk solvent and r_{probe} is the radius, in Å, of the water probe molecule employed to calculate the solvent accessible surface.
[c]Hydrogen bond parameters: r and α are the minimum radii (in Å) and angle (in degrees) between donor and acceptor, and donor–acceptor–acceptor antecedents, respectively. a, b, c, and d corresponds to the fitted parameters to account for the hydrogen bond correction (see Methods) according to the equation: $\Delta G_{\text{corr}} = a + b \cdot hb_{\text{cc}} + c \cdot hb_{\text{nc}} + d \cdot hb_{\text{nn}} + e \cdot npn$.
[d]Final parameters obtained from the comparison between PE and SCP-ISM according to the equation: $PB = C + D(A \exp(B(ISM))$, see Table III and main text.

been employed to test the new solvation model described in this paper (Table I). Decoys cover a wide range (around 40 kcal/mol) of electrostatics binding free energies. There is also ample structural variety in the complexes employed, both in protein architectures as well as in the ligand functional groups. The set includes neutral, zwitterionic, as well as formally charged (both positively and negatively) ligands. Finally, there is also a considerable number of representative orientations of each complex within each binding site, about 20 on average, covering a large spectrum of RMSD values, ranging from close natives to more than 10 Å RMSD. Thus, we feel confident that our dataset, while not perfect, has enough variety to warrant the generality of our results. The LOO tests (see below) seem to confirm this.

A second important aspect is the number of parameters to fit. Our SCP–ISM model has a relatively small number of generic parameters (Table II). Leaving aside the parameters related to charges, radii, and those involved in the surface calculation, the model has a total of 17 or 19 parameters (Table II), depending on the exact choice of the model (see below). The basic model, corresponding to Eq. (6), contains eight parameters: h_+, h_-, g, λ_+, λ_-, ε, r_{probe}, and the scale factor of the atomic radii (see Methods and Table III for definitions and values, respectively). Six out of these eight were considered for optimization, while ε and r_{probe} were kept fixed. In order to properly reproduce PE results with the version of the SCP–ISM model developed here, a hydrogen bond correction was deemed necessary. The definition of the hydrogen bond itself added two parameters to the model. In order to obtain a reasonable fit, an additional set of five parameters, accounting for the nature of the hydrogen bond, were required (Table II). A comparison of the SCP–ISM and PE electrostatics binding free energies with the optimized set of parameters is shown in Figure 2(a). The direct comparison suggests that an exponential-type relationship exists between them. The reason for this dependence is unclear to us, and its investigation will be left for future work. Noting this dependence, two different fittings between the two sets of data were attempted: an exponential one (model1, Poisson $= A \exp(B \times$ SCP–ISM)); and a second linear fitting of the exponential model (i.e., model1) was also investigated (model2, Poisson $= C + D(A \exp(B \times$ SCP–ISM)), to account for systematic deviations from the exponential behavior. Adding the fitting parameters (two or four, depending on which model is used) yields the final set of 17 or 19 parameters comprising our complete SCP–ISM model.

Results for the two fittings (model1 and model2), including LOO tests, can be found in Table III. The results for the ALL row correspond to the standard crossvalidation case, where each set of decoys for a given protein were removed, a model derived, and based on the model the removed complexes were predicted. In this case, as expected, the crossvalidated RMSD is slightly larger than fitted one. The rest of the rows in Table III correspond to the partial results of the ALL case. For example, the first row shows the in-fitting columns, the model obtained after removing decoys for 1HVI, with all other decoys as training set. The LOO crossvalidation columns show the result of this model as applied to the 1HVI decoys. A slightly better RMSD value is obtained with model2 as compared to model1 with 4.16 versus 4.20 kcal/mol, respectively. A crossvalidated r^2, or q^2 of 0.81, a slope of 0.97, and an intercept of 1.06 kcal/mol was found for the best model. A comparison with PE data is shown in Figure 2(b). The similarity for the RMSD values between the fitted and crossvalidated electrostatics binding free energies shown in the LOO tests (4.20 versus 4.33, see Table III) indicates that there is no evidence of overfitting. Thus, our results are likely to hold using different sets of complexes. Satisfactory results are also found for the different components of the electrostatics binding free energy. The squared corre-

TABLE III. Evaluation of the SCP-ISM Model

	Fitting								LOO crossvalidation		
	PE = A exp(B × ISM)				PE = C + D(A exp(B × ISM))						
Complex	A	B	r^2	RMSD[a]	C	D	r^2	RMSD[a]	Model1 RMSD[a]	Model2 RMSD[a]	q^2
1HVI	5.33	0.09	0.80	4.11	0.79	0.99	0.86	4.07	5.88	5.70	0.70
1HVJ	5.34	0.09	0.80	4.04	0.85	0.98	0.87	4.00	6.14	5.96	0.40
1HVK	5.31	0.09	0.80	4.17	0.85	0.98	0.86	4.14	4.76	4.61	0.81
1HIH	5.32	0.09	0.80	4.15	0.82	0.98	0.87	4.12	4.96	4.69	0.76
1HPX	5.30	0.09	0.80	4.14	0.84	0.98	0.87	4.10	4.69	4.53	0.50
1MCJ	5.47	0.09	0.83	4.13	1.29	0.96	0.87	4.08	5.21	5.76	0.71
1RBP	5.52	0.09	0.81	4.22	1.10	0.97	0.86	4.18	3.37	4.12	0.44
2UPJ	5.25	0.09	0.81	4.16	0.92	0.98	0.87	4.12	5.04	5.00	0.36
1ABE	4.69	0.09	0.85	4.29	1.54	0.94	0.87	4.22	5.19	4.19	0.93
1AJX	5.46	0.09	0.81	4.28	1.31	0.96	0.86	4.23	2.46	3.18	0.73
5ABP	4.86	0.09	0.84	4.23	1.46	0.95	0.87	4.16	5.96	4.97	0.86
1DBB	5.30	0.09	0.81	4.21	1.06	0.97	0.86	4.17	1.55	2.28	0.98
1FKG	5.32	0.09	0.81	4.25	1.12	0.97	0.86	4.20	1.48	2.16	0.66
1FKH	5.32	0.09	0.81	4.25	1.11	0.97	0.86	4.20	1.26	1.98	0.95
1MRK	5.28	0.09	0.81	4.20	1.06	0.97	0.87	4.16	2.44	2.17	0.62
1STP	5.57	0.09	0.82	4.25	1.06	0.97	0.86	4.21	2.59	3.22	0.80
1B9V	5.29	0.09	0.81	4.24	1.00	0.98	0.87	4.19	4.38	3.92	0.92
1DBM	5.24	0.09	0.82	4.23	1.05	0.97	0.87	4.18	3.67	3.08	0.68
1TNG	5.32	0.09	0.81	4.20	1.03	0.97	0.86	4.16	2.59	3.41	0.87
1TNI	5.41	0.09	0.80	4.23	0.91	0.98	0.86	4.19	1.17	1.77	0.68
1TNK	5.37	0.09	0.81	4.22	1.07	0.97	0.86	4.18	2.49	3.08	0.69
1TNL	5.30	0.09	0.81	4.21	1.07	0.97	0.86	4.17	1.36	1.80	0.73
1BMA	5.27	0.09	0.81	4.24	1.12	0.97	0.86	4.19	1.85	1.53	0.19
ALL	5.30	0.09	0.81	***4.20***	1.06	0.97	0.87	***4.16***	***4.40***	***4.33***	***0.81***

Fitted (for model1, PE = A exp(B × ISM); and model2, PE = C + D(A exp(B × ISM)), see main text for more details) as well as crossvalidated results are shown. The results for the ALL row correspond to the standard crossvalidation case, where each set of decoys for a given protein were removed, a model derived, and based on the model the removed complexes were predicted. Partial results obtained excluding decoys of specific system shown in the corresponding row during the fitting phase are also presented. In these cases the fitting values correspond to those obtained with the model generated using the rest of the decoys. Reported q^2 values correspond to model2.
[a]Root mean square deviation (in kcal/mol) between the PE and SCP-ISM results.

lation coefficients oscillate between 0.79 and 0.88 (see Fig. 3). Slopes are also close to unity (1.13 for the coulombic term and 1.01 for the ligand desolvation term), except for receptor desolvation term (1.72) (see Fig. 3). Intercepts are close to zero in all cases (see Fig. 3). Thus, not only the total energy, but also its contributions are well reproduced by the SCP–ISM model.

As to the computation times, Figure 4 shows a histogram of the computing times for the 826 decoys. The average computing time is 40 ms, with a mode at 30 ms. These times are well below most GB approaches. The dependency of the computing times with the number of heavy atoms in the ligand can be found in Figure 5, which shows a "box and whisker plot". For each bin, the data is divided into four intervals: a quarter of the data (25% percentile) is between the lower-lying whisker and the baseline of the box, another quarter is between this line and the median line, other quarter is between the median line and the top line of the box, and finally, the last quarter is between this last line and the end of the higher-lying whisker. An approximately linear dependency between number of heavy atoms and computing time is observed.

DISCUSSION

Herein we present a new model for the fast calculation of electrostatics binding free energies in protein–ligand binding problems. The formulation is a modification of the original model proposed by Hassan et al. to treat electrostatics effects in proteins.[25,26] As in their case, no boundary surface between the high (solvent) and low (receptor and/or ligand) dielectric media is required. This is achieved by defining the dielectric function in a sigmoidal distance-dependent manner. Similarly, the effective Born radii are readily computed only from the exposed surface accessible area of the atom of interest. In order to properly account for the PE results, a hydrogen bond correction term in the SCP–ISM model was necessary. At face value, this requirement may seem odd, since both models (PE and SCP–ISM) attempt describe the same process, the electrostatics binding free energy, and hydrogen bonding has a strong electrostatics component which should be captured by the model. However, recent studies by McCammon and coworkers[41] have clearly established that the use of atom-centered surfaces, such as the ones employed here the LCPO method, or the presence of smooth transitions between

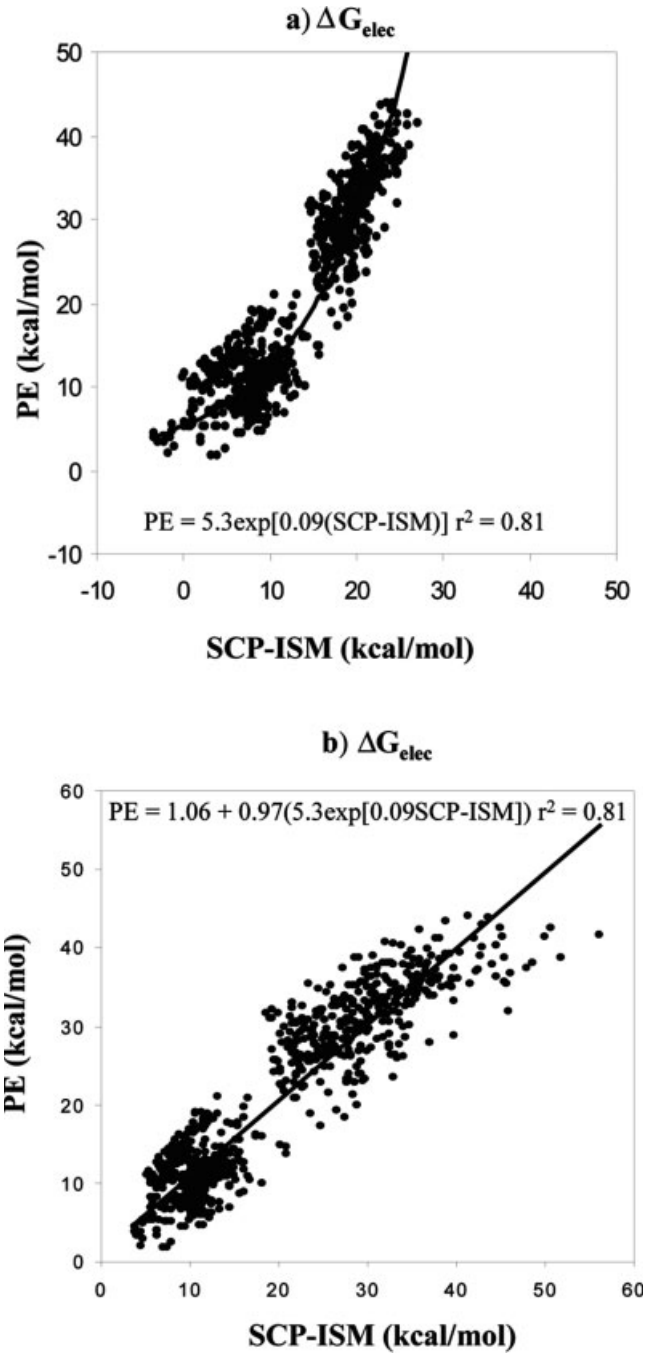

Fig. 2. Correlation between the total electrostatics binding free energy (in kcal/mol), as obtained by numerically solving the PE (*y*-axis), and by the SCP–ISM method (*x*-axis). Each point represents the corresponding energy pair for a different decoy. About 826 different decoys (summarized in Table I) have been employed. (**a**) Direct correlation. (**b**) Correlation after logarithmic correction. See text for details.

low and high dielectric regions, as in our sigmoidal dielectric function, increase the fraction of interstitial high dielectric regions in the protein interior. The presence of these regions has been shown to suppress the electrostatic free energy barriers characteristic of hydro-

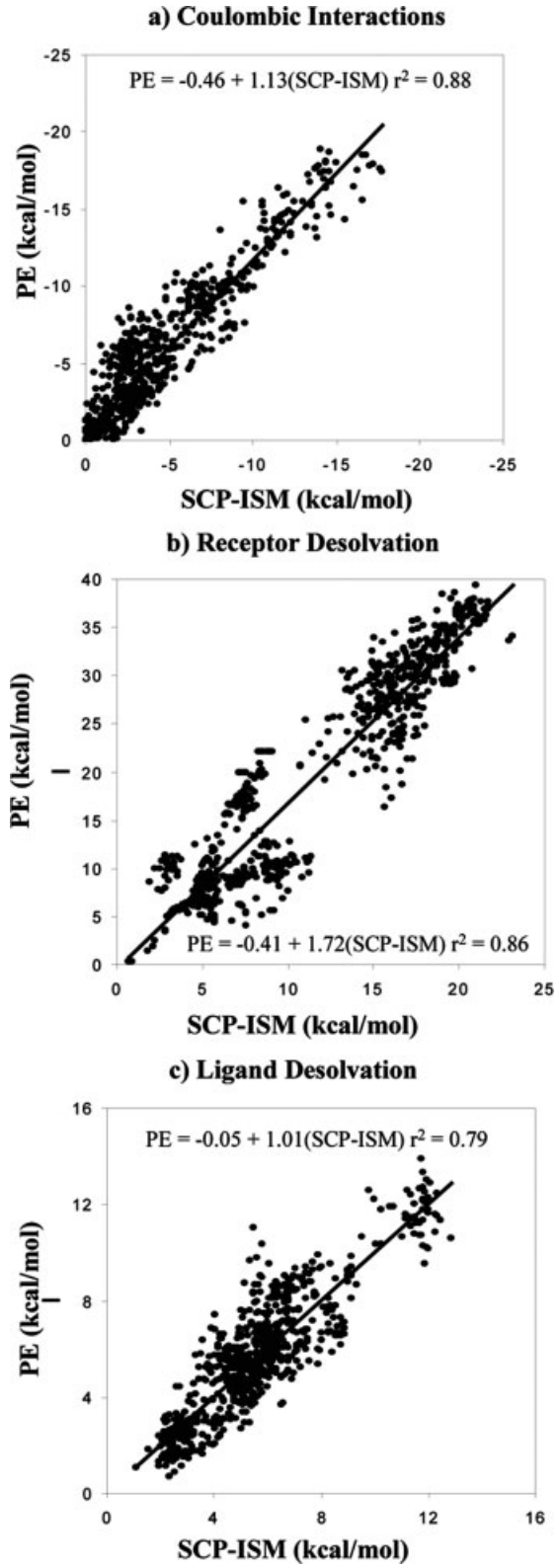

Fig. 3. Comparison of the different energy contributions to the electrostatics binding free energy, as obtained by solving the PE and by using SCP–ISM. (**a**) Coulombic contribution; (**b**) receptor desolvation, and (**c**) ligand desolvation. Each point represents the corresponding energy pair for a different decoy. About 826 different decoys were used.

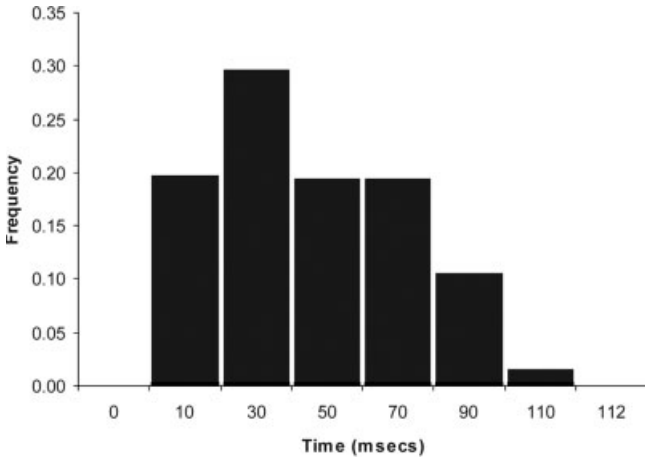

Fig. 4. Frequency distribution of the SCP–ISM computing time required to obtain the electrostatics binding free energy per decoy. The *x*-axis shows the range of computing times (in ms), while the *y*-axis shows the corresponding frequency in the set. The set of 826 decoys summarized in Table I has been employed. Calculations were performed in a 3.0 GHz Pentium IV computer.

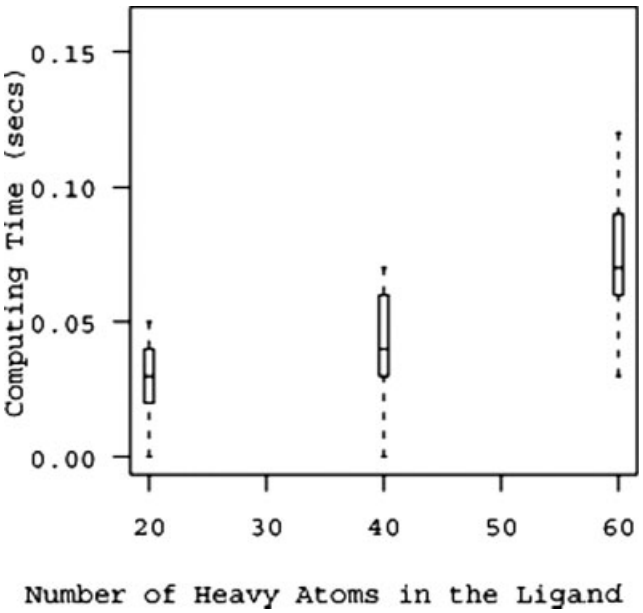

Fig. 5. Box and whisker plot showing the relationship between ligand size and computing time. The number of heavy atoms in the ligand is plotted versus the corresponding computing time. The set of 826 decoys summarized in Table I has been employed. Calculations were performed in a 3.0 GHz Pentium IV computer. See text for details.

gen bond formation when compared to atomistic potential of mean force simulations, leading to overestimation of solvation energies, particularly for groups involved in hydrogen bond interactions.[41] We speculate that our empirical hydrogen bond term might act as an "ad-hoc" correction to account for this effect. Nevertheless, future studies will address this matter in detail.

On the basis of studies with a large set of 826 decoys, covering substantial structural variety both in targets and ligands, satisfactory results have been obtained with the SCP–ISM model. The new method has a squared crossvalidated correlation coefficient with the electrostatics binding free energies obtained with the PE of 0.81, a slope of 0.97, an intercept of 1.06 kcal/mol, and a RMSD of about 4.33 kcal/mol [Fig. 2(b) and Table III]. The different contributions to the electrostatics binding free energies are also reproduced with similar accuracy (see Fig. 3). These results compare well with those recently obtained by Liu and Zou,[37] who studied the ability of GB to reproduce electrostatics binding free energies computed with PE. In their study, using crystal structures for 15 complexes in fitting and another 15 in crossvalidation, Liu and Zou obtained a squared correlation coefficient of 0.81 and a RMSD of 4.05 kcal/mol in fitting phase, while in crossvalidation the values obtained were 0.81 and 5.14, respectively. The comparison suggests that GB and SCP–ISM achieve similar performances in modelling protein–ligand binding energetics. Nevertheless, a log transformation in the SCP–ISM model was required to linearly fit the total electrostatics binding free energy to the PE results. We have found that the reason for this non linear effect rests on a relative (i.e., with respect the reference value computed with PE) overestimation of the desolvation free energy for proteins hosting long, narrow hydrophobic channels in the ligand binding site, such as the retinol binding protein or the biotin binding protein (see Table I for a description of the complexes). The reasons for this overestimation are unclear and are being investigated at present. On the other hand, we wish to emphasize that PE calculations are employed here more as a guideline, upon which our method should qualitatively conform, than as a reference quantitative golden standard. The PE method itself depends on a number of empirical parameters such as internal and external dielectric constants, boundary definitions, and so forth, which are not uniquely determined and are subjected to debate. For this reason we have not attempted to further "improve" fitting by introducing new sets of parameters or more empiricism into our model, and we have restricted our parameter search to physically meaningful quantities. The fact that in these conditions we obtain reasonable fits attests to the physical soundness of the SCP–ISM model.

The mean pose calculation time for the SCP–ISM model is about 30–40 ms (see Fig. 4), and an approximately linear relationship between ligand size and computing time is observed (see Fig. 5). This time is expected to be reduced further by employing a look-up table containing neighbouring atoms at each grid point, accelerating the calculation of the effective Born radii. Thus, the new method is shown, both in terms of timing and accuracy, to be good enough to be implemented directly into a docking algorithm, and compares favorably with other approaches. For example, the GB method implemented originally by Kuntz and coworkers in DOCK required ~10 s per complex on a SGI Octane workstation.[17] The same group later proposed a pair-

wise approach to compute the Born radii, reducing the computational time to 0.5 s per complex.[20] These timings prevent its direct use in the docking step, remaining only as a post-DOCK filter. On the other hand, Caflisch and coworkers proposed a simplified continuum method based on the assumption that electrostatic desolvation can be approximated by the removal of the first layer of water molecules at the binding interface, and the coulombic contribution can be approached by a distance dependent dielectric model.[23] Precomputation of the energy contributions on a set of grids allowed the authors to estimate the electrostatics binding free energy in solution in about 3–4 ms for fragments of 5–10 heavy atoms on a 550 MHz Pentium III. However, their method is restricted to docking of rigid molecules, since both ligand and receptor need to be grid-preprocessed, while our SCP–ISM can be employed for both rigid and flexible docking cases. This limits the applicability of the method of Caflisch and coworkers mainly to the docking of small rigid fragments in rigid binding sites. The accuracy of the total fitted electrostatics binding free energy obtained with the method of Caflisch and coworkers is also slightly worst than the one obtained with SCP–ISM, as judged by the squared correlation coefficients when compared with PE results ($\sim$0.75 vs. $\sim$0.81, respectively). Contributions to the electrostatics binding free energy are similarly reproduced by both methods (r^2 of $\sim$0.81 in both cases), but the SCP–ISM provides slopes close to 1.0 (see Fig. 3), while in the method of Caflisch and coworkers, the slopes are larger and show more dispersion (from 1.49 to 2.95).

In summary, although some descriptions are available to consider solvent effects in protein–ligand binding, they are either time consuming, inaccurate, or only applicable in very restricted conditions. This limits their usefulness in virtual screening projects, where millions of molecules with different conformers, tautomeric, and protonation states, need to be considered. The method presented here is a step in the direction of incorporating realistic, but fast, solvent models in large scale docking. We are currently incorporating the ISM method in our in-house docking program. Impact of the new electrostatics model in docking and virtual screening is being evaluated and will be presented in due time.

ACKNOWLEDGMENTS

Generous allocation of computer time at the Barcelona Supercomputer Center is gratefully acknowledged.

REFERENCES

1. Mohan V, Gibbs AC, Cummings MD, Jaeger EP, DesJarlais RL. Docking: successes and challenges. Curr Pharm Des 2005;11: 323–333.
2. Schwarzl SM, Tschopp TB, Smith JC, Fischer S. Can the calculation of ligand binding free energies be improved with continuum solvent electrostatics and an ideal-gas entropy correction? J Comput Chem 2002;23:1143–1149.
3. Sims PA, McCammon JA, Wong CF. A computational model of binding thermodynamics: the design of cyclin-dependent kinase 2 inhibitors. J Med Chem 2003;46:3314–3325.
4. Wang W, Donini O, Reyes CM, Kollman PA. Biomolecular simulations: recent developments in force fields, simulations of enzyme catalysis, protein–ligand, protein–protein, and protein–nucleic acid noncovalent interactions. Annu Rev Biophys Biomol Struct 2001;30:211–243.
5. Checa A, Ortiz AR, de Pascual-Teresa B, Gago F. Assessment of solvation effects on calculated binding affinity differences: trypsin inhibition by flavonoids as a model system for congeneric series. J Med Chem 1997;40:4136–4145.
6. Bernacki K, Kalyanaraman C, Jacobson MP. Virtual ligand screening against *Escherichia coli* dihydrofolate reductase: improving docking enrichment using physics-based methods. J Biomol Screen 2005;10:675–681.
7. Kalyanaraman C, Bernacki K, Jacobson MP. Virtual screening against highly charged active sites: identifying substrates of α–β barrel enzymes. Biochemistry 2005;44:2059–2071.
8. Shoichet BK, Leach AR, Kuntz ID. Ligand solvation in molecular docking. Proteins Struct Funct Genet 1999;34:4–16.
9. Huang D, Caflisch A. Efficient evaluation of binding free energy using continuum electrostatics solvation. J Med Chem 2004;47: 5791–5797.
10. Kuhn B, Gerber P, Schulz-Gasch T, Stahl M. Validation and use of the MM-PBSA approach for drug discovery. J Med Chem 2005;48:4040–4048.
11. Perez C, Ortiz AR. Evaluation of docking functions for protein–ligand docking. J Med Chem 2001;44:3768–3785.
12. Ferrara P, Gohlke H, Price DJ, Klebe G, Brooks CL, III. Assessing scoring functions for protein–ligand interactions. J Med Chem 2004;47:3032–3047.
13. Orozco M, Luque FJ. Theoretical methods for the description of the solvent effect in biomolecular systems. Chem Rev 2000;100: 4187–4225.
14. Honig B, Nicholls A. Classical electrostatics in biology and chemistry. Science 1995;268:1144–1149.
15. Bashford D, Case DA. Generalized born models of macromolecular solvation effects. Annu Rev Phys Chem 2000;51:129–152.
16. Still WC, Tempczyk A, Hawley RC, Hendrickson T. Semianalytical treatment of solvation for molecular mechanics and dynamics. J Am Chem Soc 1990;112:6127–6129.
17. Zou X, Sun Y, Kuntz ID. Inclusion of solvation in ligand binding free energy calculations using the generalized-born model. J Am Chem Soc 1999;121:8033–8043.
18. Majeux N, Scarsi M, Apostolakis J, Ehrhardt C, Caflisch A. Exhaustive docking of molecular fragments with electrostatic solvation. Proteins 1999;37:88–105.
19. Taylor RD, Essex JW, Jewsbury PJ. FDS: flexible ligand and receptor docking with a continuum solvent model and soft-core energy function. J Comput Chem 2003;24:1637–1656.
20. Liu H-Y, Zou X, Kuntz ID. Pairwise GB/SA scoring function for structure-based drug design. J Phys Chem B 2004;108:5453–5462.
21. Wang J, Kang X, Kuntz ID, Kollman PA. Hierarchical database screenings for HIV-1 reverse transcriptase using a pharmacophore model, rigid docking, solvation docking, and MM-PB/SA. J Med Chem 2005;48:2432–2444.
22. Arora N, Bashford D. Solvation energy density occlusion approximation for evaluation of desolvation penalties in biomolecular interactions. Proteins 2001;43:12–27.
23. Majeux N, Scarsi M, Caflisch A. Efficient electrostatic solvation model for protein-fragment docking. Proteins 2001;42:256–268.
24. Hassan SA, Mehler EL. A critical analysis of continuum electrostatics: the screened Coulomb potential–implicit solvent model and the study of the alanine dipeptide and discrimination of misfolded structures of proteins. Proteins 2002;47:45–61.
25. Hassan SA, Guarnieri F, Mehler EL. Characterization of hydrogen bonding in a continuum solvent model. J Phys Chem B 2000;104:6490–6498.
26. Hassan SA, Guarnieri F, Mehler EL. General treatment of solvent effects based on screened Coulomb potentials. J Phys Chem B 2000;104:6478–6489.
27. Morris GM, Goodsell DS, Huey R, Olson AJ. Distributed automated docking of flexible ligands to proteins: parallel applica-

tions of AutoDock 2.4. J Comput Aided Mol Des 1996;10:293–304.
28. Weiser J, Shenkin PS, Still WC. Approximate atomic surfaces from linear combinations of pairwise overlaps (LCPO). J Comput Chem 1999;20:217–230.
29. McDonald IK, Thornton JM. Satisfying hydrogen bonding potential in proteins. J Mol Biol 1994;238:777–793.
30. Rocchia W, Sridharan S, Nicholls A, Alexov E, Chiabrera A, Honig B. Rapid grid-based construction of the molecular surface and the use of induced surface charge to calculate reaction field energies: applications to the molecular systems and geometric objects. J Comput Chem 2002;23:128–137.
31. Sitkoff DS, Sharp KA, Honig B. Accurate calculation of hydration free energies using macroscopic solvent models. J Phys Chem 1994;98:1978–1988.
32. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. J Am Chem Soc 1995;117:5179–5197.
33. Stewart JJ. MOPAC: a semiempirical molecular orbital program. J Comput Aided Mol Des 1990;4:1–105.
34. Schutz CN, Warshel A. What are the dielectric "constants" of proteins and how to validate electrostatic models? Proteins 2001;44:400–417.
35. Misra VK, Sharp KA, Friedman RA, Honig B. Salt effects on ligand–DNA binding. Minor groove binding antibiotics. J Mol Biol 1994;238:245–263.
36. Misra VK, Honig B. On the magnitude of the electrostatic contribution to ligand–DNA interactions. Proc Natl Acad Sci USA 1995;92:4691–4695.
37. Liu HY, Zou X. Electrostatics of ligand binding: parametrization of the generalized Born model and comparison with the Poisson–Boltzmann approach. J Phys Chem B Condens Matter Mater Surf Interfaces Biophys 2006;110:9304–9313.
38. Grater F, Schwarzl SM, Dejaegere A, Fischer S, Smith JC. Protein/ligand binding free energies calculated with quantum mechanics/molecular mechanics. J Phys Chem B Condens Matter Mater Surf Interfaces Biophys 2005;109:10474–10483.
39. Roche O, Kiyama R, Brooks CL, III. Ligand–protein database: linking protein–ligand complex structures to binding data. J Med Chem 2001;44:3592–3598.
40. Murcia M, Ortiz AR. Virtual screening with flexible docking and COMBINE-based models. Application to a series of factor Xa inhibitors. J Med Chem 2004;47:805–820.
41. Swanson JM, Mongan J, McCammon JA. Limitations of atom-centered dielectric functions in implicit solvent models. J Phys Chem B Condens Matter Mater Surf Interfaces Biophys 2005;109:14769–14772.