J Comput Aided Mol Des (2011) 25:913–930
DOI 10.1007/s10822-011-9471-8

Charge density distributions derived from smoothed electrostatic potential functions: design of protein reduced point charge models

Laurence Leherte · Daniel P. Vercauteren

Received: 19 January 2011 / Accepted: 30 August 2011 / Published online: 14 September 2011

Abstract To generate reduced point charge models of proteins, we developed an original approach to hierarchically locate extrema in charge density distribution functions built from the Poisson equation applied to smoothed molecular electrostatic potential (MEP) functions. A charge fitting program was used to assign charge values to the so-obtained reduced representations. In continuation to a previous work, the Amber99 force field was selected. To easily generate reduced point charge models for protein structures, a library of amino acid templates was designed. Applications to four small peptides, a set of 53 protein structures, and four KcsA ion channel models, are presented. Electrostatic potential and solvation free energy values generated by the reduced models are compared with the corresponding values obtained using the original set of atomic charges. Results are in closer agreement with the original all-atom electrostatic properties than those obtained with a previous reduced model that was directly built from the smoothed MEP functions [Leherte and Vercauteren in J Chem Theory Comput 5:3279–3298, 2009].

Keywords Molecular electrostatic potential · Charge density · Smoothing · Point charge model · Coarse grain · Protein

Abbreviations

AA	Amino acid
AMBER	Assisted model building and energy refinement
APBS	Adaptive Poisson-Boltzmann Solver
CD	Charge density
CG	Coarse grain(ed)
COM	Center of mass
DNA	Desoxyribonucleic acid
ECEPP	Empirical conformational energy program for peptides
ED	Electron density
FF	Force field
GA	Genetic algorithm
LJ	Lennard-Jones
MC	Monte Carlo
MD	Molecular dynamics
MM	Molecular mechanics
MEP	Molecular electrostatic potential
PB	Poisson-Boltzmann
PDB	Protein data bank
rmsd	Root mean square deviation
SA	Simulated annealing
SLIRP	Structural library of intrinsic residue propensities
SMMP	Simple molecular mechanics for proteins
3D	Three-dimensional

Introduction

Electrostatic interactions involving charged groups in a protein may be of great importance for particular structures, e.g., in the binding of ligands [1], in the interaction

Electronic supplementary material The online version of this article (doi:10.1007/s10822-011-9471-8) contains supplementary material, which is available to authorized users.

L. Leherte (✉) · D. P. Vercauteren
Laboratoire de Physico-Chimie Informatique, Unité de Chimie Physique Théorique et Structurale, University of Namur (FUNDP), Rue de Bruxelles 61, 5000 Namur, Belgium
e-mail: laurence.leherte@fundp.ac.be

and assembly of proteins [2], in the (de)stabilization of protein folds [3, 4], or in the control of ion transportation through membranes [5]. These phenomena can, additionally, be strongly influenced by the surrounding dielectric medium, i.e., the solvent. Electrostatic interactions are thus of major importance in computational approaches such as protein dynamics simulations [6, 7], protein docking [8–10], or protein design [11–14]. However, the modeling of the solvent through explicit water molecules may be highly time-consuming for large biomolecular systems. Continuum model theories, such as Poisson-Boltzmann (PB), are consequently largely used to evaluate, e.g., solvation/desolvation energies [11, 15, 16]. The solvent can also be described using a dipolar lattice as detailed in [17]. A generalized procedure, such as the dipolar Poisson-Boltzmann-Langevin equation [18] allows to model the solvent as a set of interacting dipoles with a non-uniform dielectric constant at the solute/solvent interface that depends on the location, the electrostatic potential, and the electric field. The design of reduced, or coarse-grained (CG), descriptions is also a current field of research in the modeling of large biomolecules [19, 20]. The reduction of the number of degrees of freedom, on one hand, leads to a smoothing of the potential energy hypersurface, and on the other hand, to a decrease in the number of inter-particle interactions required in large systems modeled over long time scales.

Common approaches used to design a reduced description of a protein consist in replacing groups of atoms into single interaction sites. These sites can be located on atoms [21, 22], on geometric centers or centers of mass (COM) of groups of atoms in an amino acid (AA) [23–28], or on a combination of both [29–32]. Zhang et al. [33] also proposed a method to define CGs as COMs of groups of contiguous Cα atoms that move in a highly correlated fashion.

The association of point charge values with CGs can be done using diverse techniques. For examples, Gabdoulline et al. [34] used a model that consisted of a small number of point charges suitable for the description of the intermolecular electrostatic interactions. As later applied by Basdevant et al. [24], these charges were derived from a fitting procedure applied to reproduce the molecular electrostatic potential (MEP) obtained by solving the PB equation. The mimick of all-atom electrostatic interactions using a limited set of point charges was also proposed by Berardi et al. [35] who applied a genetic algorithm (GA) to determine the location and values of a given number of charges for molecules involved in liquid-crystalline materials. A common approach to assign charges to the CGs of AAs consists in describing each structure by unit or nul electric charge values, as achieved in the work of Skepö et al. [36], in the MARTINI force field (FF) [26], or in the more recent work of DeVane et al. [27] where the initial unit values ±1 were further scaled down to compensate for the solvent that is represented by uncharged spheres. In other models, each AA is represented by dipolar/multipolar centers [25, 37, 38]. Reviews on the progresses of CG models can be found in additional references [39–44].

Our first studies on the interaction potential of CG molecular representations were dedicated to a ligand-DNA system [45]. Based on a topological analysis of the promolecular electron density (ED) of both a DNA sequence and a drug molecule, calculated at a crystallographic resolution of 3 Å, a Lennard-Jones (LJ) type interaction potential was implemented to evaluate the interaction energy of a spherical probe and a DNA structure represented by a limited number of ellipsoids centered on the peaks (local maxima) of the ED distribution function. In a further work, this strategy was expanded to dock protein–protein and protein–DNA partners using a GA [46–48]. The electrostatic interaction potential consisted of a summation over Coulomb terms involving unit ±1 charges assigned to residues Arg, Asp, Glu, and Lys. Our first attempt to assign non-unit electric charges to ED peaks was achieved in a further work applied to the adenine binding site of the human Aldose reductase protein structure and its cofactor $NADP^+$ [49]. ED peaks were located by following the atom trajectories in progressively smoothed ED distributions using a merging/clustering algorithm [50]. Each maxima could be associated with a molecular fragment whose charge was determined by summing over the charge values of its constituting atoms.

Following the development of our approach to hierarchically decompose a protein structure into fragments from its ED distribution [49, 50], the method was applied to MEPs, calculated from point charges as implemented in well-known FFs [51]. The procedure allowed to locate minima and maxima in a smoothed MEP to design AA reduced point charge models. While efficient in approximating all-atom MEPs of a potassium ion channel, the models however presented drawbacks. Indeed, the obtained CG points were located away from the skeleton of the AA structure. This involves that both their location and charge value can be dependent on the AA conformation, as presented later in the 'Side chain modeling' part of the paper.

In the present work, the approach is modified so as to determine the point charge centers of AAs from smoothed charge density (CD) distribution functions $\rho_{A,t}$ obtained using the Poisson equation:

$$-\nabla^2 \Phi_{A,t} = \frac{\rho_{A,t}}{\varepsilon_0} \quad (1)$$

where $\rho_{A,t}$ is associated with the electrostatic potential $\Phi_{A,t}$ generated by an amino acid A and smoothed with a degree t.

A single CG model is established for each AA, taking into account several of its most frequent conformations,

and charges are then assigned through a charge fitting procedure versus weighted all-atom MEP values. Finally, a library of 3D atom coordinates of the AA and their corresponding CGs is built and used, through an automated superposition algorithm, to generate the CG models of any arbitrary protein structure.

In this paper, we will show that using the extrema of CD distribution functions rather than those of MEPs [51] allows to overcome the drawbacks mentioned above, particularly when extrema are located in distributions $\rho_{A,t}$ that are calculated from MEPs built on negative and positive charges separately, as later detailed. The new AA models allow to reproduce protein MEP maps, and additionally, provide solvation free energies as obtained using the program APBS [52, 53] that are in close agreement with the all-atom ones. The charge values allowing the calculation of the initial all-atom MEP Φ function are taken from the Amber99 FF [54] to allow comparisons with previous results [51]. Applications were achieved on rigid structures retrieved from the Protein Data Bank (PDB) [55, 56], i.e., four small peptides (PDB access codes 1BC5, 1BXX, 2EVQ, 2RD4), a set of 53 protein structures as listed in [57], and four KcsA channel structures (PDB access codes 1BL8, 1S5H, 2ATK, 2P7T).

A description of the basic theory is reported in the following Section. Then, results are discussed for the small peptides and larger protein systems. Finally, conclusions and perspectives are presented in the last Section.

Theoretical background

In this section, we present the mathematical formalisms that were used to design the CG representation of AAs from their smoothed CD distribution functions and to calculate the corresponding point charges.

Location of CGs from the analysis of a CD distribution function

Molecular electrostatic potential and charge density distribution functions

The electrostatic potential function generated by a molecule A is approximated by a summation over its atomic contributions using the Coulomb equation:

$$\Phi_A(r) = \sum_{a \in A} \frac{q_a}{r} \tag{2}$$

q_a being the net charge of atom a, $r = |\mathbf{r} - \mathbf{R}_a|$, and $\mathbf{R}_a$, the position vector of atom a. A smoothed version of the potential generated by atom a, $\Phi_{a,t}(r)$ can be expressed as [58]:

$$\Phi_{a,t}(r) = \frac{q_a}{r} \mathrm{erf}\left(\frac{r}{2\sqrt{t}}\right) \tag{3}$$

where t is the smoothing parameter and erf stands for the error function. From the potential given in Eq. 3, the corresponding analytical CD distribution function $\rho_{a,t}(r)$ can be obtained from the Poisson equation (1), and expressed as:

$$\rho_{a,t}(r) = \frac{q_a}{(4\pi t)^{3/2}} e^{-r^2/4t} \tag{4}$$

The implementation of the iterative search of the extrema in the molecular CD distribution function, i.e., the CD function derived from the MEP calculated using all atomic charges, led to results that were very similar to those obtained previously [51], i.e., smoothed MEPs with extrema that were located, as illustrated further in the text, away from the molecular skeleton. In order to generate points located on or close to the molecular structure, extrema are now searched for in two separate stages, i.e., from the CD distribution built through Eq. 1 using (1) positive atom net charge only, and (2) negative charges only.

Location of extrema in smoothed charge density distribution functions

To follow the pattern of local extrema in a Gaussian CD distribution function, as a function of the degree of smoothing t, we adapted an algorithm initially described by Leung et al. [59]. The authors proposed a method to model the blurring effect in human vision, which is achieved (1) by filtering a digital image $p(x)$ through a convolution product with a Gaussian function $g(x, t)$:

$$g(x, t) = \frac{1}{t\sqrt{2\pi}} e^{-x^2/2t^2} \tag{5}$$

where t is the scaling parameter, and (2) by assigning each data point of the resulting $p(x,t)$ image to a cluster via a dynamical equation built on the gradient of the convoluted image. We applied this idea to three-dimensional (3D) images such as CD functions $\rho_{A,t}$:

$$r_{\rho_{A,t}} = r_{\rho_{A,t-\Delta t}} + \frac{\Delta}{\rho_{A,t}} \nabla \rho_{A,t} \tag{6}$$

where the ratio $\Delta/\rho_{A,t}$ is the step length, and $\mathbf{r}$ stands for the location vector of a point in a 3D function. The various steps of the resulting merging/clustering algorithm are [50]:

1. At scale $t = 0$, all atoms are considered as the starting points of the merging procedure described below.
2. As t increases from 0.0 to a given maximal value $t_{\max}$, each point moves continuously along a gradient path to

reach a location in the 3D space where $\nabla\rho_{A,t} = 0$. On a practical point of view, this consists in following the trajectory of the extrema on the CD distribution surface calculated at t according to Eq. 4. The trajectory search is stopped when $\nabla\rho_{A,t}$ is lower or equal to a limit value, $grad_{lim}$. Once all extremum locations are found, close points are merged if their interdistance is lower than the initial value of $\Delta^{1/2}$. The procedure is repeated for each selected value of t. If the initial Δ value is too small to allow convergence towards a local extremum within the given number of iterations, its value is doubled (a scaling factor that is arbitrarily selected) and the procedure is repeated until final convergence.

Calculation of CG point charges

CG charge values were obtained using the charge fitting program QFIT [60] as already detailed previously [51]. All-atom MEP grids were built using the Coulomb law, for a system in vacuum, considering Amber99 point charges [54] assigned using the software PDB2PQR [61, 62], with a grid step of 0.5 Å. The quality of the fittings was evaluated by two root mean square deviation (*rmsd*) values, i.e., *rmsdV* determined between the MEP grid values obtained using the fitted charges and the reference unsmoothed all-atom MEP grid values, and *rmsdμ* evaluated between the dipole components calculated from the fitted CG charges and the reference dipole components of the molecular structure. All dipole moments were calculated with the origin of the atom or CG coordinates set to (0. 0. 0.).

Calculation of electrostatic solvation free energies

An additional originality as compared to our previous work [51] is that we also compared electrostatic solvation free energies calculated from the CG models to the corresponding values obtained from the all-atom descriptions of the molecules. To solve the linear version of the PB equation, we used the program "Adaptive Poisson-Boltzmann Solver" (APBS) [52, 53]. In the PB approach, the protein is treated as a low dielectric cavity containing point charges, each characterized by a sphere of a given radius. The cavity is surrounded by the solvent, itself represented by a continuum of high dielectric constant, i.e., 78.54 for water. The solvent volume is defined as a summation over spheres of radius set to 1.4 Å that do not overlap the atoms constituting the solute. All calculations were achieved with grids centered on the solute, of 129 × 129 × 129 and 193 × 193 × 193 points with a mesh of 0.5 Å, for the small peptides and the larger protein models, respectively. A cubic spline discretization is applied to map the point charges of the molecule on their nearest- and next-nearest-neighbor grid points. The dielectric constant of the solvated molecule was set equal to 1 for the small peptides, and to 2 for the larger proteins. The temperature was fixed to 298.15 K. The boundary conditions were such as the potential at the boundary is set to the values prescribed by a Debye-Hückel model for multiple, non-interacting charged spheres.

The electrostatic solvation energy ΔG_{elec} was obtained from the energy E difference between water and vacuum:

$$\Delta G_{elec} = E_{water} - E_{vacuum} \tag{7}$$

where:

$$E = \frac{1}{2}\sum_{i}^{\text{Charges}} q_i\Phi \tag{8}$$

and the electrostatic potential $\Phi(\mathbf{r})$ is given using the linear PB equation:

$$\nabla.(\varepsilon(\mathbf{r})\nabla\Phi(\mathbf{r})) = -4\pi\rho + \kappa^2\Phi(\mathbf{r}) \tag{9}$$

where ρ is the solute CD, i.e., $\rho = \sum_{i}^{\text{Charges}} q_i\delta(\mathbf{r} - \mathbf{r}_i)$, and κ is a parameter that depends on the ionic strength of the solution. In the present paper, a zero bulk ionic strength was considered.

Results and discussion

This section is dedicated to the elaboration of our reduced point charge models of proteins based on the local extrema located in their computed smoothed CD functions. After selection of the smoothing degree to work at, the two first steps of our strategy rely, first, on a CG description of the protein backbone, and then on the development of side chain CG models. Both steps involve the determination of the CG locations and corresponding electrostatic point charges. The final part of the section focusses on the application of our CG models to, as already mentioned in the "Introduction", four small peptides, a set of 53 protein structures [57], and four models of tetrameric ion channels KcsA selected to get a first insight about the ability of our model to differentiate mutants.

To determine the backbone and side chain reduced representations, we adopted the same procedure as detailed in our previous work [51]. For the backbone, we limited our study to a fully extended peptide model made of fifteen AAs, i.e., β-Gly_{15}. As in that first work, end AAs were not charged. The CGs associated with the central Gly residue, Gly8, were chosen as the common motif for all AA backbones. For the side chains, we considered isolated AAs. The consideration of isolated AAs is part of the strategy to minimize the influence of the backbone atoms

on the side chain CGs. The structures of the isolated AAs involved the $(H\alpha C\alpha–C=O)_{AA}(N–H)_{AA+1}$ backbone atoms so as to allow the merging of the two $(C=O)_{AA}$ and $(N–H)_{AA+1}$ moieties as observed in β-Gly_{15}. To generate the 3D structure of all AAs studied in this work, the simulated annealing (SA) procedure implemented in the program SMMP05 [63, 64] was applied to pentadecapeptide models, i.e., Gly_7-AA-Gly_7 structures, with Ω, Φ, Ψ, and χ dihedrals constrained to pre-defined values. Each SA run, carried out with default running parameters and the ECEPP/3 FF [65], consisted in a first 100-step equilibration Monte Carlo (MC) Metropolis stage carried out at 1,000 K. Then the procedure was continued for 50,000 MC Metropolis iterations until the final temperature, 100 K, was reached. The lowest potential energy structure generated during each run was kept. Isolated AA structures were then obtained by pruning the optimized pentadecapeptides.

Selection of the smoothing degree

The search for the local extrema in the MEP-derived molecular CD distribution functions Eq. 4 was carried out using the hierarchical merging/clustering algorithm with the following parameters: $t = 0.05–3.0$ bohr^2, $\Delta_{init} = 10^{-4}$ bohr^2, $grad_{lim} = 10^{-6}$ e$^-$/bohr^2. As already mentioned before, we treated separately positive and negative sets of charges so as to avoid the drawbacks described in the "Introduction".

The merging pattern of the trajectories is as follows (Online Resource 1). First, six atoms of each AA, i.e., N–H, CαHα, and C=O, lead to three extrema at $t = 0.1$ bohr^2 in both the negative and positive MEP functions. The first two of these three extrema are further grouped at $t = 0.35$ and 0.7 bohr^2, in the negative and positive MEPs, respectively. At the final stage, a single extremum resulting from the merge of the trajectories followed by the atoms involved in $(H\alpha_2C\alpha–C=O)_{AA}(N–H)_{AA+1}$ appears at $t = 1.7$ and 1.05 bohr^2, respectively. The number of extrema found in the CD distribution functions of β-Gly_{15} as a function of t remains *quasi* constant starting at $t = 1.70$ bohr^2 when negative and positive charges are treated separately (Online Resource 1). That smoothing degree corresponds to a total number of two extrema per glycine residue (Fig. 1a). As CG model for AA backbones, we thus selected that two-point representation obtained at $t = 1.70$ bohr^2, which keeps the dipolar character of the backbone. With respect to the approach based on a CD built from all atomic charges, the separate treatment of negative and positive charges also provides a difference in the location of the extrema. This is illustrated in Fig. 1 where it is shown that in the all-atom CD (Fig. 1b), the distribution of the extrema is similar to the one obtained in our previous study of smoothed MEPs [51], i.e., CGs are located rather away

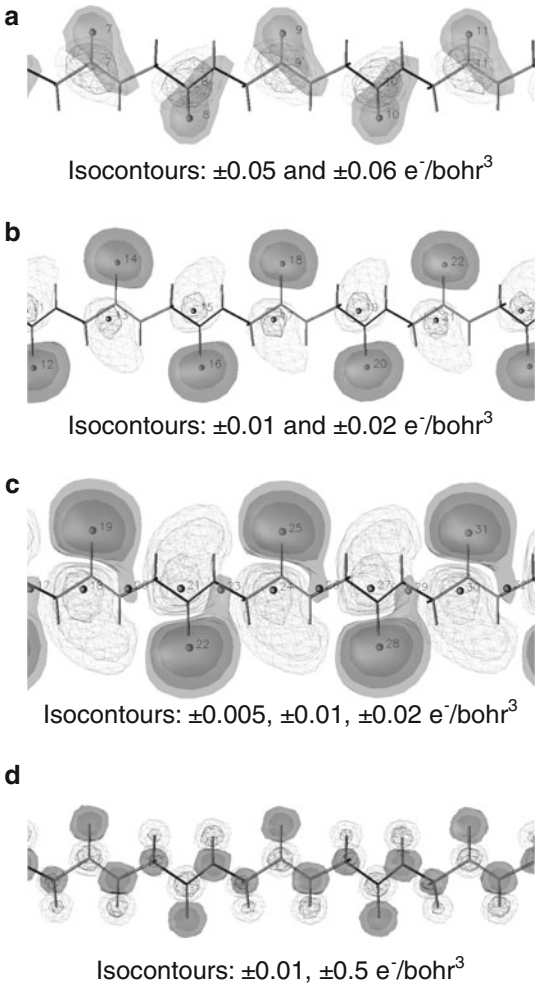

Fig. 1 Local extrema (*spheres*) and isocontours of the smoothed CD distribution function of β-Gly_{15} calculated using the Poisson equation applied to **a** the Amber99 MEPs calculated using the negative and positive atom charges separately and smoothed at $t = 1.7$ bohr^2, **b** the all-atom Amber99 MEP smoothed at $t = 2.0$ bohr^2, **c** the all-atom Amber99 MEP smoothed at $t = 1.7$ bohr^2, **d** the all-atom Amber99 MEP smoothed at $t = 0.05$ bohr^2. Only residues Gly6 to Gly10 are shown. Negative and positive isocontours are displayed using plain surfaces and meshes, respectively

from the molecular skeleton. At $t = 2.0$ bohr^2, points 17 and 18 are indeed located along the C=O bond of Gly8, at distances of 0.796 and 0.558 Å from the C and O atoms, respectively. A separate treatment of the charges (Fig. 1a) allows to get extrema that are closer to the molecular skeleton, one of the aims of the present work. Indeed, both points 9, each obtained from the negative and positive charges, are very close to the C and O atoms of residue Gly8, at distances of 0.122 and 0.161 Å, respectively. At the same value of $t = 1.7$ bohr^2, the all-atom CD leads to three peaks per AA residue, e.g., points 23–25 for Gly8 (Fig. 1c). For comparison, isocontours of the CD obtained at a very low smoothing degree, i.e., $t = 0.05$ bohr^2, illustrate the higher number of CD extrema (Fig. 1d).

Backbone modeling

As in our previous work [51], we considered an extended Gly_{15} sequence characterized by $\Omega = 180°$, $\Phi = -139°$, and $\Psi = 135°$ to generate a regular point charge distribution for the backbone. Then, only the central AA residue Gly8 was kept with backbone atoms $(H\alpha_2C\alpha–C=O)_8$ $(N–H)_9$. The resulting fitted CG charges of that isolated structure of Gly, consisting of two CGs separated by a distance of 1.20 Å, are equal to ± 0.812 e^-, a value that is larger than in the model presented previously [51], ± 0.244 e^-. In that previous model, the two point charges were separated by a distance of 2.52 Å. Current *rmsdV* and *rmsdμ* values are equal to 13.81 kJ/mol and 1.47 D ($\mu_{CG} = -1.81$, 2.86, -3.22; $\mu_{all\text{-}atom} = -3.23$, 2.74, -2.89 D). As mentioned above, that two-CG motif for Gly was further considered to model the backbone of all other AAs and the program QFIT [60] was used to evaluate their charges.

To derive the charge values of the end CGs, located on N and OXT, we also selected the β-Gly_{15} model. Thirty CGs were assigned to the backbone atoms, through the template of Gly, and two end points were added on N and OXT. Keeping the charges of the Gly template on the 30 first points to their value of ± 0.812 e^-, the two end charges were free to vary under the use of the program QFIT [60]. This led to values of ± 0.9288 e^-, with $rmsdV = 12.13$ kJ/mol and $rmsd\mu = 0.45$ D ($\mu_{CG} = -194.71$, -127.43, -10.13; $\mu_{all\text{-}atom} = -194.51$, -127.64, -9.79 D).

Side chain modeling

CG representations of each of the 20 AA side chains were obtained by considering the AAs in specific conformational states, as detailed in our previous paper [51]. Except for AA = Gly and Ala, most recurrent rotamers were generated by taking into account the angular constraints given in Table 2 of [51]. These rotamers were selected according to their occurrence degree in protein structures reported in the *Structural Library of Intrinsic Residue Propensities* (SLIRP) [66, 67]. To model the disulfide bridge, the particular case of Cys3-Cys19 in protein structure 2ERL was arbitrarily selected. As already mentioned, from the β-Gly_7-AA-Gly_7 chain models generated using SMMP05 [63, 64], only the backbone atoms $(H\alpha C\alpha–C=O)_{AA}$ $(N–H)_{AA+1}$ were considered to represent the central AA residue. This was achieved to avoid the generation of side chain CGs that might depend on a particular secondary structure motif or neighborhood. We considered the following protonation states: Arg(+1), Asp(−1), Glu(−1), and Lys(+1).

We next determined the charge values for the CG descriptions of each AA through a fitting procedure carried out using QFIT [60] versus unsmoothed all-atom MEP grids. In this procedure, and for each of the AAs, all rotamer descriptions in terms of extrema observed in the CD distribution functions derived from MEPs smoothed at $t = 1.70$ $bohr^2$ were considered according to their occurrence probability (Table 2 of [51]). This step was carried out in four stages. First, atom charges were assigned to the isolated AA structures using PDB2PQR [61, 62]. Side chain extrema were then located using the hierarchical merging/clustering algorithm applied separately to the negative and positive atom charge sets. Second, the corresponding CG charge values were fitted versus the all-atom MEP generated from the side chain atoms only. Third, the backbone CGs were added in accordance with the motif found for the central AA Gly8 in β-Gly_{15} and, fourth, a second CG charge fitting procedure, now carried out versus the MEP calculated using all the AA atoms, was achieved to determine the charge values of the two backbone CGs while preserving the side chain CG charge values first obtained. As a result, each AA is described by a single CG model retrieved from the extrema analysis of their smoothed CD distribution functions. A conformation independent CG model is obviously of great interest when considering further implementation towards Molecular Mechanics (MM) applications and, particularly, Molecular Dynamics (MD) simulations.

In Fig. 2, we report the so-obtained original or simplified CG representations for the 20 AA residues. Corresponding CG charges are given in Online Resource 2 and deviations of the electrostatic properties versus the all-atom ones are reported in Online Resource 3. As already mentioned in the "Introduction", the extrema of the CD distribution functions are closer to the atoms of the AAs than they were in the MEP functions. It is clearly seen when comparing Fig. 2, 3, 4, 5 and 6 of Ref. [51]. All non-cyclic C–H based residues, i.e., Ala, Ile, Leu, and Val, have no side chain points. This was chosen because of the low charge values obtained initially for their side chain CGs, and as an easy way to model those specific residues in possible further MM/MD-based applications. For Lys, originally constituted of two close CGs, we also simplified the model by setting the positive charge exactly on the Nζ atom (point 3 in Fig. 2). Similarly, for the disulfide bridge, Cyx, points were fixed on the S atoms (points 3 and 6 in Fig. 2). For all other AAs, the original point locations observed in the smoothed MEP functions were kept for the design of the templates and, thus, for the charge fitting procedures. It is however noticed that for Asn, Gln, Phe, and Trp, the CG number and locations were those obtained at different smoothing degrees, i.e., $t = 1.10$, 1.10, 2.60, and 2.35 $bohr^2$, respectively. This allowed to adjust the number of CG points that would otherwise have consisted of 4, 4, 9, and 10 points, instead of 5, 5, 6, and 7 points,

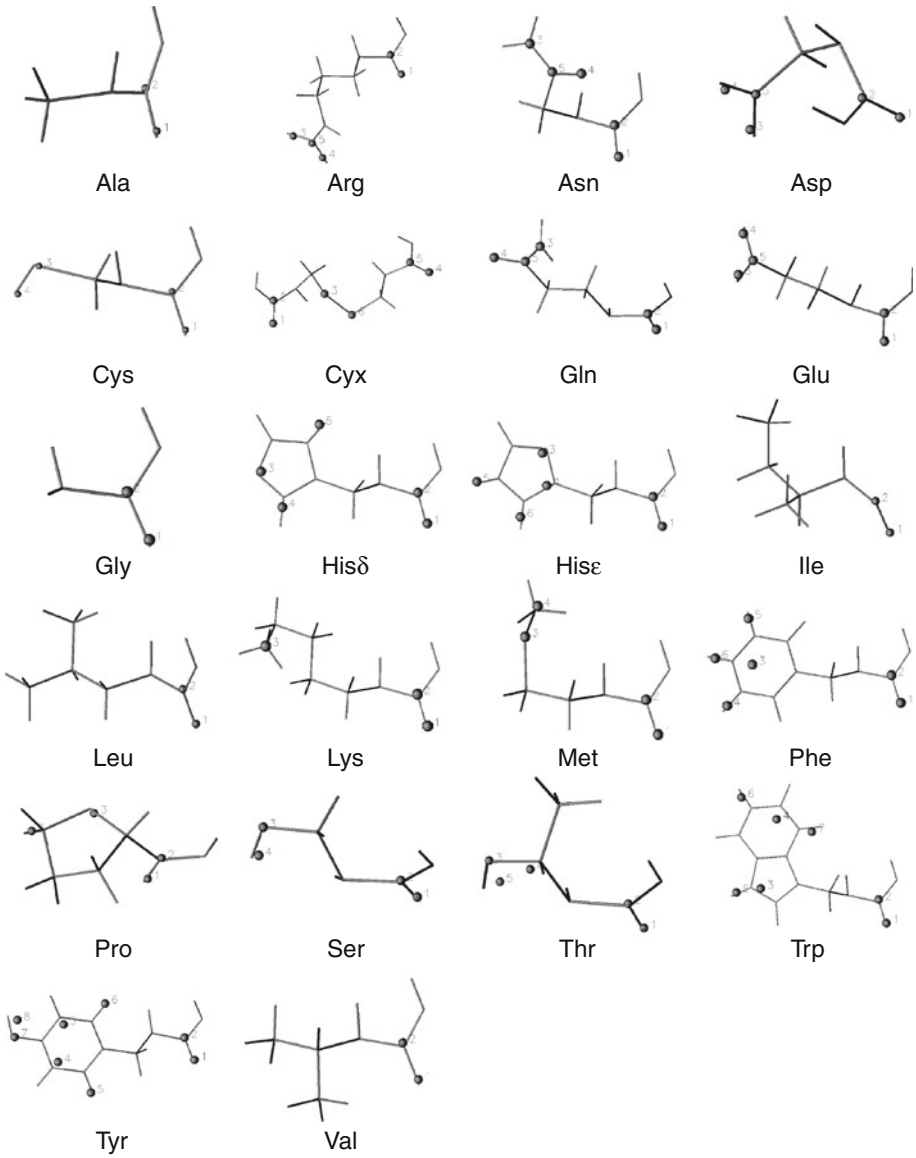

Fig. 2 CG model for each of the 20 AA residues and for the disulfide bridge as established at $t = 1.70$ bohr^2 from the hierarchical merging/clustering algorithm applied to the CD distribution function built from smoothed Amber99 MEPs. CG points are numbered as in Online Resource 2

respectively. In Fig. 2, we also note that for hydroxyl containing residues, i.e., Ser, Thr, and Tyr, there is a negative charge located on the O atom, while the positive charge associated with H is strongly delocalized towards the side chain skeleton (points 4, 5, and 8 for Ser, Thr, and Tyr, respectively). A dipolar description is also obtained for the sulfur containing residues with a CG charge close to the S atom and one in the neighborhood of the H atom (point 4 for Cys) or CH_3 group (point 4 for Met). For the negatively charged residues, i.e., Asp and Glu, each carboxylate group leads to two negative CG charges located near the O (points 3 and 4), and a positive charge on the C of the carboxylate function. Positively charged residues, Arg and Lys, present different behaviors. While the side chain of Lys leads to only one CG positive charge value (point 3), the Arg side chain is characterized by a 2-point motif (points 3 and 4), wherein each charge is located along the C–NH_2 bonds of the guanidinium group. This largely differs from the previous model based on MEP extrema where CGs were located between C–NH_2 bonds [51]. Let us additionally mention that for some AAs, an identity in the charge values was imposed such as, for Arg ($q_3 = q_4$), Asp ($q_3 = q_4$), Cyx ($q_3 = q_6$), Glu ($q_3 = q_4$), Phe ($q_4 = q_5$), and Tyr ($q_3 = q_4$, $q_5 = q_6$), as reported in Online Resource 2. On the whole, all obtained CG models are as good, in terms of *rmsdV* and *rmsdμ* values, as the previous ones [51], with the exceptions of Arg, Lys, Pro, Ser, and Thr, for which the new models are better approximations of the all-atom electrostatic properties (Online Resource 3).

As already mentioned in the "Backbone modeling" section, the degree of detail of the CG models of the AAs is chosen to preserve the dipolar character of their backbone. This eventually led to CG models of about 1 point per 4–5

atoms. This value is close to, for example, the one selected by the developers of the MARTINI FF [26]. In a previous publication [49], a method based on the topological analysis of the promolecular ED led to a model with a lower granulometry level, which appeared to provide MEP results that agree less well with all-atom descriptions.

Automated CG generation procedure

To easily generate CGs for a protein structure, we implemented an automated procedure that is fully based on the application of a superimposition algorithm of CG motif templates of each AA onto the corresponding AA structures of the protein under study. We used the program QUATFIT [68, 69] to, first, superimpose a limited set of atoms from the template on the studied structure, and then applied the resulting transformation matrix to generate the corresponding CG coordinates. For the most flexible AAs, i.e., Ser, Thr, and Tyr, the set of atoms used in the first stage is limited to three bonded atoms so as to obtain a CG model that fits the AA side chain, whatever its conformation is. On the contrary, the generation of a CG model of a more rigid AA side chain, i.e., His, Trp, and Phe, can be based on a larger set of atoms and does not involve any conformational problem.

The templates that were selected in this study are described in Online Resource 4 with all charge values given in Online Resource 2.

Applications to small peptide structures

Four small peptide structures, with electrostatic properties already reported in the literature, were selected (Fig. 3). The 3D structure of the 12-residue β-hairpin HP7 was retrieved from PDB [55, 56] (PDB code 2EVQ) following the work of Basdevant et al. [24] The primary sequence of that peptide is Lys-Thr-Trp-Asn-Pro-Ala-Thr-Gly-Lys-Trp-Thr-Glu, with a global net charge of $+1$ e^-. It is an interesting reference system because a fragment-based description and the corresponding point charges were already available [24]. The 3D structure of two other peptides, the Tgn38 internalization peptide Dyqrln with sequence Asp-Tyr-Gln-Arg-Leu-Asn (PDB code 1BXX) and the C-terminal fragment of the chemotaxis receptor (PDB code 1BC5) with sequence Asn-Trp-Glu-Thr-Phe, were studied following the work of Exner and Mezey [70]. Additionally, we selected a phospholipase inhibitor, with sequence Leu-Val-Phe-Phe-Ala (PDB code 2RD4) involved in the Aβ7 structure studied by Pizzitutti et al. [25].

For each of those peptides, CG models were obtained by applying the automated procedure specified above. End charges were added on the terminal atoms N and OXT. In a first stage, the corresponding charge values were set equal to ± 1 e^-. A visualization of the *rmsd* values obtained between the atom positions of the AA templates and the corresponding ones of the protein crystal structure, for each of the superimpositions achieved using QUATFIT [68, 69] during the CG generation, is presented in Fig. 4. All but one fit lead to a *rmsd* value that is lower than 0.16 Å. The value of 0.28 Å is due to the fit of the Pro side chain model in peptide 2EVQ. These *rmsd* values are actually dependent on the structural discrepencies that might occur between the molecular structures built using the program SMMP05 [63, 64], applied to generate the CD distribution functions of the AAs, and the structures retrieved from the PDB. As seen in Fig. 4, the *rmsd* are rather low. The quality of the CG models, evaluated versus the all-atom one, is presented in Table 1. It is achieved in terms of the *rmsdV* and *rmsd*μ deviation values. One first sees that there is an average ratio "# of atoms/# of CGs", or reduction factor, of 3.5/1. When no charge fitting is applied, the new CG models provide, for each of the four peptides, a better approximation of the all-atom electrostatic properties than the previous model based on the MEP extrema [51]. Let us however notice that, for 2EVQ and 1BXX, the sign of μ_y is kept inversed, like it was previously. While keeping all charges constant but the two end ones, the charge fitting procedure led to improved reduced models, with end charges q_{end} lower than unity. For example, for 2EVQ, the dipole moment value was drastically improved with *rmsd*μ equal to 3.90 D instead of 7.95 D. The model even led to a $rmsdV = 15.61$ kJ/mol, a value that is considerably lower than for the two Basdevant's descriptions, 37.45 and 30.38 kJ/mol. According to Basdevant et al. [24], the first set of charges was established without any charge constraints on the AAs, while the second set was obtained by fixing the individual AA charges to 0 or ± 1 e^-. The protein models, and especially their dipole approximation, can thus be largely improved by fixing the end charges to absolute values of about 0.9 e^-. This is consistent with the value of ± 0.9288 e^- obtained earlier in this work. Thus, from this point on, all end charges will be set equal to ± 0.9288 e^-. In Fig. 3, we report the MEP isocontours calculated under various conditions, i.e., unsmoothed all-atom Coulomb MEP and CG MEP in vacuum. Differences are hardly visible at the level of the MEPs.

To evaluate how our CG models approach the all-atom solvation energies, we applied the APBS 1.2.1 code [52, 53] through two different procedures. First, the electrostatic contribution to the solvation energy was computed for the all-atom representation of the four protein structures, using the atomic charges and radii generated by the program PDB2PQR [61, 62]. Second, the corresponding solvation energy was calculated for the CG models with the molecular surface defined exactly as for the all-atom description. These

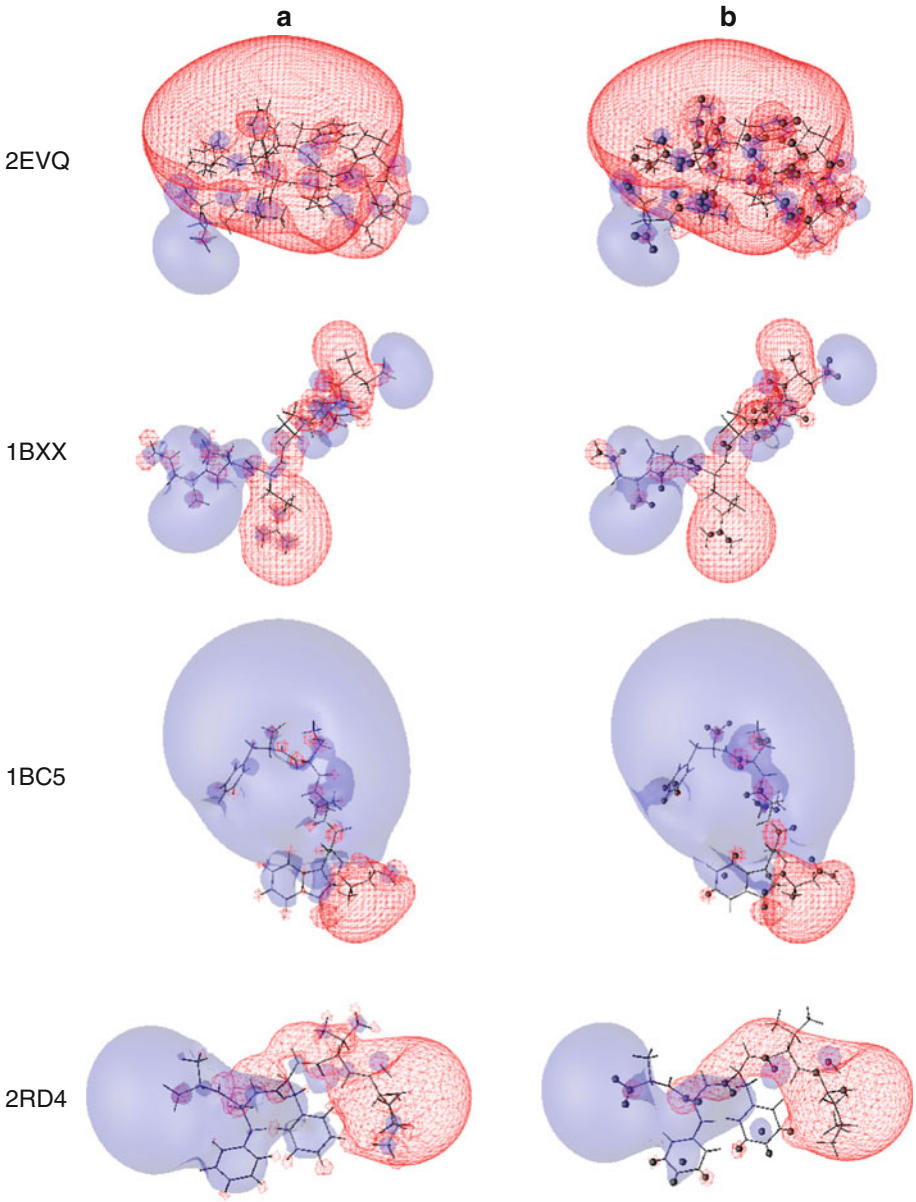

Fig. 3 MEP isocontours (negative and positive isocontours are displayed using plain surfaces and meshes, respectively) for the four small peptides (with PDB code): **a** unsmoothed all-atom Coulomb MEP at ±0.07 e^-/bohr, **b** CG Coulomb MEP with CGs (*black spheres*) at ±0.07 e^-/bohr (1 e^-/bohr is equivalent to 1.06 10^3 kT/ε at $T = 298.15$ K and $\varepsilon = 1$)

two solvation energy results were named AA_{aa} and CG_{aa}, respectively (Table 2). The error on ΔG_{elec}, given between parentheses in Table 2, is largely lower (in absolute value) for the current model CG_{aa} than it was for the previous model [51], named p-CG_{aa}, especially for 2EVQ and 1BXX. For the two other peptides, 1BC5 and 1RD4, the relative error of CG_{aa} models is close to the p-CG_{aa} values.

Applications to a set of 53 protein structures

Fifty-three protein structures were selected from the work of Tjong and Zhou [57]. All His residues were considered in their Hisδ protonation state. Incomplete AA residues were deleted, i.e., Gly150 and Glu117 in structures 1OD3 and 1UNQ, respectively, and Amber99 charge values were assigned to each atom through the program PDB2PQR [61, 62]. The building procedure of the CG models was checked by controling the *rmsd* values obtained between the 3D atomic coordinates of the AA templates and the corresponding coordinates in the protein crystal structures (Online Resource 5a). It is noted, similarly to the four small peptide cases, that deviations are generally lower than 0.05 Å, and that the highest *rmsd* values are associated with several Pro residues. The worst case, characterized by *rmsd* = 0.352 Å, is observed for the side chain of Pro28 in structure 2A6Z (Online Resource 5b). Therefore a *rmsd* value above 0.1 Å does not reflect a severe deviation versus the original all-atom structures.

To validate the end CG charge value determined previously, i.e., $q_{end} = \pm 0.9288$ e^-, deviation values *rmsdV*

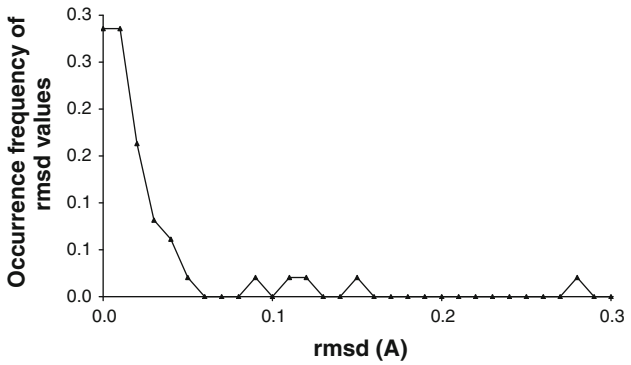

Fig. 4 Occurrence frequency of the root mean square deviation (*rmsd*) values calculated between the atomic positions of the AA template motifs and the corresponding ones of the actual AA backbones or side chains, calculated over all superimpositions achieved for the generation of the CGs of the four small peptides

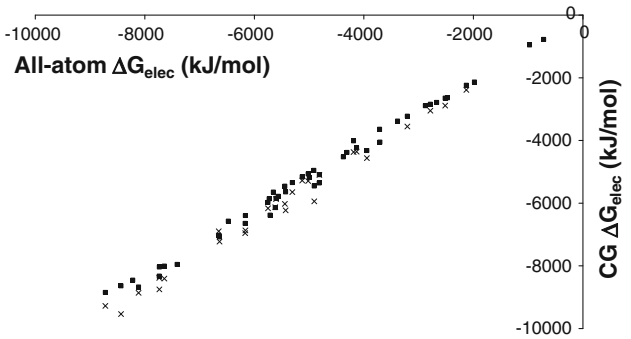

Fig. 5 Electrostatic solvation free energy ΔG_{elec} calculated using various models as a function of the corresponding all-atom values (Online Resource 7), for the 53 protein structures: CG_{aa} (plain squares), p-CG_{aa} (crosses). Linear correlations are: 1.01*x*–117.36 with $R^2 = 0.998$, 1.02*x*–367.99 with $R^2 = 0.997$, respectively

and *rmsdμ* were calculated for CG models characterized by various end charges (Table 3 and Online Resource 6). We fixed the end charges to be equal to 1., 0.9288, 0.9, and 0.8 e^-. We further evaluated the quality of the four representations by two descriptors: (1) the number of best models, i.e., whose *rmsdμ* deviation values are the lowest ones (bold numbers in Table 3 and Online Resource 6) (2) the number of models with *rmsdμ* within 1 D of the *rmsdμ* of the best model obtained (italic numbers in the two tables). When $q_{end} = \pm 1$ e^-, the model is best for 14 of the 53 structures, and is close to the best models for 6 structures. When $q_{end} = \pm 0.9288$ e^-, the corresponding numbers are 12 and 12; when $q_{end} = \pm 0.9$ e^-, one obtains 9 and 16 structures, respectively, and when $q_{end} = \pm 0.8$ e^-, one gets 18 and 1 structures. While using end charge values of 0.8 e^- leads to the highest number of best models, the use of the pre-determined value of 0.9288 e^- allows to obtain a total of 24 (12 and 12) good models, a good compromise. The mean *rmsdV* and *rmsdμ* values presented in Table 3 and Online Resource 6 also show that models with

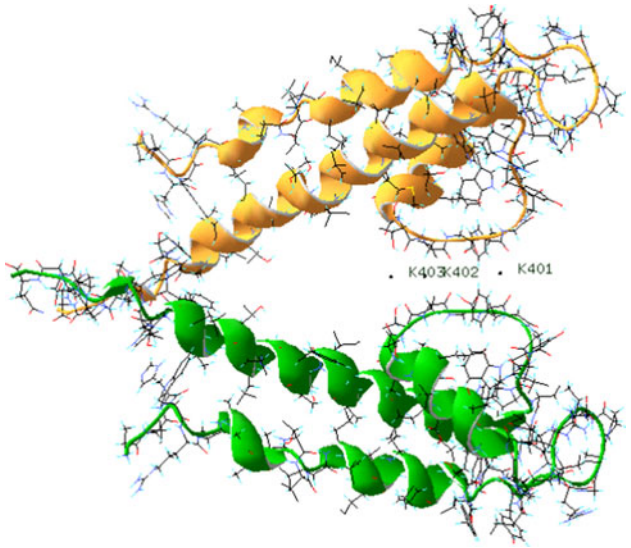

Fig. 6 3D conformation and secondary structure of the potassium channel KcsA (PDB access code 1BL8). Two monomers only, chains A and C, are displayed. Ions K401 and K403 are separated by a distance of 10.62 Å

$q_{end} = \pm 0.9288$ e^- and $q_{end} = \pm 0.9$ e^- are very close in their ability to approach all-atom electrostatic properties. Additional comparisons are proposed versus the previously developed model [51] for the 29 protein structures that do not involve any disulfide bridges. A end CG charge value of 0.9288 e^- was considered, as determined earlier. Over those 29 structures, four cases, i.e., 1IQZ, 1UCS, 1UNQ, and 1ZZK, are in better or close agreement with the all-atom representations (underlined numbers in Online Resource 6). However, the agreement is not drastically better than with the new model established herein.

To evaluate how the CG models are efficient in approaching the all-atom solvation energies, we again applied the program APBS 1.2.1 [52, 53] through the procedure already described for the four small peptides. All values are reported in Online Resource 7. Mean error values are given in Table 2. It is observed that the relative error generated by the two CG models are rather low, i.e., mean errors of 3.86% for CG_{aa} and 8.80% for p-CG_{aa}.

The plot of the electrostatic solvation free energy ΔG_{elec} of the various models versus the all-atom corresponding values is shown in Fig. 5. The best linear correlation is observed between the CG_{aa} model and the all-atom values, with $R^2 = 0.998$. The model p-CG_{aa} also leads to a good linear correlation with $R^2 = 0.997$.

Applications to KcsA channel structures: mutation effect

The large protein system selected to test our CG models was the KcsA potassium channel (PDB access code 1BL8)

Table 1 Electrostatic properties of the CG model of the four small peptides (with PDB code) versus their corresponding all-atom version. *rmsdV* and *rmsdμ* are given in kJ/mol and D, respectively

	No charge fitting	Fit of end charges only	Basdevant's model #1	Basdevant's model #2
2EVQ	197 atoms			
q	1.0			
μ (all-atom)[a]	4.78, −2.21, −66.43			
# of CGs	55	55	28	28
Reduct. factor	3.6	3.6	7.0	7.0
rmsdV	19.75	15.61	37.45	30.38
rmsdμ	7.95	3.90	4.04	3.96
μ[a]	1.30, 0.06, −73.21	2.59, −3.38, −69.45	1.50, −4.14, −65.05	4.74, 0.70, −69.11
q_{end}	±1.0000	±0.8525		
1BXX	110 atoms			
q	0.0			
μ (all-atom)[a]	11.34, −0.96, −15.73			
# of CGs	32	32		
Reduct. factor	3.4	3.4		
rmsdV	22.22	19.00		
rmsdμ	6.80	1.30		
μ[a]	16.86, 2.73, −14.26	12.02, −1.80, −15.02		
q_{end}	±1.0000	±0.9183		
1BC5	90 atoms			
q	−1.0			
μ (all-atom)[a]	−310.92, −287.53, 7.06			
# of CGs	30	30		
Reduct. factor	3.0	3.0		
rmsdV	17.03	12.09		
rmsdμ	5.61	2.29		
μ[a]	−310.85, −293.01, 5.86	−312.80, −287.36, 5.78		
q_{end}	±1.0000	±0.9173		
2RD4	88 atoms			
q	0.0			
μ (all-atom)[a]	35.12, 22.63, −44.04			
# of CGs	20	20		
Reduct. factor	4.4	4.4		
rmsdV	12.84	10.88		
rmsdμ	11.67	11.55		
μ[a]	36.74, 21.52, −46.03	35.12, 20.64, −44.22		
q_{end}	±1.0000	±0.9630		

Electric charges are given in e^-

[a] x, y, and z components of μ

and three of its mutants (PDB access codes 1S5H, 2ATK, and 2P7T). KcsA is a transmembrane protein structure that is commonly used to model biological ion channels and to evaluate computational approaches in the study of protein electrostatics [71–74]. It is formed by four identical chains, each chain containing two α-helices connected by a loop located in the channel region (Fig. 6). The channel consists of the so-called selectivity filter, that is about 18 Å long, pointing to the extracellular region, a larger cavity of about 10 Å, and a 15 Å long narrow gating pore opened towards the intracellular region. The gating pore and the cavity are hydrophobic regions, while the selectivity filter, mainly formed by five residues, Thr74-Thr75-Val76-Gly77-Tyr78, is covered by in-line carbonyl O atoms of the protein backbone, which build a structure that is similar to a water solvation shell around a K^+ ion.

In the present work, the 3D models of the protein systems were prepared according to the X-ray crystal structure

Table 2 Electrostatic solvation free energy ΔG_{elec} (kJ/mol) calculated using APBS 1.2.1 for the various protein descriptions (see definitions in the text) of the four small peptides, five of the 53 protein structures selected from Tjong and Zhou [57], and the four KcsA channels (with PDB code)

	Description				
	AA_{aa}	CG_{aa}	p-CG_{aa}	$Bas1_{aa}$	$Bas2_{aa}$
Small peptides					
2EVQ	−1,004.18	−1,101.26 (9.7)	−1,183.72 (17.9)	−1,098.40 (9.4)	−1,433.76 (42.8)
1BXX	−883.66	−869.30 (−1.6)	−1,278.54 (44.7)		
1BC5	−1,100.37	−1,045.59 (−5.0)	−1,142.32 (3.8)		
2RD4	−586.16	−548.34 (−6.5)	−623.18 (6.3)		
Protein set [57]					
Mean error versus AA_{aa} (%)		3.86 ± 3.59	8.80 ± 4.69		
Ion channels					
1BL8	−11,211.26	−11,764.79 (4.9)	−13,380.26 (19.3)		
1S5H	−10,123.54	−10,975.92 (8.4)	−12,087.67 (19.4)		
2ATK	−11,280.49	−12,005.50 (6.4)	−12,997.71 (15.2)		
2P7T	−10,619.83	−11,219.73 (5.6)	−12,321.56 (16.0)		

The two Basdevant's sets of charges (Bas1 and Bas2) [24] are also used for structure 2EVQ. The relative error versus the all-atom value is given in parentheses (in %). Mean values are calculated over the initial set of 53 protein structures. Subscript aa stands for "all-atom molecular surface"

Table 3 *rmsdV* and *rmsdμ* values, given in kJ/mol and D, respectively, obtained from the CG model used with various end CG charges for five of the 53 proteins structures (with PDB code) selected from Tjong and Zhou [57]

		1G66	1HJE	1PQ7	1VB0	1XMK	Means
Charge		−2	+1	+4	+3	+1	
# of atoms		2,794	175	3,065	913	1,268	
$\mu_{\text{all-atom}}$		202.77	77.88	66.95	686.22	273.30	
# of CGs		743	48	785	248	299	
$q_{end} = \pm 1.0000$	*rmsdV*	26.74	21.00	26.15	20.71	33.14	26.12 ± 3.78
	rmsdμ	29.65	*5.88*	20.05	**5.71**	48.15	21.38 ± 12.35
$q_{end} = \pm 0.9288$	*rmsdV*	25.77	20.84	25.90	24.73	29.83	25.21 ± 3.69
	rmsdμ	*27.53*	**5.45**	*15.31*	7.67	35.48	18.79 ± 11.30
$q_{end} = \pm 0.9000$	*rmsdV*	25.61	21.34	25.94	25.36	28.83	25.03 ± 3.75
	rmsdμ	**27.16**	5.62	**14.95**	9.62	30.41	18.19 ± 11.12
$q_{end} = \pm 0.8000$	*rmsdV*	25.98	25.36	27.03	28.74	26.99	26.13 ± 5.54
	rmsdμ	28.26	7.49	20.97	17.29	**13.58**	20.37 ± 12.55
Previous model [51]	*rmsdV*					31.46	26.09 ± 4.92
with $q_{end} = \pm 0.9288$	*rmsdμ*					42.70	22.76 ± 13.87

Dipole moments and electric charges q are given in D and e^-, respectively. For each protein structure, the lowest *rmsdμ* value is in bold; the other values within 1 D of the best one are in italic. A full version is available (Online Resource 6). Means are reported for the set of 53 proteins

as follows. The design of the histidine residues into a Hisδ configuration was achieved with the program PDB2PQR [61, 62]. For structure 1S5H, a T75C and C90L mutant of 1BL8, residues 22 and 120–124 were deleted to generate a sequence of the same length as all other protein systems, i.e., sequence 23–119. Structures 2ATK and 2P7T are E71A and E71S mutants of 1BL8, respectively. The total charge of the two last structures is thus equal to +2 e^-/chain rather than +1 e^-/chain. Pair superimpositions of the Cα atoms led to low values of *rmsd*, below 0.8 Å (Online Resource 8a). Atomic charges were assigned using PDB2PQR. The K^+ ions were not considered in the calculations.

From original structures of about 5,900 atoms, the application of our automated procedure, completed by the addition of end CG charges on the N and OXT atoms of the end residues of each of the four monomers, led to the generation of about 1,320 CGs (Online Resource 8b). The obtained

reduction ratios are slightly smaller for the current model, about 4.4–4.5/1, than for the previous model [51], i.e., 4.5–4.6/1. It stays, for example, close to the 4/1 value reported by Bond et al. [75, 76] who studied the interaction of membrane proteins with lipid molecules through MD simulations using the MARTINI FF. The visualization of the *rmsd* values obtained between the 3D positions of the AA template atoms and the corresponding ones of the protein crystal structure, for each of the superimpositions achieved using QUATFIT [68, 69] during the CG generation, is very similar to the values obtained for the set of 53 protein structures (Online Resource 8c). For the previous and the current CG models, the large *rmsd* values, i.e., beyond 0.1 Å, correspond to a less efficient fit of the four end residues Gln119 due to the terminal OXT atoms. The lowest *rmsd* values, around 0.01 Å, characterize the superimpositions of the backbone templates, while all larger *rmsd* values, from 0.03 to 0.06 Å, characterize the superimpositions of the side chain templates. For the new CG model, the largest values of *rmsd*, i.e., about 0.34 Å, are due to proline side chain superpositions, as explained before.

The resulting full KcsA CG models are characterized by dipole moments and total charges that are reported in Online Resource 8b, for both the current and previous CG models with $q_{end} = \pm 0.9288$ e$^-$. The current CG model provides better approximations of the all-atom dipole moments, with *rmsd*μ between 32.04 and 40.86 D versus between 38.67 and 64.63 D with the previous model [51]. *rmsdV* values are less affected by the type of CG representation, with values around 25 kJ/mol.

Regarding the electrostatic solvation free energies ΔG_{elec} (Table 2 and Online Resource 8d), one again observes the better approximation provided by the current CG_{aa} model. Indeed, the relative error on ΔG_{elec} varies between 4.9 and 8.4% only, to be compared to values between 15 and 20% for p-CG_{aa}.

MEP profiles were established along the channel central axis, defined by the Cartesian coordinates of the K^+ ions (Fig. 7). In each case, the closest ion to residue 75 was selected to be the origin of the axis, i.e., ion 403, 403, 127, and 505 in structures 1BL8, 1S5H, 2ATK, and 2P7T, respectively. For each KcsA channel, the current CG model leads to a profile that is slightly closer to the corresponding all-atom profile as compared with our previous version [51]. For all structures, one also notices that the pattern adopted by the all-atom profiles is obeyed by the CG profiles. Finally, the order of magnitude of the MEP values is respected by the CG approximations. For structure 1BL8 (Fig. 7), the axis region of the selective filter region is characterized by two MEP minima, followed by a large energy barrier which covers the hydrophobic cavity and narrow pore regions. This had already been observed by Bliznyuk et al. [71].

To identify the effect of a single mutation only, the four Tyr75 residues of 1BL8 were mutated in four Cys75 whose conformation was taken from structure 1S5H. No other conformational changes were considered, and the resulting protein structure was named 1BL8_C75. This led to the profile modifications illustrated in Fig. 8. The single T75C mutation leads to a deepening of the main minimum, as well as to a reduction of the electrostatic barrier separating the two minima. In the all-atom version, the first minimum is replaced by a plateau at about −930.56 kJ/mol. All these changes are also detected by the CG model, with the following characteristics. The distance between the two minima in 1BL8 is equal to 7.22 Å, while the distance between the plateau and the second minimum is equal to 5.94 Å for 1BL8_C75. These two values are close to the all-atom values of 7.43 and 5.73 Å. The two energy barriers, $\Delta 1$ and $\Delta 2$ with $\Delta 2$ being the barrier separating the deepest minimum from the maximum, are also consistent between the all-atom and CG models (Table 4). The 1BL8_L90 mutant, built by taking the Ala71 residue from structure 1S5H and transferring it to the 1BL8 structure, did not show any clear profile differences with 1BL8, while both 1BL8_A71 and 1BL8_S71 mutants were characterized by similar profiles however shifted by an amount of about +444 kJ/mol (Fig. 8). From the study of these artificially built mutants, one concludes that a mutation at the level of residue 75 involves a modification of the MEP pattern inside the channel. There is no such changes when a farther residue is involved, but the change of the electric charge, like in 1BL8_A71 and 1BL8_S71 structures, implies a shift in the energy values and an increase of the $\Delta 1/\Delta 2$ ratio (Table 4). While the absolute values of such ratios are different for the all-atom and CG description levels, e.g., 65.23 and 143.35 kJ/mol for the all-atom and CG models of 1BL8, respectively, the trends followed by $\Delta 1/\Delta 2$ is similar. Indeed, the values of $\Delta 1/\Delta 2$ are (0.12, –, 0.12, 0.29, and 0.25 kJ/mol) and (0.26, 0.07, 0.26, 0.45, and 0.41 kJ/mol) for structures 1BL8, 1BL8_C75, 1BL8_C90, 1BL8_A71, and 1BL8_S71, respectively. Finally, a modification at the outer part of the channel, at the level of residue 90, does not have any significant effect on the MEP profile.

For the other structures, 1S5H, 2ATK, and 2P7T, even if the backbone is preserved (Online Resource 8a), mutations and side chain orientations involve modifications in the MEP profile inside the channel. Nevertheless, one notices close similarities between the all-atom and CG profiles (Fig. 7). In that sense, the models presented in this paper led to better approximations than those obtained in a previous approach [77] wherein the AA CG models were generated using β-pentadecapeptide structures rather than isolated ones. Decoupling backbone and side chain contributions in the elaboration of a CG model is thus clearly

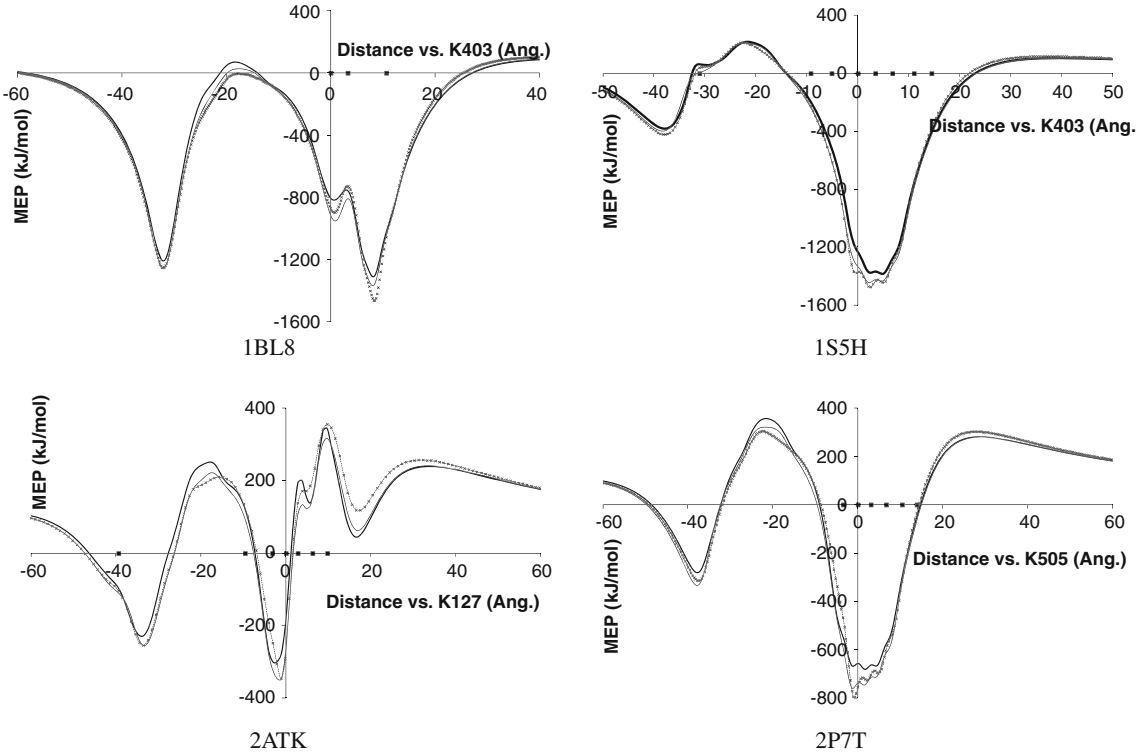

Fig. 7 MEP profiles along the central axis of the four KcsA channels (with PDB code) calculated using the all-atom Amber99 set of charges (*thick plain line*), the Amber99-based CG model with $q_{end} = \pm 0.9288$ e$^-$ (*thin plain lines*), and the previous Amber99-based CG model with $q_{end} = \pm 0.9288$ e$^-$ (*crosses*). The location of the K^+ ions is shown using plain squares

interesting for reproducing all-atom MEP properties, as previously achieved in [37].

All ΔG_{elec} results presented in this paper suggest that the shape of the original all-atom molecular surface should be preserved. Indeed, trial calculations made with a molecular surface defined using CG radius values were not satisfying. One may thus deduce that the steric centers of a molecule differ from the electric charge centers obtained in this work, as also mentioned in [78]. When considering possible applications of the present reduced point charge models in MM and/or MD calculations, two points of view thus appear. The first one consists in replacing the original all-atom electrostatic Coulomb contributions by the reduced model contributions so as to decrease the calculation times. This is the approach to favor if one wishes to preserve the molecular surface corrugation at a high level of detail. A second point of view consists in, additionally, setting the steric contributions of the FF by defining CGs on centers that will differ from the point charge locations, for example, centers that might be obtained from a topological analysis of the full ED distribution function [45–50]. A direct implementation, that does not require any modification of an existing MD program, is simply to consider the CGs as additional particles held to the AA structure through harmonic bonds. First tests carried out in that direction are encouraging but obviously lead to a modification of the total mass of the system. We are considering a second alternative that consists in updating, through geometric and/or fitting rules, the CG description of the system at each step of the MD, thus allowing the calculation of electrostatic energy terms without any other modifications brought to the peptide energy function. The two hereabove described schemes present the advantage to avoid any CG to all-atom reconstruction procedure.

Conclusions

In this work, we applied a hierarchical merging/clustering algorithm to charge density (CD) distribution functions to generate reduced point charge representations of proteins. The CD functions are calculated using the Poisson equation applied to smoothed molecular electrostatic potential functions (MEPs). Through the use of such a procedure, the reduction of a molecular structure representation, particularly a protein structure, was achieved by following the trajectories of its constituting atoms in its progressively smoothed CD distribution function. A protein structure can thus be described by a limited set of points, which correspond to the local extrema of the CD. The aim of such

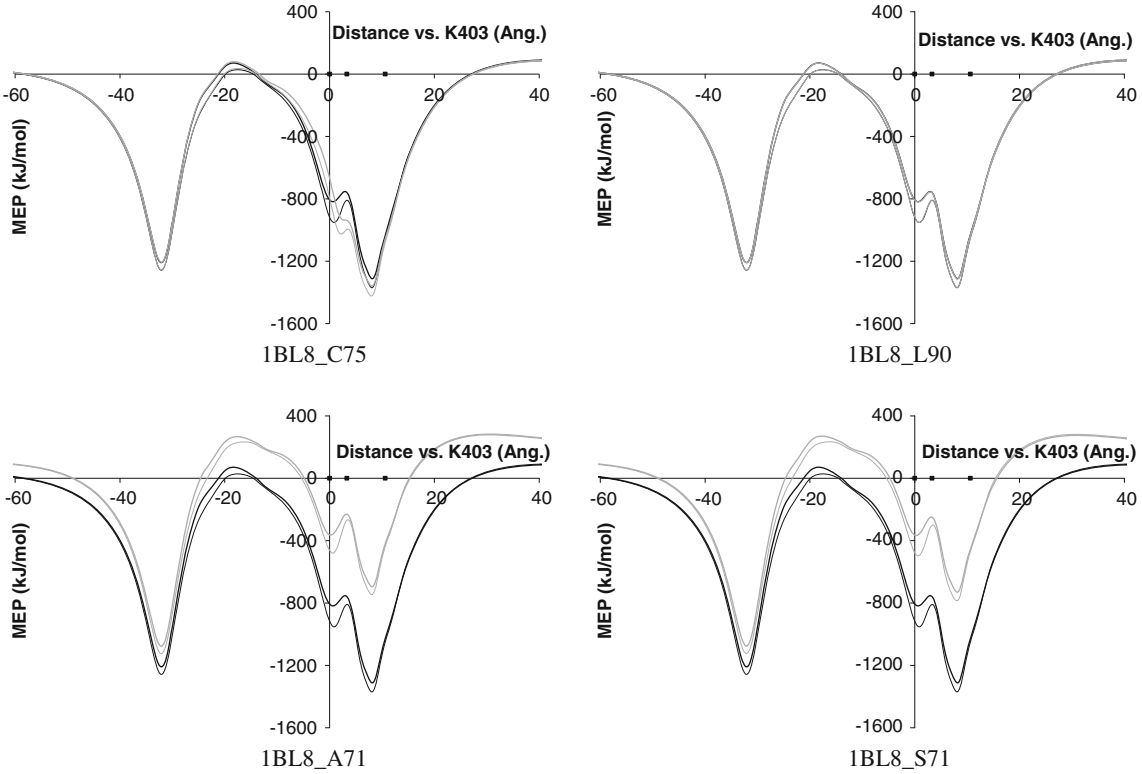

Fig. 8 MEP profiles along the central axis of the KcsA channel 1BL8 (*black*) and of the four artificial mutants (*gray*) calculated using the all-atom Amber99 set of charges (*thick lines*), and the Amber99-based CG model with $q_{end} = \pm 0.9288$ e$^-$ (*thin lines*). The location of the K^+ ions is shown using plain squares

Table 4 Distances *d* between potential wells (in Å) and MEP barriers Δ_i (in kJ/mol) observed along the KcsA channel axis for protein structures 1BL8 and 1BL8_C75

	All-atom	CG (present work)
1BL8		
d	7.43	7.22
Δ1	65.23	143.35
Δ2	558.57	560.16
1BL8_C75		
d	5.73	5.94
Δ1		31.40
Δ2	428.69	427.77
1BL8_L90		
d	7.43	7.22
Δ1	65.12	143.74
Δ2	558.68	559.78
1BL8_A71		
d	7.85	7.43
Δ1	133.84	213.20
Δ2	465.51	478.94
1BL8_S71		
d	7.85	7.43
Δ1	118.60	197.35
Δ2	481.92	485.17

calculations further consisted in the evaluation of electrostatic properties such as MEPs, dipole moments, and solvation energy of a protein using coarse-grain (CG) descriptions, and their comparison with all-atom values.

The present work especially focused on the use of the set of charges Amber99, but is readily applicable to other charge sets that are available in the literature [51, 77]. Reduced descriptions were obtained for each of the 20 natural amino acid (AA) residues and the disulfide bridge with the following specific protonation states: Arg(+1), Asp(−1), Glu(−1), and Lys(+1). Each of the 20 AAs was modeled through various rotamers (except for Ala, Asp, Cyx, Gly, and Pro). The first stage was to apply the merging/clustering algorithm to determine the CG locations of the AA backbone and side chain, separately. In a second stage, charges were assigned to these AA CG representations through a charge fitting algorithm, and were further tabulated as reference values to be used for CG modeling of protein structures of any size. CD-based CG descriptions were shown to lead to points located close to the molecular skeleton, and to be less sensitive to the molecular conformation than MEP-based CG descriptions.

An automated procedure was implemented and tested on four small peptides (PDB access codes 2EVQ, 1BXX, 1BC5, and 2RD4), a set of 53 protein structures taken from

the work of Tjong and Zhou [57], and on a larger system KcsA, a tetrameric potassium ion channel made of four 97-residue long monomers (PDB access code 1BL8), and three of its mutants (PDB access codes 1S5H, 2ATK, and 2P7T). Those artificially built mutants were studied to assess the effect of a single mutation. The generation of the CG representation of each residue was achieved through a superimposition algorithm of CG template motifs on the 3D PDB structure. On the whole, one observes a reduction factor of 4.4/1. Knowing that the calculation of Coulomb interactions involving N point charges asks for a computing time that evolves with $N^2/2$, it means that the gain in calculation time is about a factor of 16.

The new CG model allows to better reproduce all-atom electrostatic properties than the model developed previously [51], i.e., built on extrema of MEPs rather than extrema of CD distribution functions. The improvements are minor for the MEP, larger for the dipole moment values, and drastic for the solvation free energy ΔG_{elec}. When calculating solvation energies, the molecular surface was modeled using the atom radii of the molecular structure. Tests were achieved with a CG-based representation but did not provide satisfying results, although allowing a reduction of the calculation time.

During the elaboration of the MEP-based CG models, three points were considered to be important to favor transferability. First, AAs were studied in the isolated state to neglect the protein backbone conformation, and second, for each AA, CG charges were obtained by considering various side chain conformations. Additionally, CD extrema are less sensitive to a change in conformation than MEP extrema.

On-going studies will focus on porin and especially aquaporin systems, more suitable for solvation studies. Also, implementation of the point charge reduced model obtained in the present work in a MM/MD program is under consideration.

Acknowledgments The authors thank the referees for very useful comments. They also acknowledge Profs. E. Clementi and M. Sansom for very fruitful discussions, as well as Prof. N. Baker for APBS assistance. The ''Fonds National de la Recherche Scientifique'' (FNRS-FRFC), the ''Loterie Nationale'' (convention no. 2.4578.02), and the ''Facultés Universitaires Notre-Dame de la Paix'' (FUNDP), are gratefully acknowledged for the use of the Interuniversity Scientific Computing Facility (ISCF) Center.

References

1. Kahraman A, Morris RJ, Laskowski RA, Favia AD, Thornton JM (2010) On the diversity of physicochemical environments experienced by identical ligands in binding pockets of unrelated proteins. Proteins 78:1120–1136
2. Dell'Orco D, Xue W-F, Thulin E, Linse S (2005) Electrostatic contributions to the kinetics and thermodynamics of protein assembly. Biophys J 88:1991–2002
3. Kumar S, Wolfson HJ, Nussinov R (2001) Protein flexibility and electrostatic interactions. IBM J Res Dev 45:499–512
4. Strickler SS, Gribenko AV, Gribenko AV, Keiffer TR, Tomlinson J, Reihle T, Loladze VV, Makhatadze GI (2006) Protein stability and surface electrostatics: a charged relationship. Biochemistry 45:2761–2766
5. Boiteux C, Kraszewski S, Ramseyer C, Girardet C (2007) Ion conductance vs. pore gating and selectivity in KcsA channel: modeling achievements and perspectives. J Mol Model 13: 699–713
6. Azia A, Levy Y (2009) Nonnative electrostatic interactions can modulate protein folding: molecular dynamics with a grain of salt. J Mol Biol 393:527–542
7. Reif MM, Kräutler V, Kastenholz MA, Daura X, Hünenberger PH (2009) Molecular dynamics simulations of a reversibly folding β-heptapeptide in methanol: influence of the treatment of long-range electrostatic interactions. J Phys Chem B 113: 3112–3128
8. Camacho CJ, Ma H, Champ PC (2006) Scoring a diverse set of high-quality docked conformations: a metascore based on electrostatic and desolvation interactions. Proteins 63:868–877
9. Corrêa F, Salinas RK, Bonvin AMJJ, Farah CS (2008) Deciphering the role of the electrostatic interactions in the α-tropomyosin head-to-tail complex. Proteins 73:902–917
10. Garden DP, Zhorov BS (2010) Docking flexible ligands in proteins with a solvent exposure- and distance-dependent dielectric function. J Comput Aided Mol Des 24:91–105
11. Vizcarra CL, Mayo SL (2005) Electrostatics in computational protein design. Curr Opin Chem Biol 9:622–626
12. Boas FE, Harbury PB (2007) Potential energy functions for protein design. Curr Opin Struct Biol 17:199–204
13. Lippow SM, Tidor B (2007) Progress in computational protein design. Curr Opin Biotech 18:1–7
14. Liang S, Li L, Hsu W-L, Pilcher MN, Uversky V, Zhou Y, Dunker AK, Meroueh SO (2009) Exploring the molecular design of protein interaction sites with molecular dynamics simulations and free energy calculations. Biochemistry 48:399–414
15. Hildebrandt A, Blossey R, Rjasanow S, Kohlbacher O, Lenhof H-P (2007) Electrostatic potentials of proteins in water: a structured continuum approach. Bioinformatics 23:e99–e103
16. Dong F, Olsen B, Baker NA (2008) Computational methods for biomolecular electrostatics. Methods Cell Biol 84:843–870
17. Papazyan A, Warshel A (1997) Continuum and dipolar lattice models of solvation. J Phys Chem B 101:11254–11264
18. Koehl P, Delarue M (2010) AQUASOL: an efficient solver for the dipolar Poisson-Boltzmann-Langevin equation. J Chem Phys 132:064101–1/064101–16
19. Voth GA (ed) (2009) Coarse-graining of condensed phase and biomolecular systems. CRC Press, Boca Raton
20. Hills RD Jr, Lu L, Voth GA (2010) Multiscale coarse-graining of the protein energy landscape. PLoS Comput Biol 6:e1000827/1–e1000827/12
21. Bereau T, Deserno M (2009) Generic coarse-grained model for protein folding and aggregation. J Chem Phys 130:235106/1–235106/15
22. Moritsugu K, Smith JC (2009) REACH: a program for coarse-grained biomolecular simulation. Comput Phys Commun 180:1188–1195
23. Marrink SJ, Risselada HJ, Yefimov S, Tieleman DP, de Vries AH (2007) The MARTINI forcefield: coarse-grained model for biomolecular simulations. J Phys Chem B 111:7812–7824

24. Basdevant N, Borgis D, Ha-Duong T (2007) Coarse-grained protein-protein potential derived from an all-atom force field. J Phys Chem B 111:9390–9399
25. Pizzitutti F, Marchi M, Borgis D (2007) Coarse-graining the accessible surface and the electrostatics of proteins for protein-protein interactions. J Chem Theory Comput 3:1867–1876
26. Monticelli L, Kandasamy SK, Periole X, Larson RG, Tieleman DP, Marrink SJ (2008) The MARTINI coarse-grained forcefield: extension to proteins. J Chem Theory Comput 4:819–834
27. DeVane R, Shinoda W, Moore PB, Klein ML (2009) Transferable coarse grain nonbonded interaction model for amino acids. J Chem Theory Comput 5:2115–2124
28. Liwo A, Czaplewski C, Oldziej S, Rojas AV, Kazmierkiewicz R, Makowski M, Murarka RK, Sheraga HA (2009) In: Voth GA (ed) Coarse-graining of condensed phase and biomolecular systems. CRC Press, Boca Raton
29. Zacharias M (2003) Protein-protein docking with a reduced protein model accounting for side chain flexibility. Prot Sci 12:1271–1282
30. Fujitsuka Y, Takada S, Luthey-Schulten ZA, Wolynes PG (2004) Optimizing physical energy functions for protein folding. Proteins 54:88–103
31. Curcó D, Nussinov R, Alemán C (2007) Coarse-grained representation of β-helical protein building blocks. J Phys Chem B 111:10538–10549
32. Hori N, Chikenji G, Berry RS, Takada S (2009) Folding energy landscape and network dynamics of small globular proteins. Proc Natl Acad Sci USA 106:73–78
33. Zhang Z, Lu L, Noid WG, Krishna V, Pfaendtner J, Voth GA (2008) A systematic methodology for defining coarse-grained sites in large biomolecules. Biophys J 95:5073–5083
34. Gabdoulline RR, Wade RC (1996) Effective charges for macromolecules in solvent. J Phys Chem 100:3868–3878
35. Berardi R, Muccioli L, Orlandi S, Ricci M, Zannoni C (2004) Mimicking electrostatic interactions with a set of effective charges: a genetic algorithm. Chem Phys Lett 389:373–378
36. Skepö M, Linse P, Arnebrant T (2006) Coarse-grained modeling of proline rich protein 1 (PRP-1) in bulk solution and adsorbed to a negatively charged surface. J Phys Chem B 110:12141–12148
37. Cascella M, Neri MA, Carloni P, Dal Peraro M (2008) Topologically based multipolar reconstruction of electrostatic interactions in multiscale simulations of proteins. J Chem Theory Comput 4:1378–1385
38. Ha-Duong T (2010) Protein backbone dynamics simulations using coarse-grained bonded potentials and simplified hydrogen bonds. J Chem Theory Comput 6:761–773
39. Ayton GS, Noid WG, Voth GA (2007) Multiscale modeling of biomolecular systems: in serial and in parallel. Curr Opin Struct Biol 17:192–198
40. Yang L-W, Chng C-P (2008) Coarse-grained models reveal functional dynamics–I. Elastic network models–theories, comparisons and perspectives. Bioinf Biol Insights 2:25–45
41. Chng C-P, Yang L-W (2008) Coarse-grained models reveal functional dynamics–II. Molecular dynamics simulation at the coarse-grained level–Theories and biological applications. Bioinf Biol Insights 2:171–185
42. Clementi C (2008) Coarse-grained models of protein folding: toy models or predictive tools? Curr Opin Struct Biol 18:10–15
43. Kamerlin SCL, Vicatos S, Dryga A, Warshel A (2011) Coarse-grained (Multiscale) simulations in studies of biophysical and chemical systems. Annu Rev Phys Chem 62:41–64
44. Wu C, Shea J-E (2011) Coarse-grained models for protein aggregation. Curr Opin Struct Biol 21:209–220
45. Leherte L, Allen FH (1994) Shape information from critical point analyses of calculated electron density maps: application to DNA-drug systems. J Comput Aided Mol Des 8:257–272
46. Becue A (2004) Development of an original genetic algorithm method dedicated to complementarity studies between protein-protein and protein-nucleic acid macromolecular partners. Ph.D. Thesis, University of Namur
47. Becue A, Meurice N, Leherte L, Vercauteren DP (2004) Evaluation of the protein solvent-accessible surface using reduced representations in terms of critical points of the electron density. J Comput Chem 25:1117–1126
48. Becue A, Meurice N, Leherte L, Vercauteren DP (2008) In: Boeyens JCA, Ogilvie JF (eds) Models, mysteries, and magic of molecules. Springer, Dordrecht
49. Leherte L, Guillot B, Vercauteren DP, Pichon-Pesme V, Jelsch C, Lagoutte A, Lecomte C (2007) In: Matta CF, Boyd RJ (eds) The quantum theory of atoms in molecules—from solid state to dna and drug design. Wiley-VCH, Weinheim
50. Leherte L (2004) Hierarchical analysis of promolecular full electron-density distributions: description of protein structure fragments. Acta Crystallogr D 60:1254–1265
51. Leherte L, Vercauteren DP (2009) Coarse point charge models for proteins from smoothed molecular electrostatic potentials. J Chem Theory Comput 5:3279–3298
52. Baker NA, Sept D, Joseph S, Holst MJ, McCammon JA (2001) Electrostatics of nanosystems: application to microtubules and the ribosome. Proc Natl Acad Sci USA 98:10037–10041
53. APBS—Adaptive Poisson-Boltzmann Solver (APBS) (2011) Software for evaluating the electrostatic properties of nanoscale biomolecular systems. http://www.poissonboltzmann.org/apbs/. Accessed 3 January 2011
54. Wang J, Cieplak P, Kollman PA (2000) How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules. J Comput Chem 21:1049–1074
55. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The protein data bank. Nucleic Acids Res 28:235–242
56. RCSB PDB Protein Data Bank (2009) http://www.rcsb.org/pdb. Accessed 3 January 2011
57. Tjong H, Zhou H-X (2008) On the dielectric boundary in Poisson-Boltzmann calculations. J Chem Theory Comput 4:507–514
58. Hart RK, Pappu RV, Ponder JW (2000) Exploring the similarities between potential smoothing and simulated annealing. J Comput Chem 21:531–552
59. Leung Y, Zhang JS, Xu Z-B (2000) Clustering by scale-space filtering. IEEE T Pattern Anal 22:1396–1410
60. Borodin O, Smith GD (2011) Force Field Fitting Toolkit, The University of Utah. http://www.eng.utah.edu/~gdsmith/fff.html. Accessed 3 January 2011
61. Dolinsky TJ, Nielsen JE, McCammon JA, Baker NA (2004) PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. Nucl Acids Res 32:W665–W667
62. PDB2PQR, An Automated Pipeline for the Setup, Execution, and Analysis of Poisson-Boltzmann Electrostatics Calculations (2007) SourceForge Project Page. http://pdb2pqr.sourceforge.net/. Accessed 3 January 2011
63. Eisenmenger F, Hansmann UHE, Hayryan S, Hu C-K (2006) An enhanced version of SMMP-open-source software package for simulation of proteins. Comput Phys Comm 174:422–429
64. Simple Molecular Mechanics for Proteins. http://www.smmp05.net/. Accessed 3 January 2011
65. Nemethy G, Gibson KD, Palmer KA, Yoon CN, Paterlini G, Zagari A, Rumsey S, Scheraga HA (1992) Energy parameters in polypeptides. 10. Improved geometrical parameters and nonbonded interactions for use in the ECEPP/3 algorithm, with application to proline-containing peptides. J Phys Chem 96:6472–6484

66. Simms AM, Toofanny RD, Kehl C, Benson NC, Daggett V (2008) Dynameomics: design of a computational lab workflow and scientific data repository for protein simulations. Prot Eng Des Sel 21:369–377
67. DYNAMEOMICS (2007) The Daggett Group at the University of Washington. http://www.dynameomics.org/. Accessed 3 January 2011
68. Heisterberg DJ (1990), Technical report, Ohio Supercomputer Center, Translation from FORTRAN to C and Input/Output by Labanowski J, Ohio Supercomputer Center
69. CCL quaternion-mol-fit (1999) Computational Chemistry List, ltd. http://www.ccl.net/cca/software/SOURCES/C/quaternion-mol-fit/. Accessed 3 January 2011
70. Exner TE, Mezey PG (2002) Ab initio-quality electrostatic potentials for proteins: an application of the ADMA approach. J Phys Chem A 106:11791–11800
71. Bliznyuk AA, Rendell AP, Allen TW, Chung S-H (2001) The potassium ion channel: comparison of linear scaling semiempirical and molecular mechanics representations of the electrostatic potential. J Phys Chem B 105:12674–12679
72. Gascon JA, Leung SSF, Batista ER, Batista VS (2006) A Self-consistent space-domain decomposition method for QM/MM computations of protein electrostatic potentials. J Chem Theory Comput 2:175–186
73. Warshel A, Kato M, Pisliakov AV (2007) Polarizable force fields: history, test cases, and prospects. J Chem Theory Comput 3: 2034–2045
74. Piccinini E, Ceccarelli M, Affinito F, Brunetti R, Jacoboni C (2008) Biased molecular simulations for free-energy mapping: a comparison on the KcsA channel as a test case. J Chem Theory Comput 4:173–183
75. Bond PJ, Sansom MSP (2006) Insertion and assembly of membrane proteins via simulation. J Am Chem Soc 128:2697–2704
76. Bond PJ, Holyoake J, Ivetac A, Khalid S, Sansom MSP (2007) Coarse-grained molecular dynamics simulations of membrane proteins and peptides. J Struct Biol 157:593–605
77. Leherte L, Vercauteren DP (2010) In: Collett CT, Robson CD (eds) Handbook of computational chemistry research. Nova Science Publishers, New York
78. Maciejczyk M, Spasic A, Liwo A, Scheraga HA (2010) Coarse-grained model of nucleic acid bases. J Comput Chem 31: 1644–1655