Contribution to the study of electrostatic properties of proteins from low-resolution electron density distributions and potential functions, Mémoire - Prix de l'Académie Royale de Belgique

Leherte, Laurence

Publication date:
2009

Document Version
Peer reviewed version

FACULTES UNIVERSITAIRES NOTRE-DAME DE LA PAIX
NAMUR

Groupe de Chimie Physique Théorique et Structurale
Laboratoire de Physico-Chimie Informatique
Rue de Bruxelles, 61 – 5000 Namur

CONTRIBUTION TO THE STUDY OF ELECTROSTATIC PROPERTIES OF PROTEINS FROM LOW-RESOLUTION ELECTRON DENSITY DISTRIBUTIONS AND POTENTIAL FUNCTIONS

Mémoire présenté dans le cadre du Concours Annuel de l'Académie Royale de Belgique

Classe des Sciences

Laurence LEHERTE

2009

Préambule

Ce mémoire est introduit auprès de l'Académie Royale de Belgique, Classe des Sciences, en réponse à la question du Groupe III – CHIMIE de l'année 2009 :

« On demande une contribution à l'étude des propriétés électrostatiques dans les molécules, les protéines ou les solides au départ de la fonction de distribution de densité électronique à basse résolution. »

Le sujet qui y est traité concerne l'élaboration de modèles « gros grains » de protéines issus de fonctions de distribution de densité électronique moléculaire lissées et de la fonction dérivée qu'est le potentiel électrostatique moléculaire. Les aspects principaux du travail concernent, d'une part, l'élaboration d'une technique de recherche et d'identification des « gros grains », la détermination de leur charge électrique, et, d'autre part, la validation des modèles au travers d'applications aux systèmes protéiniques.

Foreword

In the present work, we develop protein coarse grain electrostatic models from electron density distribution functions and molecular electrostatic potentials. The main aspects of this work regard, on the one hand, the elaboration of a procedure for the location, the identification, and the charge determination of the coarse grains and, on the other hand, the validation through applications to protein systems.

List of abbreviations

3D	Three-dimensional
AA	Amino Acid
Amber	Assisted Model Building and Energy Refinement
a.u.	atomic unit
BAK	Backbone
CCDC	Crystallographic Data Centre
CG	Coarse Grain
c.o.m.	Center of mass
CP	Critical Point
DNA	Desoxyribonucleic Acid
ED	Electron Density
ENM	Elastic Network Model
FF	Force Field
Gromos	GROningen MOlecular Simulation
hAr	human Aldose reductase
HP7	12-residue β-hairpin peptide
KcsA	Potassium Ion Channel
LJ	Lennard-Jones
MD	Molecular Dynamics
MEP	Molecular Electrostatic Potential
MOF	Minimal Objective Function
NADP	Nicotinamide Adenine Dinucleotide Phosphate
NMA	Normal Mode Analysis
OF	Objective Function
PASA	Promolecular Atom Shell Approximation
PDB	Protein Data Base
rmsd	Root Mean Square Deviation
rmsdV	Root Mean Square Deviation of the electrostatic potential grid values
rmsdμ	Mean Square Deviation of the molecular dipole moment value
SCH	Side Chain
vdW	van der Waals
XRD	X-Ray Diffraction

I. Introduction

Applications of interaction potential functions, parametrized for small or large molecules, require the definition of the electrostatic contributions that commonly involve the determination of atomic point charges. Those contributions are fundamental in that they govern local and global properties, *e.g*., molecular stability, flexibility, … For macromolecules, the sampling of conformational space is however a complex and highly time-consuming task due to the large number of degrees of freedom of the systems and the complexity of the interaction potential functions. It is nevertheless a major interest to relate a protein function to its microscopic description, notably for the study of protein-protein and protein-ligand interactions. For recent years, much effort has been put into accelerating computational techniques such as Molecular Dynamics (MD) and Normal Mode Analysis (NMA) for simulating large biological systems [emp08, hin08, mor08]. Enhancements to these well-known algorithmic procedures are based, notably, on a spatial coarse graining of the molecular structures [vot09]. Rather than simulating the molecules at their atomic level, one reduces their description to a limited set of points, either centered on selected sites/atoms such the Cα atoms of a protein backbone [dor02, emp08], the center of mass (c.o.m.) of specific groups of atoms like amino acid (AA) residues [bas07], the heavy atoms (united atom description) [fuk01, yan06], or a set of merged atoms [goh06]. Elastic Network Models (ENMs) are among those NMA methods wherein coarse grains (CGs) are interacting through harmonic potential functions. Despite their simplicity, sometimes based on the topology of the protein structure only (excluding inter-CG distance information), they have shown to be extremely useful for the modeling of slow large amplitude motions of proteins [kon06, cle08]. It has even been demonstrated that grouping up to 40 residues into a single node essentially produced the same low-frequency modes as the original single Cα node per residue [dor02]. In a very recent work, Zhang *et al*. [zha08] proposed a method to define CGs that reflect the collective motions computed by a Principal Component Analysis of an atomistic MD trajectory. Each CG site is the c.o.m. of a domain, *i.e*., a group of contiguous Cα atoms that move in a highly correlated fashion. Reviews of the progresses on CG-ENM and -MD models can be found in references [chn08, yan08] and in the Introduction of Section III.

Besides the use of simple harmonic functions, the development of CG interaction potential functions is generally made either from atomistic interaction potential [par05] or MD results [izv05,

liu07, car08], *via* experimental data such as B-factors [kon06], or through the fitting of a potential function achieved by matching CG and atomistic distributions [fuk01, car08]. For example, Lyman *et al.* [lym08] presented a new method for fitting spring constants to mean square CG-CG distance fluctuations computed from atomistic MD. One can also cite the Inverse Monte Carlo approach [lyu95], used for iteratively adjusting an effective CG potential function until it matches a target radial distribution function. Consistency between CG and all-atom models can be checked through a statistical mechanics theory as proposed by Noid *et al.* [noi08a, noi08b]. Another example is the parametrization of the MARTINI force field (FF), dedicated to MD simulations of biomolecular systems, and based on the reproduction of partitioning free energies between polar and apolar phases of a large number of chemical systems [mar07, mon08]. The model is based on a four-to-one mapping, *i.e.*, four heavy atoms are represented by a single interaction center, except for small ring-like fragments (Figure I.1). Specifically, AAs consist of one to four side chain beads and one backbone bead [mon08]. Only four main types of interaction sites are defined: polar (P), non-polar (N), apolar (C), and charged (Q). Each particle type has a number of subtypes, which allow for an accurate representation of the chemical nature of the underlying atomistic structure. In the MARTINI FF, only AA residues Arg, Asp, Glu, and Lys, are charged. Such a description was for example applied to protein channels embedded in a lipid membrane environment [tre08].

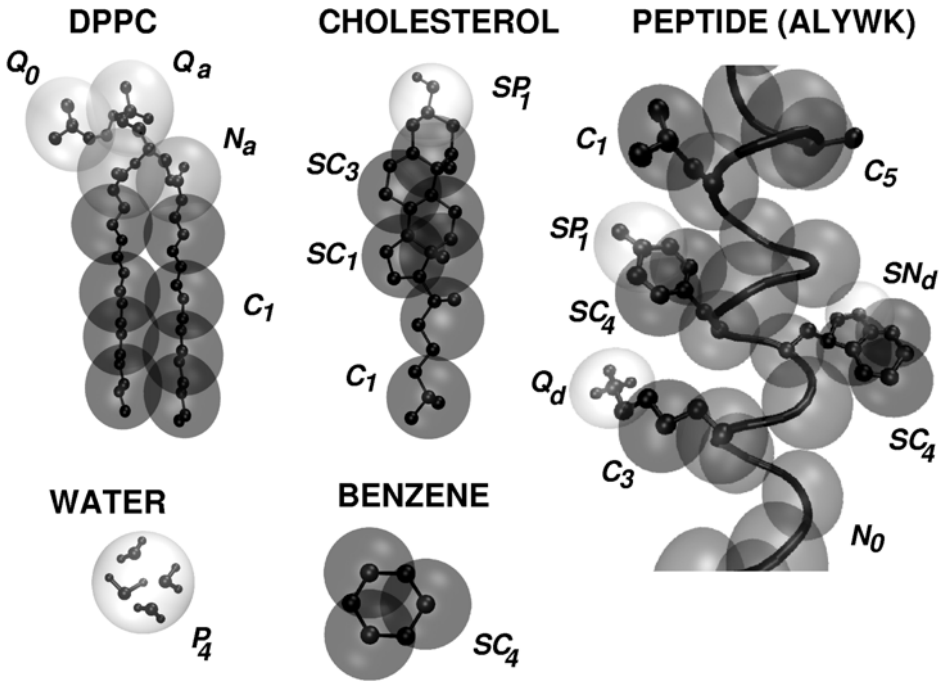


Figure I.1. Illustration depicting MARTINI CG models for molecular structures of various sizes. The illustration is taken from [http://md.chem.rug.nl/~marrink/coarsegrain.html].

In the UNRES model [liw09], a peptidic chain is represented by a sequence of backbone beads located at peptide bonds, while side chains are modelled as single beads attached to the Cα atoms, which are considered only to define the molecular geometry. In the so-called SimFold CG description and energy function, a mixed representation is used. Residues of aqueous proteins are represented by backbone atoms N, Cα, C, O, and H, and one side chain centroid [fuj04, hor09]. In UNRES and SimFold, electrostatic interactions are not explicitely calculated using the Coulomb term like they are in the MARTINI FF.

Multiscale methods, that combine several levels of description, are also appealing since they allow to model limited regions of space with details while limiting the outer regions to coarser models [cle08, she08]. Besides their limitations, *i.e.*, simplified interaction potential functions, neglect of fast motions, faster dynamics than all-atom systems, or partial rigidity of the structure, many studies have found good conformational sampling agreement between features predicted by NMA, for instance, and the observed or simulated conformational change of protein structures, as reported in [dob08]. The consideration of external influences, such as external stresses [eya08] or solvent effects [zho08] can also be treated with CG approaches.

Our first studies on the interaction potential of CG molecular representations, achieved in the frame of a post-doctoral stay at the Cambridge Crystallographic Data Centre (CCDC), were dedicated to a DNA-drug system [leh94]. In that work, a computational method was described for mapping the volume within the DNA double helix structure that is accessible to netropsin, an antitumor antibiotic drug molecule that binds in the minor groove of DNA. Based on a topological analysis of the electron density (ED) of both the DNA and the 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. It was concluded that the global shape of a molecule could be described using local information associated with its centers of high ED, *i.e.*, peaks expanded in terms of ellipsoids. The idea was later extended to the study of supramolecular cyclodextrin-based systems [leh95], zeolitic frameworks [leh97], and protein and DNA complexes [bec03].

In a further comprehensive work, this theory was expanded to model protein-protein and protein-DNA complementarity [bec04b]. The strategy implemented to dock the partners was based on the use of the hereabove mentioned reduced dimensionality representations of biological macromolecules combined with a genetic algorithm. One of the main objectives consisted in the development of an intermolecular interaction function specifically adapted to reduced molecular representations; a recognition score between macromolecules was constructed from statistical

studies of protein-protein and protein-DNA complexes of known structures. The interaction function was a combination of the contact interface area, an electrostatic interaction potential, a steric clash detection procedure, and a contribution related to the macromolecular recognition. This last term was based on a set of distribution tables of preferential distances constructed from statistical analyses of 475 protein-protein complexes and 165 protein-DNA complexes [bec04a, bec07]. The electrostatic potential consisted in a summation over unit charges assigned to the charged residues such as Arg, Lys, Asp, and Glu.

Our first attempt to assign non unitary electric charges to ED peaks was achieved through a collaborative work with the members of the Laboratoire de Cristallographie of the Université de Nancy, directed by Prof. Cl. Lecomte [leh07]. ED distribution of the adenine binding site of the human Aldose reductase (hAr) protein structure and its cofactor $NADP^+$ were calculated using a promolecular analytical approach. ED peaks were located by following the atom trajectories in progressively smoothed ED distributions using a merging/clustering algorithm. To each maxima, it was possible to define their corresponding molecular fragment through a clustering procedure. Molecular electrostatic potentials (MEPs) generated by the adenine binding site in the hAr structure and the electrostatic interaction energies of the adenine moiety with the protein binding site were calculated using several charge models. Two models were built from two sets of atomic charges derived from subatomic resolution XRD data. Each of these two sets was used to calculate the electric charge of ED-based protein fragments centered at ED maxima. An additional charge model was built by assigning formal unit charges to the Arg(+1), Lys(+1), Glu(-1), and Asp(-1) side chains.

Later, the modelling of flexible protein structure was achieved using NMA-based approaches, more specifically ENMs with force constants weighted by the overlap integral value of the fragment ED distribution functions [leh08a, leh08b].

Following our development of an original approach to hierarchically decompose a protein structure into fragments from its ED distribution [leh04, leh07], the method is here applied to MEPs, calculated from point charges as implemented in well-known force fields. To follow the pattern of local maxima and minima in a MEP, as a function of its resolution/degree of smoothing t, the following strategy was adopted. First, each atom of a molecule is considered as a starting point. As the smoothing degree increases, each point moves along a path to reach a location where the MEP gradient value vanishes. Convergence of trajectories leads to a reduction of the number of points. Practically, to determine the protein backbone representations, we analyzed CG models obtained for a β-strand of 15 glycine residues. A fitting algorithm was used to assign charges to the obtained

local maxima and minima *vs*. the unsmoothed MEP, as a function of *t*. The best fit obtained allowed to determine the degree of smoothing to be considered. Then, the influence of the different AA side chains was studied at the selected value of *t* for different rotamers by substituting the central glycine residue.

A description of the basic theory is reported in Section II of the present document. Section III consists in the reproduction of our first comprehensive research work about the analysis of smoothed Amber-based MEP and the determination of CG point charge models applicable to protein structures. In Sections IV and V, we report how to improve CG models built from smoothed MEP, and we extend that approach to the set of charges used in the FF Gromos43A1. An automated procedure to generate electrostatic CG representations of proteins is then described in Section VI and applied to a large ion channel protein system. Finally, general conclusions and perspectives, that include a discussion about transferability, are presented in Section VII.

II. Topology of Low-Resolution or Smoothed Molecular Electron Density Distributions - Applications

In this Section, we present the fundamental concepts and procedures that are useful for the understanding of our approach. They complete the Theoretical Background part that appear in Section III; some redundancies may thus occur.

Electron Density Calculation – Crystallography-Based Approach

The intensity of X-rays diffracted by a crystalline structure is proportional to the modulus of their corresponding structure factor $F(\mathbf{h})$:

$$F(\mathbf{h}) = \left|F(\mathbf{h})\right| e^{-i\varphi(\mathbf{h})} \tag{1}$$

where **h** is a reciprocal space vector with indices *h, k*, and *l*, an $\varphi(\mathbf{h})$ is the phase of the diffracted wave. Within the crystallographic approach, the electron density (ED) distribution function $\rho(\mathbf{r})$ is calculated as the Fourier Transform of $F(\mathbf{h})$:

$$\rho(\mathbf{r}) = \frac{1}{V} \sum_{\{\mathbf{h}\}} F(\mathbf{h}) e^{-2\pi i \mathbf{h}.\mathbf{r}} \tag{2}$$

V being the volume of the unit cell. Such ED maps can be calculated at various resolution levels through the simulation of X-ray diffraction experiments using programs such as XTAL [hal90].

In practice, the number of known structure factors occurring in equation (2) is not infinite and varies with the resolution. In crystallography, the resolution factor d_{min} is a well-known concept which is theoretically defined using Bragg's law:

$$(\frac{\sin\theta}{\lambda})_{\max} = \frac{1}{2d_{\min}} \tag{3}$$

where 2θ is the angle between the diffracted and the primary beams of wavelength λ, and d_{min} depends on different parameters including the quality of the crystal, the chemical composition, the radiation used, and the temperature of the experiment. For example, Figure II.1 depicts the ED distributions of the Diazepam molecule calculated using XTAL at various resolution levels.

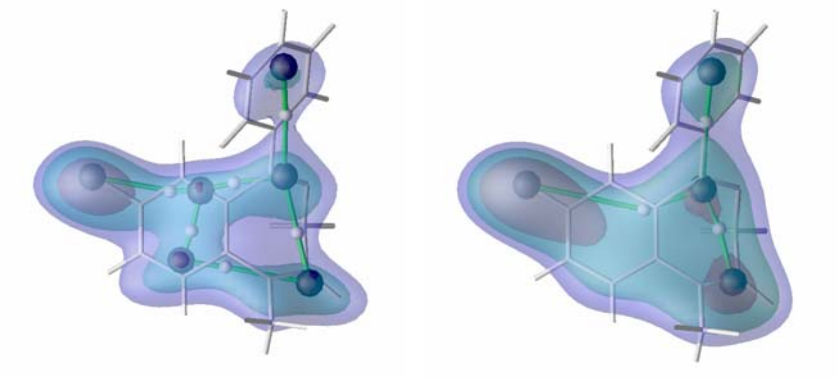


Figure II.1. Iso-contours of the ED distributions calculated for Diazepam using XTAL at a resolution of (left) 2.5 Å (iso = 1.5, 2, 3 $e^-/Å^3$) and (right) 3.0 Å (iso = 1.2, 1.5, 1.9 $e^-/Å^3$), with superimposition on the local maxima (black spheres) and saddle points (white spheres).

Critical Point Analysis

An ED distribution $\rho(\mathbf{r})$ can be described in terms of the location and identification of its critical points (CPs), *i.e.*, points where the gradient of the density is equal to zero. They are thus characterized as maxima, minima, or saddle points depending upon the sign of the second derivatives of $\rho(\mathbf{r})$. The Hessian matrix **H** of a continuous 3D function such as the ED is built from its second derivatives:

$$\mathbf{H} = \begin{pmatrix} \frac{\partial^2\rho}{\partial x^2} & \frac{\partial^2\rho}{\partial x\partial y} & \frac{\partial^2\rho}{\partial x\partial z} \\ \frac{\partial^2\rho}{\partial y\partial x} & \frac{\partial^2\rho}{\partial y^2} & \frac{\partial^2\rho}{\partial y\partial z} \\ \frac{\partial^2\rho}{\partial z\partial x} & \frac{\partial^2\rho}{\partial z\partial y} & \frac{\partial^2\rho}{\partial z^2} \end{pmatrix} \tag{4}$$

This real and symmetric matrix can be diagonalized. The three resulting eigenvalues provide informations relative to the local curvature; the Laplacian $\nabla^2\rho(\mathbf{r})$, which is the summation over the three eigenvalues, gives details about the local concentration (sign < 0) or depletion (sign > 0) of the ED. If the rank (number of non zero eigenvalues) of the diagonalized matrix is 3, then four cases are met. The signature (sum of the sign of the eigenvalues) s = -3 corresponds to a local maximum or peak, *i.e.*, the ED function adopts maximum values along each of the three principal directions x', y', and z'. s = -1 corresponds to a saddle point or pass where two of the eigenvalues are negative; these are also called Bond Critical Points. s = +1 corresponds to a saddle point or pale characterized by only one negative eigenvalue. s = +3 corresponds to a local minimum or pit, *i.e.*, the ED function adopts minimum values along each of the three principal directions.

Morse theory allows to determine whether the set of critical points is topologically consistent. It is applicable to functions which are everywhere twice differentiable, and wherein there is no degenerate critical points, *i.e.*, no zero eigenvalues of the Hessian matrix at the critical point locations. Considering M_k as the number of CPs with index k of the function $\rho(\mathbf{r})$, then:

$$M_3 - M_2 + M_1 - M_0 = 1 \tag{5}$$

where M_3, M_2, M_1, and M_0 stand for the number of peaks, passes, pales, and pits, respectively [lio93]. In the case of crystals, the CP network is defined not only by the molecular structure but also by the lattice periodicity and the space group symmetry. Due to periodic boundary conditions, a unit cell can be considered as a 3D torus, each pair of opposite faces being connected. This means that the motif of CPs is not isolated but interacts with its periodic images and, therefore, the number of CPs is constrained by the relationship:

$$M_3 - M_2 + M_1 - M_0 = 0 \tag{6}$$

At atomic resolution, peaks and passes are normally associated with the presence of atoms and chemical bonds, respectively, while pales and pits occur as a result of the geometrical arrangement of the atoms and the corresponding networks of bonds. Pales and pits are found in the interior of rings and cages, respectively [bad95]. Figure II.2 represents a cubic network of CPs, with one peak located at each of the 8 corners. The 8 Gaussian functions built on these peaks generate a pass on each of the edges, pales centered on the 6 faces, and one pit located in the center of the cube.

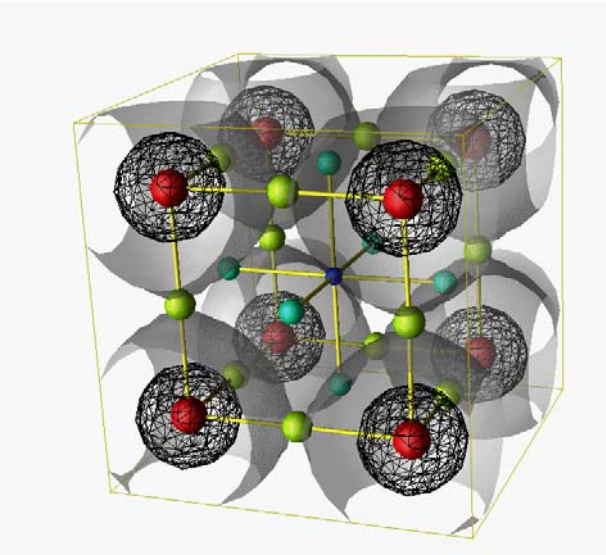


Figure II.2. Critical point network 'peak (red) – pass (yellow)' and 'pit (dark blue) – pale (light blue)' of a cubic arrangement of three-dimensional Gaussian functions.

Shape Reconstruction

At the CP locations, the three main curvatures of the ED function are the eigenvalues of the Hessian matrix constructed from the second derivatives. It is assumed that this local information can be transferred to the space surrounding the CP concerned; hence it is possible to evaluate (or reconstruct) the 3D function in the close neighbourhood of each point. Each maximum of the ED function, *i.e.*, each peak, is considered as the center of expansion of a Gaussian function and such a mathematical expression is fitted in order to define a volume around each peak taking into account its three characteristic eigenvalues:

$$\rho(\mathbf{r}) = \rho_0 \, e^{\frac{\alpha \mathbf{r}^T \mathbf{H}' \mathbf{r}}{\rho_0}} \tag{7}$$

where **H**' is the diagonalized form of **H**, and **r** is defined in a reference frame built on the three corresponding eigenvectors. In order to evaluate the volume associated with a particular peak, the exponential term of the Gaussian function can be integrated over the space within the frame of an ellipsoid:

$$\int e^{\alpha \mathbf{r}^T \mathbf{H}' \mathbf{r} / \rho_0} d\mathbf{r} \tag{8}$$

characterized by three main axes r_X, r_Y, and r_Z:

$$V = \frac{\pi^{3/2} \rho_0^{3/2}}{2^{3/2} \sqrt{|h_X h_Y h_Z|}} = \frac{4\pi}{3} r_X r_Y r_Z \tag{9}$$

and hence provides a method of representing shape anisotropy of the CPs. This shape description is extended to a whole molecule by considering a set of ellipsoids, and a descriptor for the resulting structure can be defined in terms of interaction energy values, as described below.

Steric Interaction Energy

In a study on the DNA-netropsin system [leh94], the total interaction energy *E* between the host ellipsoids and a guest probe was expressed within a pseudo-pair potential approximation, wherein the dispersive interaction between an ellipsoid *i* and a sphere *j* is proportional to their volume product:

$$E_{ij} = -\frac{A_{ij}}{r_{ij}^6} + \frac{B_{ij}}{r_{ij}^{12}} \text{ where } A_{ij} > 0 \text{ and } B_{ij} > 0 \tag{10}$$

$$A_{ij} = V_i V_j \tag{11}$$

$$B_{ij} = A_{ij}(r_i + r_j)^6 \tag{12}$$

r_{ij} being the separation distance between particles *i* and *j*, and r_i and r_j, their radius calculated along the interdistance vector *i-j*. In such a formula, it is considered that the equilibrium distance between *i* and *j* is given by $2^{1/6}(r_i + r_j)$ and $E_{ij} = 0$ when $r_{ij} = (r_i + r_j)$. The idea of using a pseudo-potential energy function to determine the optimal steric location of a guest molecule was also developed by Kuntz and coworkers [sho93, gro94, goo95] in the program DOCK. These authors simplified the overlap energy between two molecules to a contribution depending upon the van der Waals (vdW) radii of the interacting atoms and their separation distance. Considering a LJ type potential allowed us to emphasize the effect of global curvature of the neighbourhood, *e.g.*, a cavity leading to more attractive energies. Fitting Gaussian functions to a higher resolution representation, *i.e.*, to atoms, has been done latter by Grant and Pickup [gra95] to overcome the limitations of hard sphere representations of molecular shapes. From such functions, these authors were able to derive gradients and Hessian of the nuclear coordinate derivatives, *i.e.*, properties similar to CP characteristics.

Electron Density Calculation – Promolecular Approach

Promolecular models have often turned out to lead to very good approximated representations of ED distributions for the purpose of a number of applications as varied as chemical bond analysis or molecular similarity applications for example [tsi98a, tsi98b, gir98, mit00, gir01, dow02, bul03]. In the Promolecular Atomic Shell Approximation (PASA) approach, a promolecular ED distribution ρ_M is calculated as a weighted summation over atomic ED distributions ρ_a, which are described in terms of series of squared *1s* Gaussian functions fitted from atomic basis set representations [ama97]:

$$\rho_a(\mathbf{r}-\mathbf{R}_a)=\sum_{i=1}^{5} w_{a,i}\left[\left(\frac{2\varsigma_{a,i}}{\pi}\right)^{3/4} e^{-\varsigma_{a,i}|\mathbf{r}-\mathbf{R}_a|^2}\right]^2 \tag{13}$$

where $\mathbf{R}_a$ is the position vector of atom a, and $w_{a,i}$ and $\zeta_{a,i}$ are the fitted parameters, respectively, as reported in the Web site at [http://iqc.udg.es/cat/similarity/ASA/funcset.html]. ρ_M is then calculated as:

$$\rho_M = \sum_a Z_a \rho_a \tag{14}$$

where Z_a is the atomic number of atom a.

In one of our approaches to generate low resolution 3D functions [leh01], an ED map is a deformed version of ρ_M that is directly expressed as the solution of the diffusion equation according to the formalism presented by Kostrowicki *et al*. [kos91]:

$$\rho_{a,t}(\mathbf{r}-\mathbf{R}_a)=\sum_{i=1}^{5} a_{a,i}(1+4b_{a,i}t)^{-3/2} e^{\frac{-b_{a,i}|\mathbf{r}-\mathbf{R}_a|^2}{1+4b_{a,i}t}} \tag{15}$$

where:

$$b_{a,i}=2\varsigma_{a,i} \qquad a_{a,i}=w_{a,i}\left(\frac{b_{a,i}}{\pi}\right)^{6/4} \tag{16}$$

In this context, t is seen as the product of a diffusion coefficient with time. It has also been shown that t is equivalent to the well-known crystallographic anisotropic displacement parameter u^2

[leh04]. Figure II.3 shows the evolution of the ED distribution of Diazepam calculated at the RHF-MO-LCAO 6-31G* level as t increases from 0.0 bohr^2 (original PASA distribution) to 2.5 bohr^2.

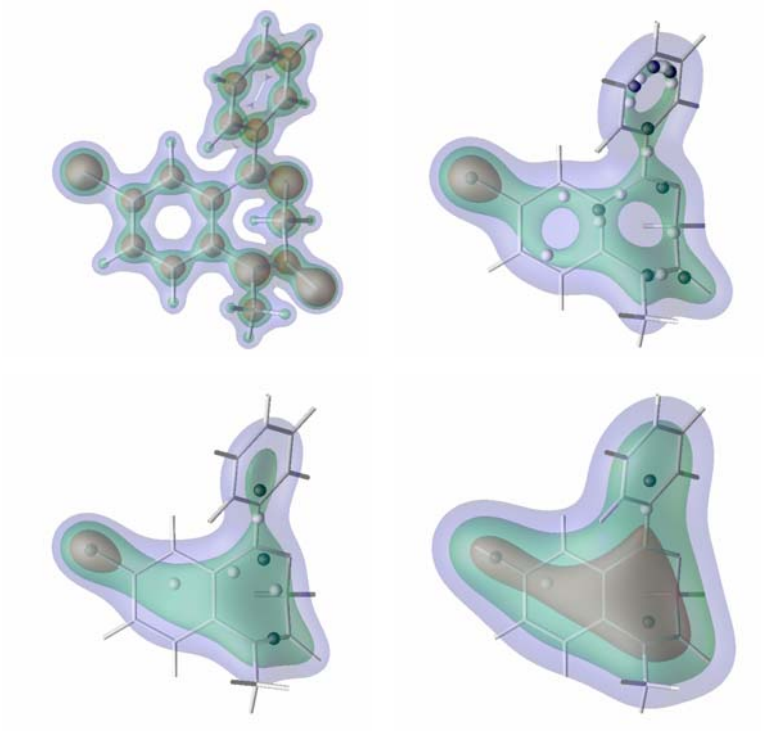


Figure II.3. Iso-contours of the ED distributions calculated for Diazepam at the PASA level with (top left) t = 0.0 bohr^2 (iso = 0.1, 0.2, 0.3 e$^-$/bohr^3), (top right) t = 1.1 bohr^2 (iso = 0.1, 0.15, 0.2 e$^-$/bohr^3), (bottom left) t = 1.6 bohr^2 (iso = 0.10, 0.125, 0.15 e$^-$/bohr^3), and (bottom right) t = 2.5 bohr^2 (iso = 0.05, 0.075, 0.10 e$^-$/bohr^3), with superimposition on the local maxima (black spheres) and saddle points (white spheres).

Merging/Clustering Technique

One way to follow patterns of CPs, and more particularly of the peaks, as a function of the degree of smoothing is to implement the algorithm proposed by Leung *et al*. [leu00]. These authors proposed a method to model the blurring effect in human vision. This was achieved (i) by filtering a digital image *p(x)* through a convolution product with a Gaussian function *g(x,t)*:

$$g(x,t) = \frac{1}{t\sqrt{2\pi}} e^{-x^2/2t^2} \tag{17}$$

where t is the scale parameter, and (ii) 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:

$$x(n+1) = x(n) + h\nabla_x p(x,t) \tag{18}$$

where h is defined by the authors as the step length. We have adapted this idea to 3D images such as PASA ED distribution functions. In this framework, the original ED corresponds to the PASA ED approximation at scale t = 0 where each atom of a molecular structure is considered as the

starting points of a merging procedure. As t increases continuously from 0.0 to a given maximal value, each peak moves continuously along the gradient path to reach a location in the 3D space where $\nabla\rho = 0$.

On a practical point of view, this consists in following the trajectory of the peaks obtained at a resolution (t - Δt) on the ED distribution surface calculated at resolution level t. Once all peak locations are found, close peaks are merged and the procedure is repeated for each selected value of t until the whole set of maxima becomes one single point.

The results obtained using this algorithm can be visualized in terms of dendrograms *via* the program Phylodendron [gil96]. The dendrogram obtained for the Diazepam molecule and the corresponding significant substructures, *i.e.*, the chlorine atom (I), the carbonyl function (II), the imine group (III), and the phenyl moiety (IV), are shown in Figure II.4.

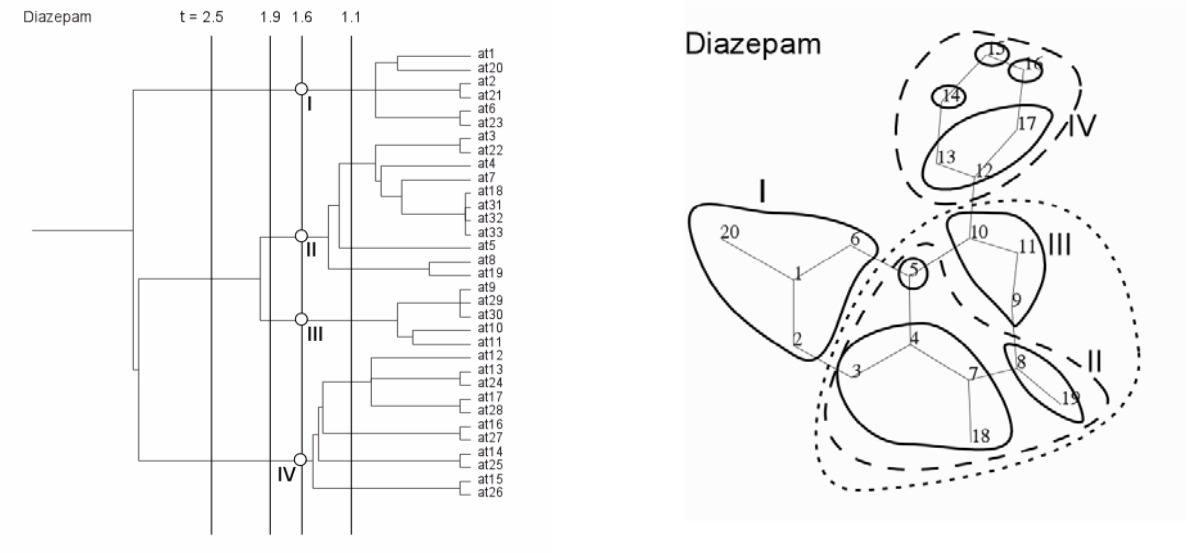


Figure II.4. (left) Dendrogram obtained from the hierarchical merging/clustering procedure applied to PASA ED maps of Diazepam; results at t = 1.1, 1.6, 1.9, and 2.5 bohr^2 are emphasized using vertical lines. (right) Contours of the molecular fragments shown at resolution values t = 1.1 (plain line), 1.6 (long dashed line), 1.9 (dash-dot line), and 2.5 bohr^2 (dotted line).

Elastic Network Models

Normal Mode Analysis (NMA) is a classical technique for studying the vibrational (and thermal) properties of molecular structures. Although this technique is widely used for molecular systems consisting of a small number of atoms, performing NMA on large-scale systems, as proteins, is computational challenging, and reduced representations, often based on Cα atoms only, are used

(Figure II.5). So-called Elastic Network Models (ENM) are, for example, built by connecting the neighboring residues of a protein by springs with a harmonic and uniform force constant.

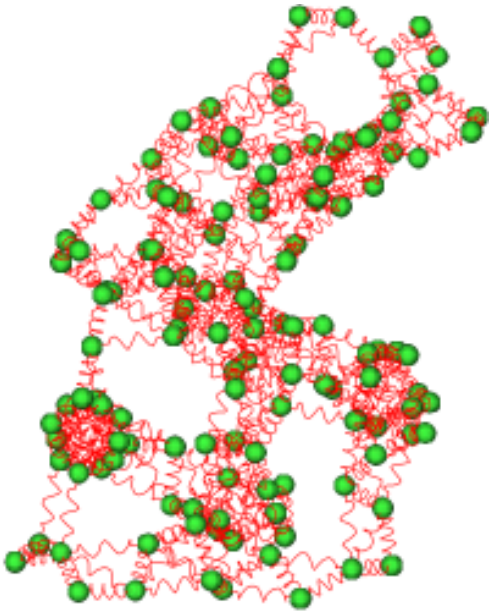


Figure II.5. Illustration of an ENM for lysozyme using a Cα representation (green spheres). The interactions are represented by springs (in red) between pairs of atoms. Figure is taken from [Hinsen, K., Elastic Network Models for Proteins, Theoretical Biophysics, Molecular Simulation, and Numerically Intensive Computation, http://dirac.cnrs-orleans.fr/plone/Members/hinsen/elastic-network-models-for-proteins, 2007].

Mathematically, the motion of the molecule can be described by a second order ordinary differential equation:

$$\ddot{\mathbf{r}} + \mathbf{F}\,\mathbf{r} = 0 \tag{19}$$

where the matrix **F** is a force constant matrix built from the second derivative of the potential with respect to the Cartesian coordinates. The standard procedure for solving this equation is to diagonalize the matrix **F** by computing its eigenvalues and eigenvectors. Each eigenvector is often referred to as a normal mode with a vibrational frequency ω, determined by the eigenvalue. The overall dynamics of the molecular system can be described by a superposition of a number of linearly independent normal modes. When working with CG representations, the low-frequency region of the spectrum is particularly interesting because it has been shown that the lowest modes are able to capture collective conformational changes that are hard to access by all-atom MD simulations. In NMA, once the normal mode vectors, their frequencies, and their amplitudes are obtained, various properties can be calculated easily, such as the mean fluctuations of atom positions and their mutual correlations.

In recent studies [leh08a, leh08b], we have evaluated the suitability of smoothed PASA ED overlap integrals:

$$\int \rho_{i,t}(\mathbf{r})\rho_{j,t}(\mathbf{r})d\mathbf{r} \tag{20}$$

to model force constants associated with springs between ED fragments centered on ED maxima (peaks). It was verified that the decomposition of a 3D protein structure based on its ED smoothed at t = 1.4 bohr^2 allowed to describe its structure in terms of backbone and side chain fragments, each associated with an ED peak. This description is comparable to a representation obtained at a crystallographic resolution value of about 3 Å. ENM built from the backbone ED maxima are thus similar to models built from Cα coordinates.

Topology of the Molecular Electrostatic Potential

Topological analysis of scalar fields other than ED distributions is not very common. Calculations applied to 3D properties, such as molecular electrostatic potentials (MEPs), have nevertheless been proposed by Gadre *et al.* [gad96] and Leboeuf *et al.* [leb99].

MEP, a well-known property of a free molecule for examining its reactivity towards nucleophiles, is derived from its molecular ED $\rho_A(\mathbf{r})$ as:

$$V_A(\mathbf{r}) = \sum_a \frac{Z_a}{|\mathbf{r} - \mathbf{R}_a|} - \int d\mathbf{r}' \frac{\rho_A(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} \tag{21}$$

where a, Z_a, and $\boldsymbol{R}_a$ stand for the atoms that constitute the molecular structure A, their atomic numbers, and their vector positions, respectively. Gadre *et al.* [gad96] used MEP topology as a predictive tool for obtaining intermolecular interaction parameters. Leboeuf *et al.* [leb99] showed how MEP CPs are related to the electronic structure (π bonds, lone pairs, …) of the investigated molecules. Pathak and Gadre [pat90] particularly showed that MEP distributions generally lack non nuclear maxima. Politzer *et al*. proposed procedures for predicting sites of nucleophilic attack, such as the use of a MEP of a molecule in a distorded geometry [pol82], or the mapping of MEP values onto a 0.002 a.u. ED isosurface [sjo90]. If promolecular representations, based on non interacting spherical atoms, are not adapted for predicting interactions sites of molecules, they however provide acceptable results for the analysis of MEP minima along the internuclear axes in a molecule [pac92, bot98]. More recently, but in a less direct work on the topology of MEPs, Popelier *et al*. [pop04] studied the electrostatic potential generated by the ED of molecular fragments of retinal and lysine defined by quantum chemical topology principles.

Mata *et al*. [mat07] reported that zero-flux surfaces occurring in a MEP, as for $\rho(\mathbf{r})$, are also observed between atoms, but the actual partition of the space in volumes is different from that of the ED. They reported a topological analysis procedure for the EP. In the case of MEP, the gradient and Laplacian operators have a particular physical significance. The gradient of $V(\mathbf{r})$ is the negative of the electrostatic field while the Laplacian is related to the density distribution $\rho(\mathbf{r})$ by the Poisson equation:

$$\nabla^2 V(\mathbf{r}) = -4\pi\rho(\mathbf{r}) \tag{22}$$

According to Leboeuf *et al*. [leb99], the Poincaré-Hopf relationship that is valid for MEPs is:

$$M_3 - M_2 + M_1 - M_0 = M_+ + M_- \tag{23}$$

where M_+ and M_- stand for the number of asymptotic maxima and minima, respectively. They correspond to regions of solid angle where $V(\mathbf{r})$ is going asymptotically to zero from positive and negative values, respectively. In the case of the ED, $M_+ = 1$ and $M_- = 0$. In their work, Mata *et al.* [mat07] associated each local maxima (nuclei) and minima to electrophilic and nucleophilic sites, respectively, with the corresponding basins indicating their influence zones.

Equation (23) has recently been revisited and generalized by Roy *et al*. [roy08]. By iteratively visualizing the direction of the MEP gradient calculated on spherical surfaces centered at various molecular points, the authors determine the critical points of the MEP function. The resulting Euler characteristic *EC* is defined as:

$$EC = n_0 - n_i \tag{24}$$

where n_0 and n_i stand for the number of regions on a spherical surface where the MEP increases (the gradient points into the sphere) and decreases (the gradient points outside the sphere), respectively.

III. Determination of Protein Coarse-Grain Charges from Smoothed Electron Density Distribution Functions and Molecular Electrostatic Potentials

Original paper accepted in "Computational Chemistry: New Research", Ed. F. Columbus, Nova Publishers, Aug. 26th 2008.

Determination of protein coarse-grain charges from smoothed electron density distribution functions and molecular electrostatic potentials

Laurence Leherte, Daniel P. Vercauteren
Laboratoire de Physico-Chimie Informatique
Groupe de Chimie Physique Théorique et Structurale
University of Namur (FUNDP)
Rue de Bruxelles, 61
B-5000 Namur (Belgium)

Abstract

The design of protein coarse-grain (CG) models and their corresponding interaction potentials is an active field of research, especially for solving problems such as protein folding, docking, … Among the essential parameters involved in CG potentials, electrostatic interactions are of crucial importance since they govern local and global properties, *e.g.*, their stability, their flexibility, …

Following our development of an original approach to hierarchically decompose a protein structure into fragments from its electron density (ED) distribution, the method is here applied to molecular electrostatic potential (MEP) functions, calculated from point charges as implemented in well-known force fields (FF). To follow the pattern of local maxima (and minima) in an ED or a MEP distribution, as a function of the degree of smoothing, we adopted the following strategy. First, each atom of a molecule is considered as a starting point (a peak, or a pit for negative potentials in a MEP analysis). As the smoothing degree increases, each point moves along a path to reach a location where the ED or MEP gradient value vanishes. Convergences of trajectories lead to a reduction of the number of points, which can be associated with molecular fragments.

Practically, to determine the protein backbone representations, we analyzed CG models obtained for an extended strand of polyglycine. The influence of the different amino acid side chains was then studied for different rotamers by substituting the central glycine residue. Regarding

the determination of charges, we adopted two procedures. First, the net charge of a fragment was calculated as the summation over the charges of its constituting atoms. Second, a fitting algorithm was used to assign charges to the obtained local maxima/minima.

Applications to a literature case, a 12-residue β-hairpin peptide, are also presented. It is observed that classical CG models are more similar to ED-based models, while MEP-based descriptions lead to different CG motifs that better fit the MEP distributions.

Introduction

The design of coarse-grain (CG) models [1] and their corresponding potential functions [2] for protein computational studies is currently an active field of research, especially in solving long-scale dynamics problems such as protein folding, protein-protein docking, … For example, to eliminate fast degrees of freedom, it has been shown that one can rely on CG representations only, or on mixtures of CG and more detailed descriptions [3,4] in order to significantly increase the time step in molecular dynamics (MD) simulations. Among the parameters involved in CG potentials, the electrostatic interactions are of major importance [5] since they govern local and global properties such as their stability [6], their flexibility [7], …

Common approaches used to design a CG description of a protein consist in reducing groups of atoms into single interaction sites. For example, in reference [8], each amino acid (AA) is represented by a single spherical site, with unit or nul electric charge. The authors studied a proline-rich protein PRP-1 interacting with a mica surface using Monte-Carlo simulations. Curcó *et al.* [9] developed a CG model of β-helical protein fragments where the AAs are represented by two, three, or four blobs depending upon the AA type, in accordance with a best fitting between Monte-Carlo based all-atom and CG energies. In their work, the AAs are depicted by the amide hydrogen atom HN, the oxygen atom, the geometric center of the side chain (except for Gly), and a fourth blob whose position depends on the AA type (except for Gly, Ala, and Val). In reference [10], each AA residue is modeled using one sphere located on the geometric center of the backbone and one or two spheres located on the geometric centers of the side chain fragments (except for Gly). Differently, Pizzitutti *et al.* [11] represented each AA of a protein sequence by a charged dipolar sphere. For each AA, one CG sphere is located on the center-of-mass (c.o.m.) of the uncharged residues, while two CG spheres are assigned to the c.o.m. of the neutral part of the AA residue and to the c.o.m. of the charged part, respectively. Charged residues are Lys, Arg, Glu, Asp, and terminal AAs. The

authors show that, in protein association, their model provides a good approximation of the all-atom potential if the distance between the protein surfaces is larger than the diameter of a solvent molecule.

As mentioned earlier, a CG potential can be combined with an all-atom potential. For example, Neri *et al*. [3] included a CG description, in which the potential energy is expressed as harmonic terms between close Cα and/or Cβ atoms. Such elastic network representations are well-known to study the slow large amplitude dynamics of protein structures [12-15]. The small biologically relevant region of the protein is modeled using an atom-based potential while the remaining part of the protein is treated using a CG model. In this context, Heyden and Thruhlar [16] proposed an algorithm allowing a change in resolution of selected molecular fragments during a MD simulation, with conservation of energy and angular momentum. A different and relatively logic way of considering the combination between all-atom and CG potentials is to use CG as a pre-processing stage carried out to establish starting conformations for all-atom MD simulations [4].

Even when one uses an all-atom representation to model a protein structure, a reduced set of Coulomb charges can still be used. For example, Gabb *et al*. [17] reported protein docking studies where electrostatic complementarity is evaluated by Fourier correlation. Charges used in Coulomb electrostatic fields were close to unit charges and placed on a limited set of atoms. Besides the use of unit charges as in [8,11,18], an approach to assign an electrostatic charge to a fragment or pseudo-atom is to sum over the corresponding atomic charges. Extended approaches involve the assignment of dipolar and quadrupolar contributions to the CGs [19]. In this last work [19], dedicated to small molecules such as benzene, methanol, or water, the charge distribution is represented by point multipolar expansions fitted to reproduce MD simulation data. Without being exhaustive, other assignment methods consist in fitting the CG potential parameters so as to reproduce at best the all-atom potential values [9,19].

In this chapter, we present two approaches to design and evaluate CG electrostatic point charges. The first one has already been described in a previous work regarding the evaluation of the electrostatic interactions between Aldose Reductase and its ligand [20]. In that first approach, the fragment content is determined through a merging/clustering procedure of atom trajectories generated in progressively smoothed electron density (ED) distribution functions. The specific use of a Gaussian promolecular representation of an ED, *i.e*., a model where a molecule is the superposition of independent and spherical atoms, allows a fast evaluation of the ED distribution as well as their derived properties such as derivatives and integrals. In the second approach, atoms are clustered according to their trajectories defined in a smoothed molecular electrostatic potential

(MEP) function. As the charge calculation approach useful in ED cases revealed to be inefficient in MEP cases, a fitting algorithm is applied to evaluate CG charges. Results are presented for the 20 AAs, first as derived from a promolecular ED representation, and second from the all-atom Amber charges reported in Duan *et al*. [21]. In this last work, the authors developed a third-generation point charge all-atom force field for proteins. Charges were obtained by a fitting to the MEP of dipeptides calculated using B3LYP/cc-pVTZ//HF/6-31G** quantum mechanical approaches in the PCM continuum solvent in a low dielectric to mimic an organic environment similar to that of the protein interior.

Finally, we will show that the CG charges obtained for each AA residue can be used to determine a CG model representation for any protein. A particular application to a literature case, a 12-residue β-hairpin HP7 [10], is described and MEP results are compared with published models.

Theoretical Background

In this section, we present the mathematical formalisms that were needed to design a protein CG representation and its point charges. First, the smoothing algorithm that is applicable to both ED and MEP functions is described. This description is followed by the mathematical expressions needed to smooth either a Gaussian-based ED distribution function, or the Coulomb electrostatic interaction function. Finally, the two approaches used to calculate CG point charges, from ED- and MEP-based CG, respectively, are detailed.

Smoothing Algorithm

An algorithm initially described by Leung *et al*. [22] was implemented to follow the pattern of local maxima in a Gaussian promolecular ED or a MEP function, as a function of the degree of smoothing. More particularly, the authors proposed a method to model the blurring effect in human vision, which is achieved (*i*) by filtering a digital image *p(x)* through a convolution product with a Gaussian function *g(x,t)*:

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

where *t* is the scale parameter, and (*ii*) 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:

$$x(n+1) = x(n) + h\nabla_x p(x,t) \quad (2)$$

where h is defined as the step length. We adapted this idea to three-dimensional (3D) images such as ED and MEP functions, f, such as:

$$\vec{r}_{f(t)} = \vec{r}_{f(t-\Delta t)} + \frac{\Delta}{f(t)}\vec{\nabla} f(t) \quad (3)$$

where $\vec{r}$ stands for the location vector of a point in a 3D function. The various steps of the resulting merging/clustering algorithm are:

1. At scale $t = 0$, each atom of a molecular structure is considered as a local maximum (peak) of the ED and/or a local minimum (pit) of the MEP function. All atoms are consequently 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 $\vec{\nabla} f(t) = 0$. On a practical point of view, this consists in following the trajectory of the peaks and/or pits on the ED or MEP distribution surface calculated at t according to Equation (3). The trajectory search is stopped when $\left|\vec{\nabla} f(t)\right|$ is lower or equal to a limit value, $grad_{lim}$. Once all peak and/or pit 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 maximum or minimum 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.

The results obtained using that algorithm are the location of the local maxima and/or minima, *i.e.*, peaks and pits, and the atomic content of all fragments, at each value of t between 0 and t_{max} [23], that can be further interpreted in terms of dendrograms as, for example, using the Web version of the program Phylodendron [24]. For information, input data were written in the adequate format using DENDRO [25], a home-made program implemented using Delphi, an object-oriented programming language that allows the representation and processing of data in terms of classes of objects.

Promolecular Electron Density Distributions

In their studies related to the Promolecular Atom Shell Approximation (PASA), Amat and Carbó-Dorca used atomic Gaussian ED functions that were fitted on 6-311G atomic basis set results [26]. A molecular or promolecular ED distribution is thus a sum over atomic Gaussian functions wherein expansion coefficients are positive to preserve the statistical meaning of the density function in the fitted structure. In the PASA approach that is considered in the present work, a promolecular ED distribution ρ_M is analytically represented as a weighted summation over atomic ED distributions ρ_a, which are described in terms of series of three squared *1s* Gaussian functions fitted from atomic basis set representations [27]:

$$\rho_a(\vec{r}-\vec{R}_a) = Z_a \sum_{i=1}^{3} w_{a,i} \left[\left(\frac{2\varsigma_{a,i}}{\pi} \right)^{3/4} e^{-\varsigma_{a,i}\left|\vec{r}-\vec{R}_a\right|^2} \right]^2 \tag{4}$$

where $w_{a,i}$ and $\zeta_{a,i}$ are the fitted parameters, respectively, as reported at the Web address http://iqc.udg.es/cat/similarity/ASA/funcset.html. ρ_M is then calculated as:

$$\rho_M = \sum_{a \in A} \rho_a \tag{5}$$

In the present approach to generate smoothed 3D ED functions, ρ_M is directly expressed as the solution of the diffusion equation according to the formalism presented by Kostrowicki *et al*. [28]:

$$\rho_{a,t}(\vec{r}-\vec{R}_a) = Z_a \sum_{i=1}^{3} s_{a,i} \quad \text{where} \quad s_{a,i} = \alpha_{a,i} e^{-\beta_{a,i}\left|\vec{r}-\vec{R}_a\right|^2} \tag{6}$$

with:

$$\alpha_{a,i} = Z_a w_{a,i} \left(\frac{2\varsigma_{a,i}}{\pi} \right)^{3/2} \frac{1}{\left(1+8\varsigma_{a,i}t\right)^{3/2}} \quad \text{and} \quad \beta_{a,i} = \frac{2\varsigma_{a,i}}{\left(1+8\varsigma_{a,i}t\right)} \tag{7}$$

where t is the smoothing degree of the ED. t can also be seen as the product of a diffusion coefficient with time or, in crystallography terms, as the overall isotropic displacement parameter [29]. Unsmoothed EDs are thus obtained by imposing $t = 0$ bohr^2.

Molecular Electrostatic Potentials

The electrostatic potential function generated by a molecule *A* is calculated as a summation over its atomic contributions:

$$V_A(\vec{r}) = \sum_{a \in A} \frac{Z_a}{\left|\vec{r} - \vec{R}_a\right|} \tag{8}$$

A smoothed version can be expressed as:

$$V_{A,t}(\vec{r}) = \sum_{a \in A} \frac{Z_a}{\left|\vec{r} - \vec{R}_a\right|} erf\left(\frac{\left|\vec{r} - \vec{R}_a\right|}{2\sqrt{t}}\right) \tag{9}$$

where the error function *erf* can be calculated using the analytically derivable expression [30]:

$$erf(x) = 1 - (a_1 t + a_2 t^2 + a_3 t^3 + a_4 t^4 + a_5 t^5) e^{-x^2}, \text{ with } t = \frac{1}{1 + px} \tag{10}$$

The values of the parameters p and a are: p = 0.3275911, a_1 = 0.254829595, a_2 = -0.284496736, a_3 = 1.421413741, a_4 = -1.453152027, and a_5 = 1.061405429, as reported in [30]. Equation (9) is identical to the expression found in potential smoothing approach, a well-known technique used in Molecular Mechanics (MM) applications [31].

Calculation of Fragment Charges

Fragment charges can, *a priori*, be calculated by summing over the point charges of the atoms *a* leading to a given fragment *F* in an ED or MEP field. This approach was, for example, initially applied for the evaluation of charges in proteins [20]:

$$q_F = \sum_{a \in F} q_a \tag{11}$$

As illustrated further in the text, the charges obtained in this way differ strongly from the values obtained using a charge fitting program. That last option was thus selected, and applied

through the program QFIT [32] to get fragment charges fitted from a MEP grid. In a conventional fitting procedure, grid points that are located too close or too far from the molecular structure under consideration are excluded from the calculation. The atomic van der Waals (vdW) radii are often the reference property to select grid points under interest. However, when using smoothed MEPs, charges are located at a reduced number of positions that do not necessarily correspond to atomic positions. Therefore, the corresponding peak/pit radius, $v_{smoothed}$, was defined as follows. Let us consider a 3D spherical Gaussian function:

$$f(r) \approx e^{-ar^2} \tag{12}$$

and its smoothed version:

$$f(r,t) \approx e^{-\frac{a}{(1+4at)}r^2} \tag{13}$$

An identification of the 3D integral of expressions (12) and (13) with the volume of a sphere built on a vdW radius v, *i.e.*:

$$\int f(r)4\pi r^2 dr = \frac{4}{3}\pi v^3 \quad \text{and} \quad \int f(r,t)4\pi r^2 dr = \frac{4}{3}\pi v_{smoothed}^3 \tag{14}$$

leads to the two following equalities, respectively:

$$a = \frac{\pi^{1/3}}{\left(\frac{4}{3}\right)^{2/3} v^2} \tag{15}$$

with v set equal to 1.5 Å for peaks and pits in a MEP grid, and:

$$(v_{smoothed})^3 = \frac{3}{4}\frac{\pi^{1/2}(1+4at)^{3/2}}{a^{3/2}} = (1+4at)^{3/2} v^3 \tag{16}$$

For example, at t = 1.4 bohr^2, $v_{smoothed}$ is equal to 2.036 Å, a value that is representative of low radius values that were previously associated with protein peaks observed in ED maps generated at a medium crystallographic resolution level [33]. In the present work, all MEP grids were built using the Amber point charges as reported in Duan *et al*. [21], with a grid step of 0.5 Å. For both

unsmoothed and smoothed MEP grids, fittings were achieved by considering points located at distances between 1.4 and 2.0 times the vdW radius of the atoms and peaks/pits, respectively. These two limiting distance values were selected as in the Merz-Singh-Kollman scheme [34].

In all fittings presented, the magnitude of the molecular dipole moment was constrained to be equal to the corresponding all-atom Amber value. The quality of the fittings was evaluated by two root mean square deviation (*rmsd*) values, *rmsdV* determined between the MEP values obtained using the fitted charges and the reference MEP values, and *rmsdμ* evaluated between the dipolar value calculated from the fitted CG charges and the reference dipole moment of the molecular structure:

$$rmsd\mu = \sqrt{\sum_{i=x,y,z} \left(\mu_{ref} - \mu_{fit}\right)^2} \qquad (17)$$

All dipole moment components were calculated with the origin set to (0. 0. 0.).

Results and Discussion

This section is dedicated to the elaboration of protein CG models, either based on the local maxima observed in smoothed ED, or on the local maxima and minima observed in smoothed MEP functions. The two main steps of our strategy rely, first, on a CG description of the protein backbone, and then on the development of side chain CG models. Each stage involves the determination of CG locations and corresponding electrostatic point charges. The final part of the section focusses on the application of our CG model to a literature case, the 12-residue β-hairpin HP7 [10].

We have restricted our studies to several fully extended peptides made of 15 amino acids, *i.e.*, Gly_7-AA-Gly_7, with the following protonation states: Lys(+1), Arg(+1), His with protonated Nε (noted Hisε further in the text), Glu(-1), and Asp(-1). The particular choice of such peptide sequences was a compromise to ensure that (*i*) the backbone of the central AA residue can interfere with neighbors. It was indeed shown previously that molecular ED-based fragments, especially protein backbone fragments, encompass atoms from the nearest residues [20,29]; (*ii*) the interference between the central AA residue and the whole peptide structure is minimized. The

concept of "interference" is solely based on the CG description obtained for various secondary structures. For example, when a α-helix is considered rather than an extended β-strand structure, atoms from the peptide backbone may merge with the side chain of the central residue. It is thus extremely difficult to define a CG model that is specific to a selected residue. We will show that the MEP-based clustering results are actually highly dependent on the peptide conformation; (*iii*) the charge on the central residue Gly8 of Gly_{15} is nul. This effect might also be obtained by considering a periodic peptide, which, up to now, is not implemented yet. For each of the pentadecapeptide studied, end residues were not charged. At first, this may sound artificial, but the presence of a large negative or positive charge in the structure strongly affects the homogeneity of the CG distribution along the peptide chain. This will be illustrated later when studying pentadecapeptide with a central charged AA residue. As also shown later, an extended structure presents an homogeneous CG distribution of a protein backbone, a specificity expected for an easy derivation of a CG model that should hopefully be transferable to any protein structure knowing its atom coordinates.

To generate all pentadecapeptides studied in this work, the simulated annealing (SA) procedure implemented in the program SMMP05 [35] was applied with dihedrals Ω, Φ, Ψ, and χ constrained to pre-defined values. The default force field (FF) ECEPP/3 [36] and SA running parameters were selected. Each SA run consisted in a first 100-step equilibration Monte Carlo (MC) Metropolis stage carried out at 1000 K. Then the procedure was continued for 50000 MC Metropolis iterations until the final temperature, 100 K, was reached. The lowest potential energy structure generated during each run was kept.

The hierarchical decomposition of molecular structures from ED distribution functions was achieved at t values ranging from 0.0 to 3.0 $bohr^2$, with a step of 0.05 $bohr^2$. The initial value Δ_{init} was set equal to 10^{-4} $bohr^2$, and $grad_{lim}$ to 10^{-5} $e^-/bohr^4$. When working with MEP functions, the steepness of the MEP at the initial atom location led to the following choice of parameters: t = 0.05 to 3.0 $bohr^2$, $\Delta_{init} = 10^{-6}$ $bohr^2$, $grad_{lim} = 10^{-6}$ $e^-/bohr^2$. Computing times for pentadecapeptide Gly_{15} and 12-residue HP7, on a PC Xeon 32-bit processor with a clock frequency of 2.8 GHz, are presented in Table III.I.

Table III.I. Calculation times (min.) for the hierarchical merging/clustering decompositions of PASA-ED and all-atom Amber MEP functions of Gly_{15} and 12-residue hairpin HP7 (PDB code: 2EVQ).

	cpu time	
	ED	MEP
α-Gly_{15}	5	45
β-Gly_{15}	3	25
HP7	27	44

It is seen that cpu times obviously increase with the number of atoms in a molecular structure but also with its packing. As Coulomb interactions are long-ranged, packing however has a limited influence on the calculation time that is required for the analysis of MEP functions.

Protein Backbone Modeling

As announced hereabove, to maximize the interatomic distances between the backbone and side chain atoms, an extended geometry characterized by $\Omega = 180°$, $\Phi = -139°$, $\Psi = 135°$ was considered. Indeed, for MEP analyses, the conformation of the peptide appeared to be extremely important on the results of the merging/clustering algorithm applied to MEP functions. This is illustrated in Figures III.1 and III.2 that respectively depict the smoothed ED and MEP obtained at t = 1.4 bohr^2 for a β-strand and a α-helix of Gly_{15}. As already established before [20,29], the ED-based decomposition of the protein backbone is rather regular, consisting mainly in fragments $(C{=}O)_{AA}(N\text{-}C\alpha)_{AA+1}$.

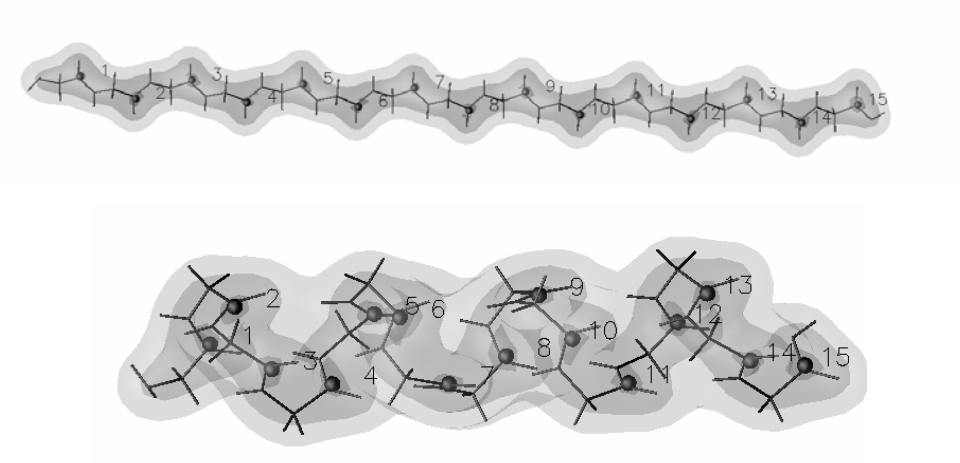


Figure III.1. ED iso-contours (0.05, 0.10, 0.15 e^-/bohr^3) of (top) ß-Gly_{15} and (bottom) α-Gly_{15} smoothed at t = 1.4 bohr^2. Local maxima at t = 1.4 bohr^2 were obtained using the hierarchical merging/clustering algorithm applied to the PASA ED distribution function. CG points are numbered as in Table III.III. Figures were generated using DataExplorer [47].

The dendrograms (Figure III.3) resulting from the application of our hierarchical merging/clustering algorithm shows that the ED-based merging of the atoms to form fragments first occurs between the H atoms and their chemically bonded neighbors at t = 0.05 bohr^2. Then, as already shown [20,29], the C and O atoms of the backbone carbonyl groups begin to merge starting at t = 0.4 bohr^2. From 0.65 to 0.9 bohr^2, the atoms of the AA backbones merge until regular fragment structures such as $(C{=}O)_{AA}(N\text{-}C\alpha)_{AA+1}$ (H atoms are not mentioned for clarity) are fully

created at about t = 1.25 bohr^2. At t = 1.4 bohr^2, there still exists one peak per residue, and an *rmsd* value of 0.216 Å is observed between the coordinates of the backbone peaks and their corresponding c.o.m. (Figure III.4). A difference between the ED peaks of the α- and β-structures does not appear before t = 2.45 bohr^2. At that smoothing level, the close packing of the residues that occurs in the helix structure leads to a faster reduction of the number of local ED maxima (Figure III.5). As just mentioned, at t = 1.4 bohr^2, one observes one ED peak per residue, regardless of the secondary structure (Figures III.1 and III.3).

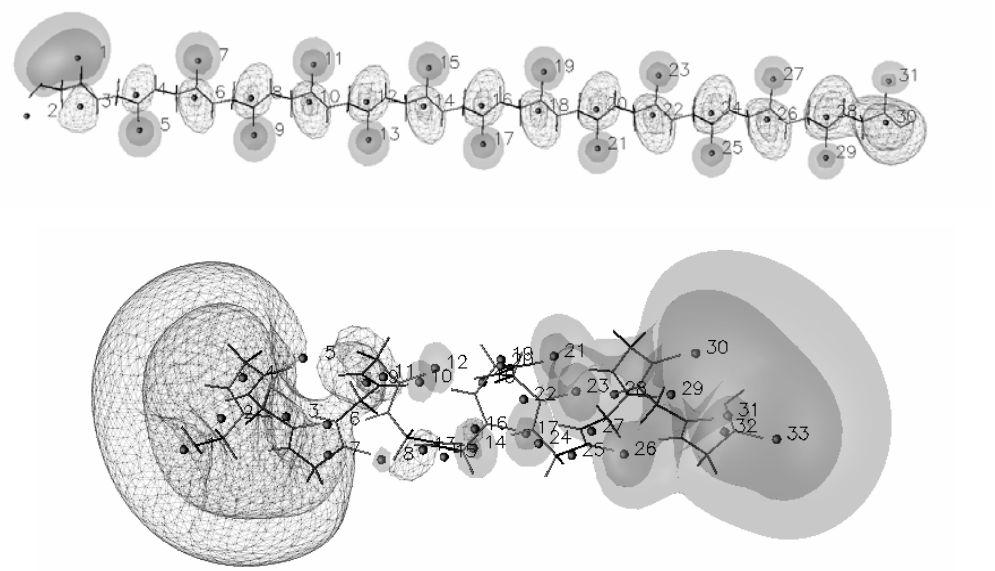


Figure III.2. MEP iso-contours (plain: -0.05, -0.03 ; grid: 0.03, 0.05 e$^-$/bohr) of (top) β-Gly$_{15}$ and (bottom) α-Gly$_{15}$ smoothed at t = 1.4 bohr^2. Local maxima and minima at t = 1.4 bohr^2 were obtained using the hierarchical merging/clustering algorithm applied to the all-atom Amber MEP function. CG points are numbered as in Table III.II. Figures were generated using DataExplorer [47].

When a MEP function is used, results differ from the ED-based ones, and are highly dependent on the backbone conformation. The dendrogram built from the results of the merging/clustering algorithm applied to the all-atom Amber MEP function illustrates that difference, and also shows that atoms are not necessarily merged according to their connectivity (Figure III.6). For example, at t = 1.4 bohr^2, a value selected because the number of peaks/pits does not vary significantly any longer beyond that smoothing degree, the points that are close to the O and C atoms (Figure III.2) are the result from the merge of the atoms (O, N, Cα) and (H, C, Hα, Hα), respectively. For an easier identification of those points, the corresponding closest atom in the molecular structure is given in Table III.II. In the case of β-Gly$_{15}$, one interestingly observes an alternating distribution of negative and positive charges around the C=O groups, while for α-Gly$_{15}$,

the dipolar character of the global structure is strongly emphasized with negative and positive charges being distributed at each end of the peptide, respectively (Figure III.2).

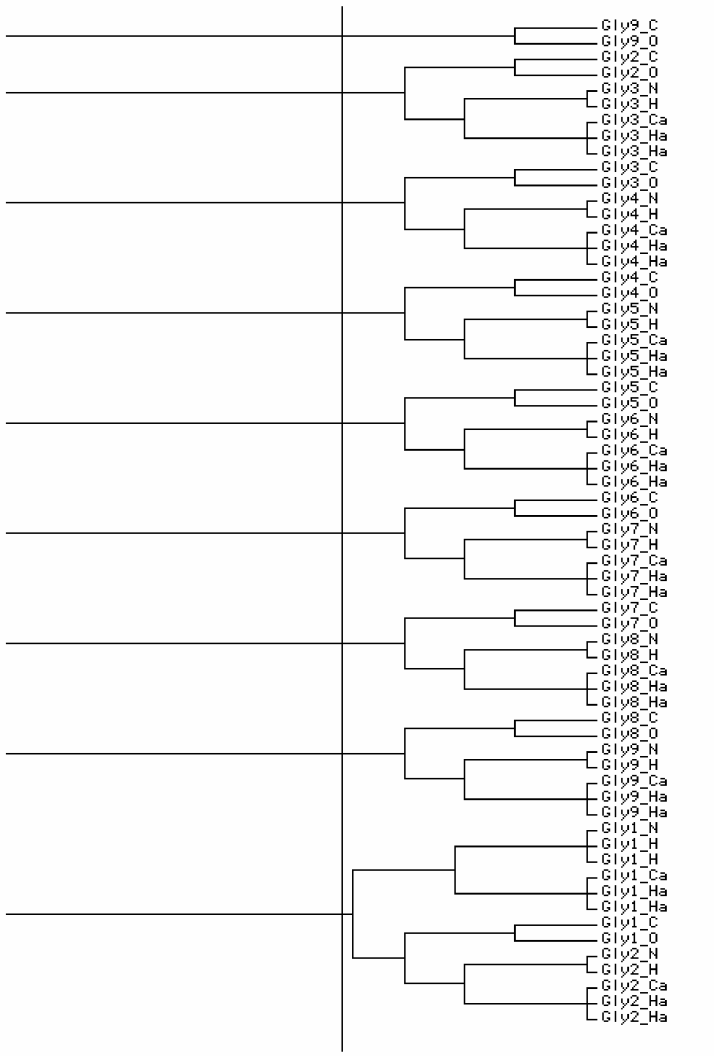


Figure III.3. Dendrogram depicting the results of the hierarchical merging/clustering algorithm applied to the PASA ED distribution function of ß-Gly_{15}. Results are displayed for the atoms of the first nine AA residues only. The vertical line locates t = 1.4 bohr^2.

Corresponding charge values, $q_{1.4}$, fitted from the MEP grids smoothed at t = 1.4 bohr^2, are presented in Table III.II. For β-Gly_{15}, the sign of the charges correspond to the expected dipolar distribution, *i.e.*, a positive and negative net charge close to the C and O atoms, respectively. For α-Gly_{15}, this expected charge distribution is observed only for residues 2, 4-7, and 15. It is thus hardly transferable from one residue to another. There are also additional charges that are close to the N atoms, with charge values being either positive (*e.g.*, point 15) or negative (*e.g.*, point 18).

Table III.II. CG charges $q_{1.4}$ and $q_{0.0}$ (in e^-) of Gly_{15} fitted from the all-atom Amber MEP grids smoothed at t = 1.4 and 0.0 bohr^2, respectively, using the program QFIT. Local maxima and minima at t = 1.4 bohr^2 were obtained using the hierarchical merging/clustering algorithm applied to the all-atom Amber MEP function. For each point, the distance *vs.* the closest atom, *d*, is given in Å. *rmsdV* and *rmsdμ* are given in kcal/mol and D, respectively. Point numbers (#) refer to Figure III.2.

	α-helix					β-strand				
#	Closest atom		*d*	$q_{1.4}$	$q_{0.0}$	Closest atom		*d*	$q_{1.4}$	$q_{0.0}$
1	N	Gly1	0.768	-0.042	-0.014	O	Gly1	0.599	-0.329	-0.311
2	H	Gly3	0.973	0.261	0.216	H	Gly1	1.212	0.011	0.026
3	O	Gly1	0.681	-0.002	-0.076	Hα	Gly1	1.120	0.218	0.189
4	C	Gly2	0.899	0.246	0.307	C	Gly2	0.797	0.261	0.256
5	O	Gly2	0.598	-0.139	-0.187	O	Gly2	0.605	-0.241	-0.236
6	O	Gly3	0.560	-0.024	-0.074	C	Gly3	0.840	0.201	0.214
7	C	Gly4	0.771	0.290	0.307	O	Gly3	0.624	-0.184	-0.205
8	O	Gly4	0.524	-0.325	-0.298	C	Gly4	0.825	0.169	0.193
9	C	Gly5	0.682	0.034	0.162	O	Gly4	0.620	-0.187	-0.204
10	O	Gly5	0.507	-0.038	-0.176	C	Gly5	0.832	0.202	0.209
11	C	Gly6	0.700	0.039	0.146	O	Gly5	0.623	-0.191	-0.202
12	O	Gly6	0.499	-0.027	-0.174	C	Gly6	0.827	0.195	0.202
13	C	Gly7	0.706	0.107	0.170	O	Gly6	0.622	-0.197	-0.204
14	O	Gly7	0.492	-0.088	-0.186	C	Gly7	0.830	0.199	0.206
15	N	Gly8	0.700	0.138	0.088	O	Gly7	0.623	-0.197	-0.205
16	C	Gly8	0.693	0.081	0.148	C	Gly8	0.828	0.197	0.205
17	O	Gly8	0.491	-0.110	-0.189	O	Gly8	0.623	-0.196	-0.205
18	N	Gly9	0.695	-0.024	0.030	C	Gly9	0.829	0.198	0.208
19	C	Gly9	0.698	0.048	0.131	O	Gly9	0.623	-0.196	-0.205
20	N	Gly10	0.691	0.005	0.066	C	Gly10	0.828	0.194	0.203
21	O	Gly9	0.499	-0.036	-0.138	O	Gly10	0.623	-0.195	-0.204
22	C	Gly10	0.716	-0.077	0.043	C	Gly11	0.829	0.204	0.210
23	O	Gly10	0.499	0.107	-0.001	O	Gly11	0.623	-0.194	-0.204
24	N	Gly11	0.675	0.005	0.039	C	Gly12	0.828	0.180	0.196
25	C	Gly11	0.747	-0.001	0.098	O	Gly12	0.623	-0.197	-0.206
26	O	Gly11	0.483	0.043	-0.050	C	Gly13	0.830	0.219	0.221
27	N	Gly12	0.678	-0.039	-0.019	O	Gly13	0.626	-0.194	-0.205
28	C	Gly12	0.824	0.186	0.209	C	Gly14	0.823	0.207	0.212
29	O	Gly12	0.581	-0.259	-0.252	O	Gly14	0.633	-0.211	-0.213
30	O	Gly13	0.619	-0.152	-0.144	C	Gly15	0.608	0.275	0.273
31	O	Gly14	0.619	-0.168	-0.159	O	Gly15	0.762	-0.205	-0.200
32	C	Gly15	0.587	0.242	0.242					
33	O	Gly15	0.648	-0.266	-0.250					
rmsdV				1.12	1.85				0.67	1.32
rmsdμ				0.46	0.11				0.19	0.34

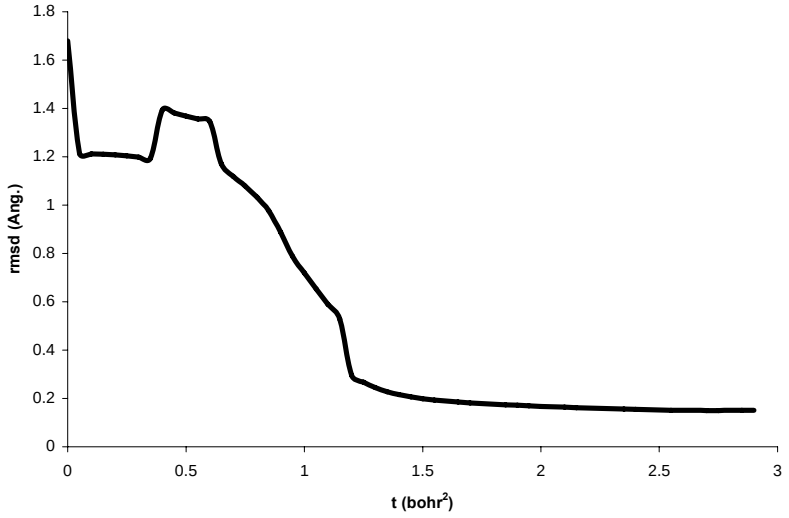


Figure III.4. *t*-dependent *rmsd* value calculated between the peaks observed in smoothed ED distribution functions of ß-Gly_{15} and their closest residue c.o.m. Local maxima were obtained using the hierarchical merging/clustering algorithm applied to the PASA ED distribution function.

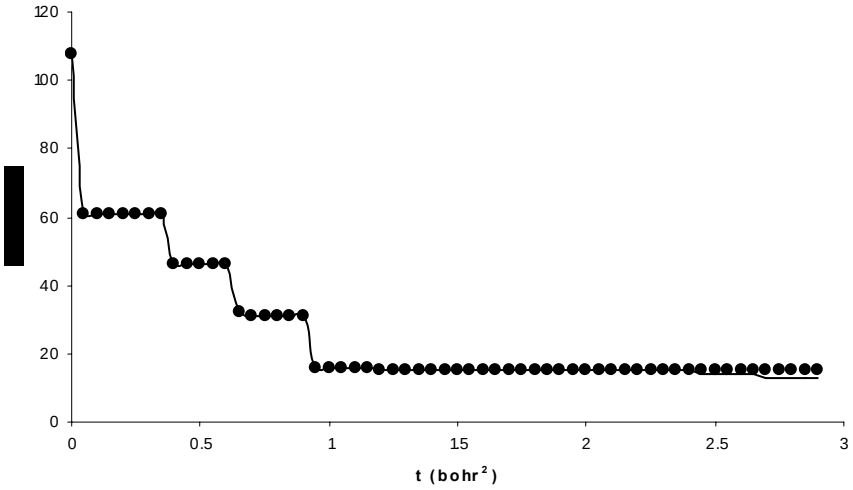


Figure III.5. *t*-dependent number of ED peaks observed for structures α-Gly_{15} (plain line) and ß-Gly_{15} (spheres). Local maxima were obtained using the hierarchical merging/clustering algorithm applied to the PASA ED distribution function.

The information about the closest atom is thus not strictly physically significant. The *rmsd* values reflect a rather good fitting result. For example, for the α-helix structure, μ(Amber) = (-28.694, -22.761, -26.473 D) and μ(fitted) calculated with $q_{1.4}$ charges = (-29.103, -22.834, -26.672 D); for the β-strand structure, μ(Amber) = (16.369, 6.365, 3.644 D) and μ(fitted) calculated with $q_{1.4}$ charges = (16.352, 6.542, 3.720 D). Except for the residues that are close to the peptide ends, it is

seen that for the extended structure, positive and negative charges are consistently located along the C=O axes, at distances of 0.83 and 0.62 Å from their closest atom.

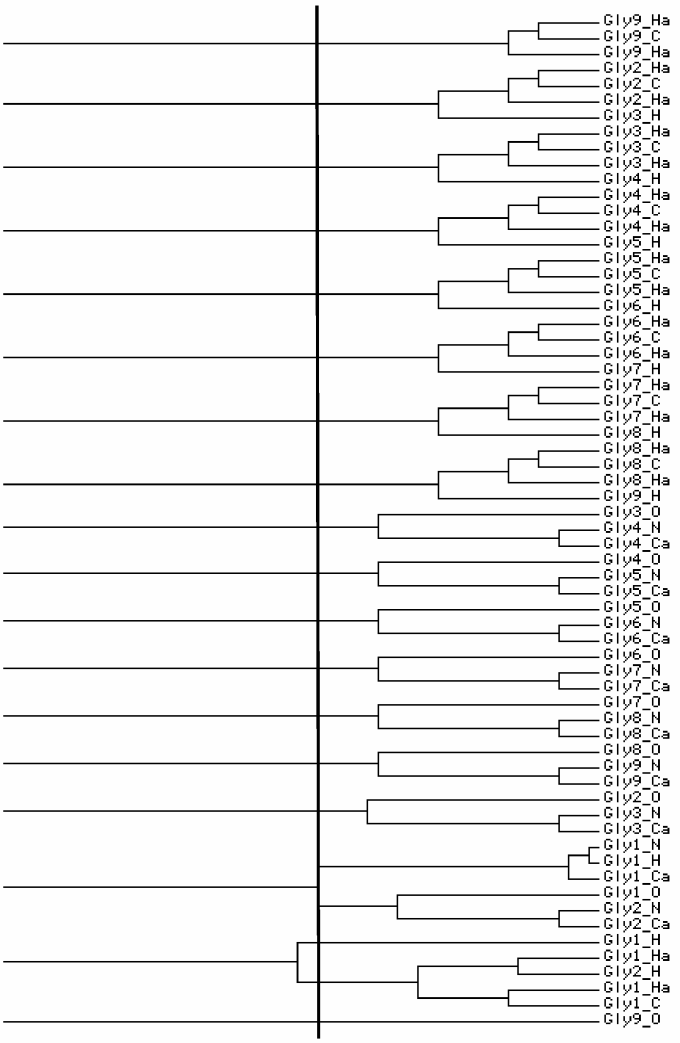


Figure III.6. Dendrogram depicting the results of the hierarchical merging/clustering algorithm applied to the all-atom Amber MEP function of β-Gly15. Results are displayed for the atoms of the first nine AA residues only. The vertical line locates t = 1.4 bohr^2.

There is a greater variability of these distances in the α-helix case. As expected, the charge values depend upon the position of the residue in the peptide sequence. This variability is largely more pronounced for the α-helix case. In the β-strand structure, the charges located on the central residue, Gly8, are close to $q_{1.4} = \pm 0.196$ e$^-$ (points 16 and 17), while there is no such local dipolar distribution in the α-helix structure. Rather, in this last case, the central glycine residue leads to 3 points (points 15-17). In α-Gly_{15}, the positive charges are predominant between points 1 and 7, while the negative charges are predominant from point 29 to 33 (Figure III.7).

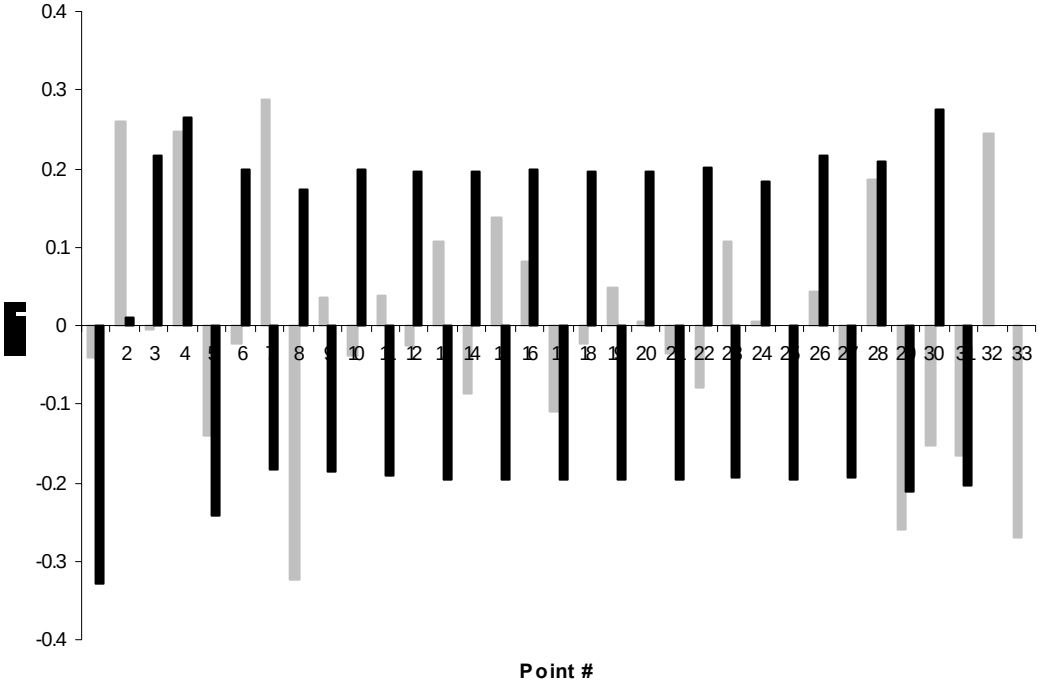


Figure III.7. CG charges $q_{1.4}$ fitted from an all-atom Amber MEP grid smoothed at t = 1.4 bohr^2 for α-Gly$_{15}$ (grey bars) and ß-Gly$_{15}$ (black bars). Local maxima and minima at t = 1.4 bohr^2 were obtained using the hierarchical merging/clustering algorithm applied to the all-atom Amber MEP function. Points are numbered as in Table III.II.

It is also clearly seen that the charge magnitude is largely reduced at the center of α-Gly$_{15}$, while the distribution is rather homogeneous for β-Gly$_{15}$. CG charges obtained though the fitting on the unsmoothed all-atom Amber MEP grids, $q_{0.0}$, are also presented in Table III.II. That approach is proposed to eliminate the effect of smoothing on the charge values. Indeed, this effect may be not considered in a conventional MM calculation. The fitting is obviously less efficient but is still of a reasonable quality, especially for the β-strand structure for which *rmsdV* is equal to 1.32 kcal/mol. The dipole moment calculated over the fitted CG charges $q_{0.0}$ is equal to (16.106, 6.557, 3.732 D) and leads to a *rmsdμ* value of 0.34 D. The dipolar character of each C=O pair is characterized by charges $q_{0.0}$ = ± 0.205 e$^-$ separated by a distance of 2.65 Å. For α-Gly$_{15}$, the fitting is slightly less convincing, with *rmsdV* = 1.85 kcal/mol, while the rather good dipole moment of (-28.651, -22.796, -26.383 D) leads to *rmsdμ* = 0.11 D.

For comparison purposes, we report, in Table III.III, the charges q_F obtained using Equation (11) applied to the fragments observed in the PASA ED distribution function smoothed at t = 1.4 bohr^2. For an easier identification, each fragment is characterized by its closest residue c.o.m. except for the last peak, described with respect to the terminal oxygen atom OXT of Gly15.

Table III.III. CG charges q_F (in e^-) of Gly_{15} obtained using Equation (11) applied to fragments determined at t = 1.4 $bohr^2$ using a hierarchical merging/clustering algorithm applied to the PASA ED distribution function. Charges $q_{1.4}$ and $q_{0.0}$ were obtained through a charge fitting algorithm using all-atom Amber MEP grids smoothed at t = 1.4 and 0.0 $bohr^2$, respectively. For each point, the distance *vs*. the closest c.o.m., *d*, is given in Å. *rmsdV* and *rmsdμ* are given in kcal/mol and D, respectively. Point numbers (#) refer to Figure III.1.

	α-helix					β-strand				
#	c.o.m.	*d*	q_F	$q_{1.4}$	$q_{0.0}$	c.o.m.	*d*	q_F	$q_{1.4}$	$q_{0.0}$
1	Gly1	0.253	-0.070	0.195	0.214	Gly1	0.214	-0.070	-0.184	-0.174
2	Gly2	0.239	0.001	0.195	0.172	Gly2	0.216	0.001	0.213	0.182
3	Gly3	0.238	0.001	0.285	0.276	Gly3	0.216	0.001	-0.135	-0.105
4	Gly4	0.239	0.001	-0.043	-0.053	Gly4	0.216	0.001	0.043	0.036
5	Gly5	0.239	0.001	-0.123	-0.113	Gly5	0.215	0.001	0.003	-0.001
6	Gly6	0.239	0.001	-0.108	-0.096	Gly6	0.216	0.001	0.001	0.006
7	Gly7	0.238	0.001	-0.045	-0.041	Gly7	0.216	0.001	-0.008	-0.008
8	Gly8	0.239	0.001	-0.013	-0.012	Gly8	0.216	0.001	0.017	0.014
9	Gly9	0.239	0.001	0.044	0.034	Gly9	0.216	0.001	-0.011	-0.009
10	Gly10	0.240	0.001	0.073	0.076	Gly10	0.216	0.001	0.010	0.011
11	Gly11	0.239	0.001	0.156	0.153	Gly11	0.215	0.001	-0.009	-0.009
12	Gly12	0.237	0.001	0.069	0.051	Gly12	0.216	0.001	0.034	0.027
13	Gly13	0.238	0.001	-0.235	-0.219	Gly13	0.216	0.001	-0.064	-0.051
14	Gly14	0.236	0.001	-0.146	-0.165	Gly14	0.216	0.001	0.084	0.073
15	OXT15	1.154	0.072	-0.289	-0.264	OXT15	1.234	0.072	0.020	0.023
rmsdV			17.57* 19.17**	3.59	3.97			4.43* 5.24**	4.29	5.10
rmsdμ			53.30	1.18	1.70			4.55	2.09	2.37

*Fitting achieved *vs*. all-atom Amber MEP grid smoothed at t = 1.4 $bohr^2$.

**Fitting achieved *vs*. unsmoothed all-atom Amber MEP grid.

It is first seen that both α- and β-structures lead to identical decomposition results. As each peak consists in a glycine residue, its total charge q_F is zero (the exact value results from the atom charges reported in [21]), except for the end points which contain only a partial number of glycine atoms. We further considered those ED-based peaks as a CG model for the pentadecapeptide, and we evaluated the CG charges, $q_{0.0}$ and $q_{1.4}$, from all-atom Amber MEP grids generated at t = 0.0 and 1.4 $bohr^2$, respectively (Table III.III). The charges q_F determined from Equation (11) do not lead to a fitting as nice as those obtained from the MEP-based CG representations. This is especially true for the α-helix case, for which *rmsdV* values are equal to 17.57 and 19.17 kcal/mol *vs.* the all-atom Amber MEP grids smoothed at t = 1.4 and 0.0 $bohr^2$, respectively, and *rmsdμ* = 53.30 D.

Protein Side Chains Modeling

Several CG representations of AA side chains were obtained by substituting the central residue Gly8 of β-Gly_{15} by a selected AA in a specific conformational state. Except for AA = Gly and Ala,

a number of rotamers were generated by considering the angular constraints given in Table III.IV. These rotamers were selected according to their occurrence degree in protein structures as reported in the *Structural Library of Intrinsic Residue Propensities* (SLIRP) [37].

Table III.IV. Geometrical parameters and occurrence probability of the selected AA side chain rotamers. *g* and *t* stand for *gauche* and *trans*, respectively (see [37] for details).

	Conformation	χ1 (°)	χ2 (°)	χ3 (°)	χ4 (°)	Occurrence (%)
Arg	*g-, t, g-, g-*	300	180	300	300	9.5
	g-, t, g-, t	300	180	300	180	11.9
	g-, t, g+, t	300	180	60	180	12.2
	g-, t, t, t	300	180	180	180	12.2
Asn	*t, Nt*	180	0			11.1
	t, Og-	180	300			21.3
	t, Og+	180	60			23.6
Asp	*t, g+*	180	60			62.8
Cys	*g-*	300				56.3
	g+	60				15.1
	t	180				28.7
Gln	*g-, t, Nt*	300	180	0		11.2
	g-, t, Og-	300	180	300		33.2
	g-, t, Og+	300	180	60		28.6
Glu	*g-, t, g-*	300	180	120		29.9
	g-, t, g+	300	180	60		25.3
His	*g-, Ng-*	300	300			35.8
	t, Ng+	180	60			15.0
Ile	*g-, g-*	300	300			22.7
	g-, t	300	180			28.3
	g+, t	60	180			42.5
Leu	*g-, t*	300	180			65.2
	t, g+	180	60			24.1
Lys	*g-, g-, t, g-*	300	300	180	300	8.5
	g-, g-, t, g+	300	300	180	60	6.5
	g-, t, t, g-	300	180	180	300	21.7
	g-, t, t, g+	300	180	180	60	14.3
Met	*g-, g-, g-*	300	300	300		15.5
	g-, g-, t	300	300	180		11.6
	g-, t, g-	300	180	300		19.4
	g-, t, g+	300	180	60		16.4
	g-, t, t	300	180	180		15.4
Phe	*g-, g-*	300	300			37.8
	t, g+	180	60			31.5
Pro	*g+*	0				66.8
Ser	*g-*	300				73.1
	g+	30				24.8
Thr	*g-*	300				51.6
	g+	30				46.3
Trp	*g-, g-*	300	90			28.2
	g-, t	300	0			16.5
	t, g-	180	60			11.6
	t, g+	180	300			13.8
	t, t	180	0			11.2

Tyr	*g-, g-*	300	120	38.3
	t, g+	180	60	31.7
Val	*g-*	300		46.4
	t	180		51.9

As already specified above, we considered the following protonation states: Lys(+1), Arg(+1), Hisε, Glu(-1), and Asp(-1). In Figures III.8 to III.10, we present the details of the MEP-based CG representations of Asn, Arg(+1), and Glu(-1) obtained from the all-atom Amber MEP function [21], smoothed at t = 1.4 bohr^2. The case of Asn (Figure III.8) illustrates that, for a neutral residue, the number of minima and maxima may depend upon the conformation.

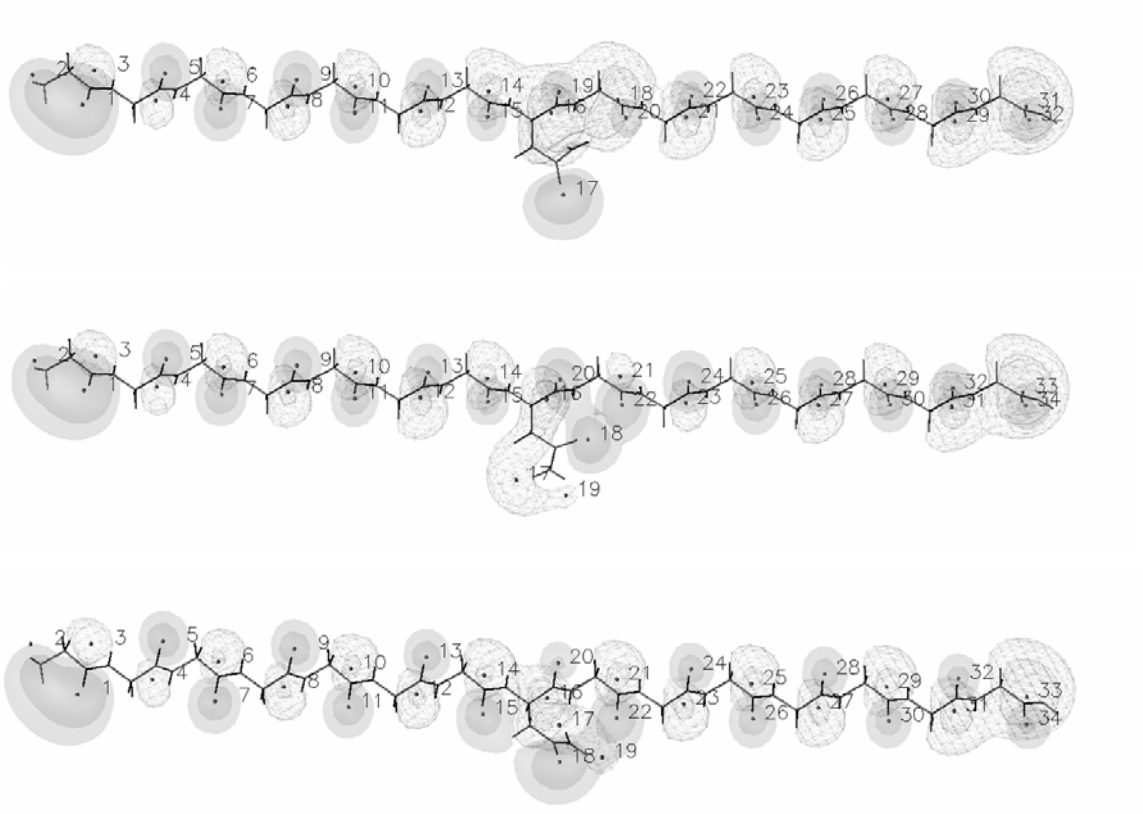


Figure III.8. MEP iso-contours (plain: -0.05, -0.03 ; grid: 0.03, 0.05 e$^-$/bohr) of Gly_7-Asn-Gly_7 smoothed at t = 1.4 bohr^2. Top: *t,Nt*, middle: *t,Og-*, bottom: *t,Og+* conformation. Local maxima and minima at t = 1.4 bohr^2 were obtained using the hierarchical merging/clustering algorithm applied to the all-atom Amber MEP function. Figures were generated using DataExplorer [47].

For conformation *t,Nt*, there is only one negative charge located at the proximity of the O atom (point 17), while for the two other selected conformations, *t,Og-* and *t,Og+*, there are two additional positive charges close to the amide H atoms (points 17 and 19). For *t,Nt*, the two H atoms of the amide group are sufficiently close to the peptide backbone to be merged in the two positively charged fragments 16 and 18. In the case of Arg(+1), all four conformers showed the same

characteristics, *i.e.*, three positive charges located at the neighborhood of the guanidinium group (points 17-19). The *g-,t,g-,g-* conformation is illustrated in Figure III.9. It is well seen that the positive charge located on the side chain strongly affects the distribution of peaks and pits at the level of the whole peptide backbone. Concerning Glu(-1), illustrated in Figure III.10 for the *g-,t,g-* conformation, all rotamers studied also showed a similar CG description with two negative charges facing the carboxylate O atoms (points 14 and 15). The peptide backbone CG representation is also strongly affected by the global negative charge of the residue side chain. The global influence of charged groups on the CG model of the pentadecapeptide models justifies, as mentioned earlier, the choice of studying peptides without charged end residues. In a further step, we determined the charge values for the CG descriptions of each AA through a fitting procedure carried out using QFIT [32] *vs.* unsmoothed MEP grids. For each of the AAs, all rotamer descriptions in terms of peaks and pits observed in all-atom Amber MEP smoothed at t = 1.4 bohr^2 were considered according to their occurrence probability (Table III.IV). The peptide backbone was constrained to be modeled by a sequence of alternating negative and positive charges, $q_{0.0}$, as previously determined for β-Gly$_{15}$ (Table III.II), and that, even for charged residues (Arg, Lys, Glu, Asp). The exception is for the central residue under consideration, for which all charges, even the backbone ones, were free to vary during the fitting procedure, under two constraints: the molecular all-atom Amber charge and the corresponding total dipole moment.

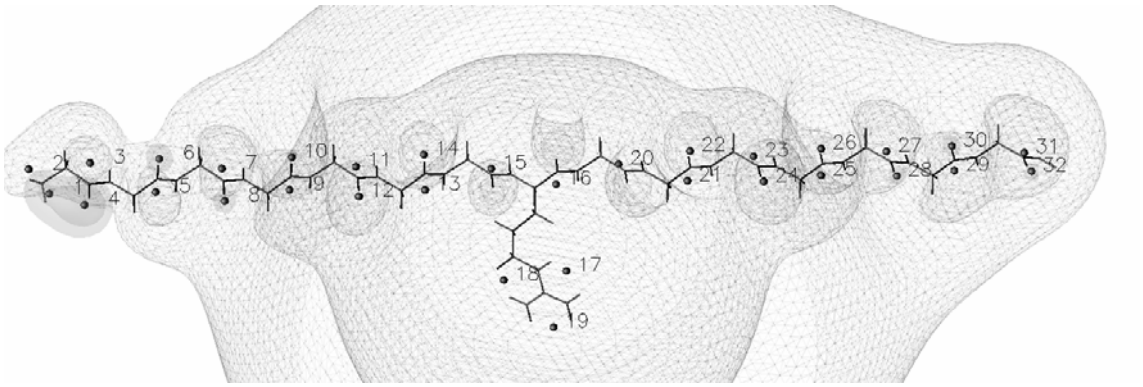


Figure III.9. MEP iso-contours (plain: -0.05, -0.03 ; grid: 0.03, 0.05 e$^-$/bohr) of Gly$_7$-Arg-Gly$_7$ in its *g-,t,g-,g-* conformation, smoothed at t = 1.4 bohr^2. Local maxima and minima at t = 1.4 bohr^2 were obtained using the hierarchical merging/clustering algorithm applied to the all-atom Amber MEP function. Figure was generated using DataExplorer [47].

It is to be specified that, for some AA residues, the initial MEP-based peak/pit CG representation obtained for the corresponding side chain was replaced by a simpler model consisting of one of several points centered on selected atoms. This was achieved as a first stage in

the easy design of a CG protein model from its atom coordinates, *e.g.*, coordinates retrieved from the PDB [38].

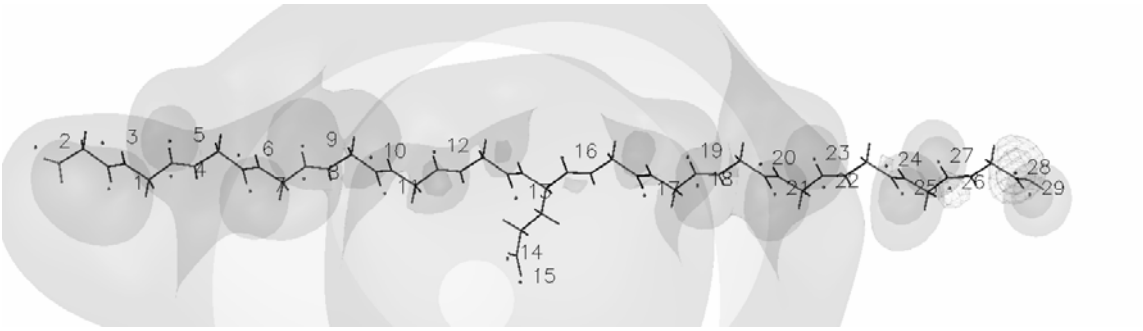


Figure III.10. MEP iso-contours (plain: -0.05, -0.03 ; grid: 0.03, 0.05 e^-/bohr) of Gly_7-Glu-Gly_7 in its *g-,t,g-* conformation, smoothed at t = 1.4 $bohr^2$. Local maxima and minima at t = 1.4 $bohr^2$ were obtained using the hierarchical merging/clustering algorithm applied to the all-atom Amber MEP function. Figure was generated using DataExplorer [47].

In Figure III.11, we report the so-obtained original or simplified CG representations for all 20 AA residues as derived from the results of our hierarchical merging/clustering algorithm applied to the all-atom Amber MEP function, smoothed at t = 1.4 $bohr^2$. Corresponding CG charges are reported in Table III.V. For all non-cyclic C-H based residues, *i.e.*, Ala, Ile, Leu, and Val, the side chain points were placed exactly on C atoms. This was chosen as an easy way to model the side chain of those specific residues in MM applications. For example, in the Ala case, the point charge that was initially observed in the all-atom Amber MEP function smoothed at t = 1.4 $bohr^2$ at a distance of 0.587 Å from Cβ (Table III.VI) is replaced by a sphere centered exactly on atom Cβ. For the specific case of Ile, the number of peaks and pits that was initially observed in the smoothed MEP functions, depends on the conformation of the Ile side chain. More precisely, there is only one CG observed in each of the three rotamers, which is close to atom Cβ for conformations *g-,g-* and *g-,t* and close to Cδ1 for conformation *g+,t*. We thus evaluated two different models composed either of three or four points. In the case of the three-point model, the side chain CG point is located either on atom Cβ or atom Cδ1; in the case of the four-point model, the two CG points of the side chain are centered on the Cβ and Cδ1 atoms. Resulting *rmsdV* and *rmsdμ* values presented in Table III.VI show that the two models perform similarly in approximating the unsmoothed all-atom Amber MEP function, due to the very low charge values associated with the side chain CG points.

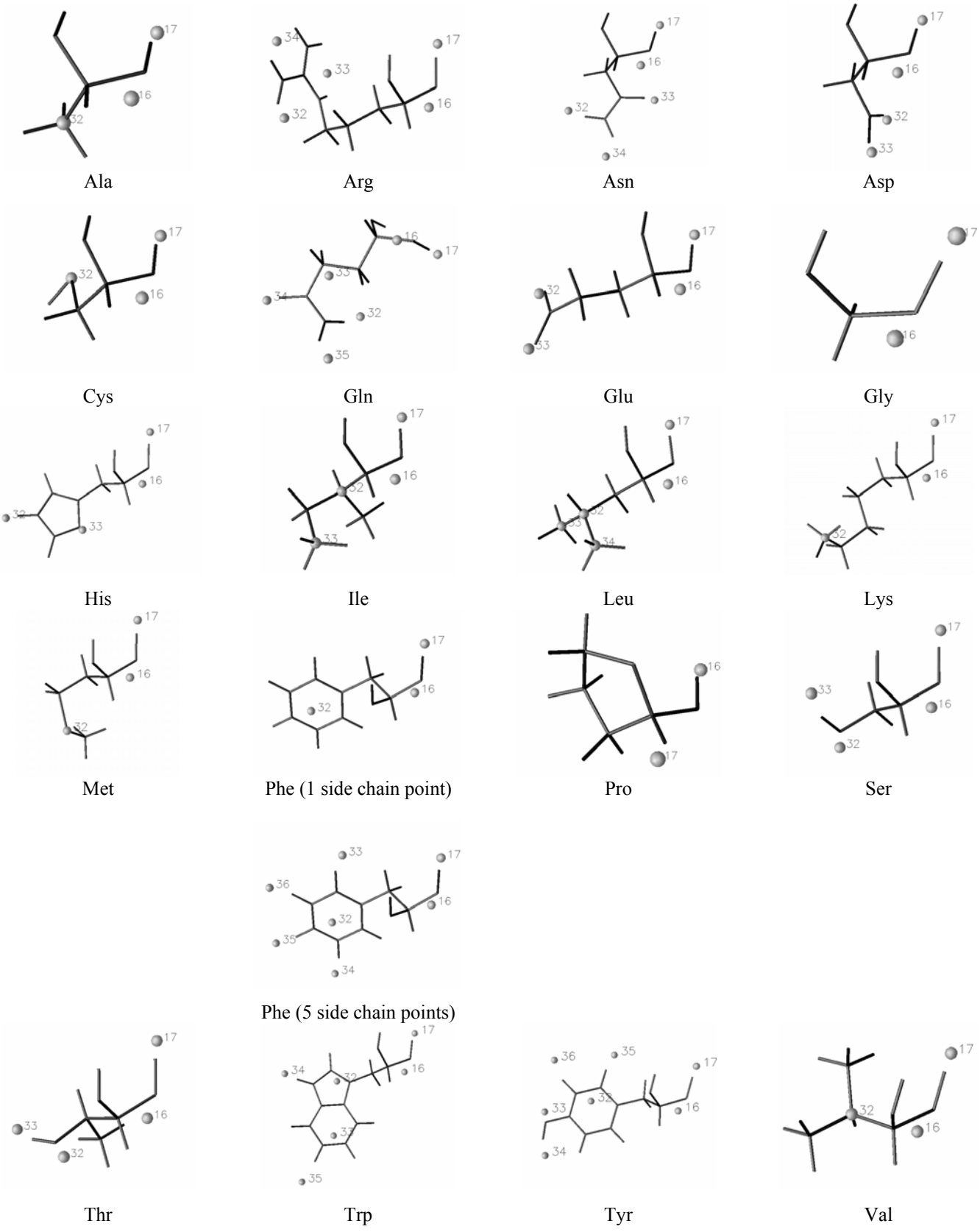


Figure III.11. CG model for each of the 20 AA residues as established at t = 1.4 bohr^2 from the hierarchical merging/clustering algorithm applied to the all-atom Amber MEP function. CG points are numbered as in Table III.V. Figures were generated using DataExplorer [47].

Similarly to the non-cyclic C-H based residues, the side chain of sulfur-containing residues, *i.e.*, Cys and Met, was modeled by a sphere placed exactly on the S atom. In the case of Cys in its *g*+ conformation, there was initially an extra charge located at 1.038 Å of the H atom, which was, however, not observed in the *g*- conformation. As the charge fitting of the single structure *g*+ onto the unsmoothed all-atom Amber MEP grid led to q_H = 0.077 e^- (other output parameters were q_O = -0.275, q_C = 0.360, q_S = -0.163 e^-, *rmsdV* = 1.85 kcal/mol, *rmsdμ* = 0.44 D), we neglected the point charge q_H to generate a unique model valid for all three Cys rotamers. For Met, most of the original side chain points were very close to the S atom, below 0.28 Å (Table III.VI), and a unique model with a sphere placed on S was built. For Lys, we also simplified the model by setting the positive charge exactly on the Nς atom. For all the other AAs, the original point locations observed in the smoothed MEP functions were kept for the charge fitting procedures. In the case of Phe, two models were evaluated (Figure III.11 and Table III.V). The first one was built from the set of CG points observed in the corresponding all-atom Amber MEP function, at t = 1.4 $bohr^2$. That model includes a point for the six-membered ring and four additional charges located close to the H atoms. This model does not reveal to be worse or better than the second one tested, consisting in a single ring point only. This is due to the very small amplitude of the H-related charges, ranging between 0.02 and 0.05 e^-. Indeed, the *rmsdV* values obtained for the 1- and 5-point side chain models are close to 1.5 and 1.4 kcal/mol, respectively. For *rmsdμ*, values are 0.1 and 0.3 D, respectively. On the whole, we can also note that for hydroxyl containing residues, *i.e.*, Ser, Thr, and Tyr, there are two charges located near, but not exactly on, the O and H atoms. For the negatively charged residues, *i.e.*, Asp and Glu, each carboxylate functional group leads to two negative charges located near the O atoms. On the contrary, positively charged residues, Arg and Lys, present different behaviors. While the side chain of Lys leads to only one positive charge value, the Arg side chain is characterized by a 3-point motif, wherein each charge is almost symmetrically located on bisectors of each of the three N-C-N angles of the guanidinium group. For all these residues, distances between the CG charge and their closest atom in the molecular structure are reported in Table III.VI.

For comparison with the MEP-based CG representations, the same exercise was achieved using ED-based CG representations that were built from the peaks observed in PASA ED distribution functions, smoothed at t = 1.4 $bohr^2$ (Figure III.12). Associated charges, calculated using Equation (11), are reported in Table III.VII. First of all, it is observed that, for a given AA, all rotamers showed the same behaviour, *i.e.,* identical hierarchical decompositions and fragment contents. A detailed description of the side chain fragments is presented in Table III.VIII. For Ala,

Gly, Ile, Pro, and Val, there is no side chain peak observed. All side chain atoms have actually been merged with backbone atoms to form a fragment whose corresponding ED maximum is closer to a backbone c.o.m. For example, atoms N, Cα, and Cβ of Ala were merged with C and O of the preceding AA residue in the peptide sequence. The same occurred for Ile and Val, where a backbone fragment was formed by $(CO)_{Gly7}$ and $(N\text{-}C\alpha\text{-}C\beta\text{-}C\gamma1\text{-}C\gamma2\text{-}C\delta1)_{Ile8}$, and $(CO)_{Gly7}$ and $(N\text{-}C\alpha\text{-}C\beta\text{-}C\gamma1\text{-}C\gamma2)_{Val8}$, respectively. The backbone peak of Pro was actually associated with atoms $(N\text{-}C\alpha\text{-}C\beta\text{-}C\gamma\text{-}C\delta)_{Pro8}$. Except for the AA under consideration and its nearest neighbor, *AA-1*, the CG model is not dependent upon the AA type, and the Gly charge remains equal to 0.001 e^- (Table III.VII), the value corresponding to the total charge of a Gly residue as reported in [21]. It can also be seen that the other nearest neighbor, *AA+1*, stays unaffected by the AA type. This ED effect is thus highly local, and might be qualified as a 'shape' effect, while the electrostatic long-range influence, that is present in MEP-based results, needs to be controlled using charge constraints during the fitting procedure. To eventually evaluate the quality of charges associated with ED-based CGs in reproducing the all-atom unsmoothed Amber MEP maps, *rmsdV* and *rmsdμ* values were calculated. They are reported in Table III.VII as well, and reflect the less precise reproduction of MEP and dipole values than the MEP-based CG charges.

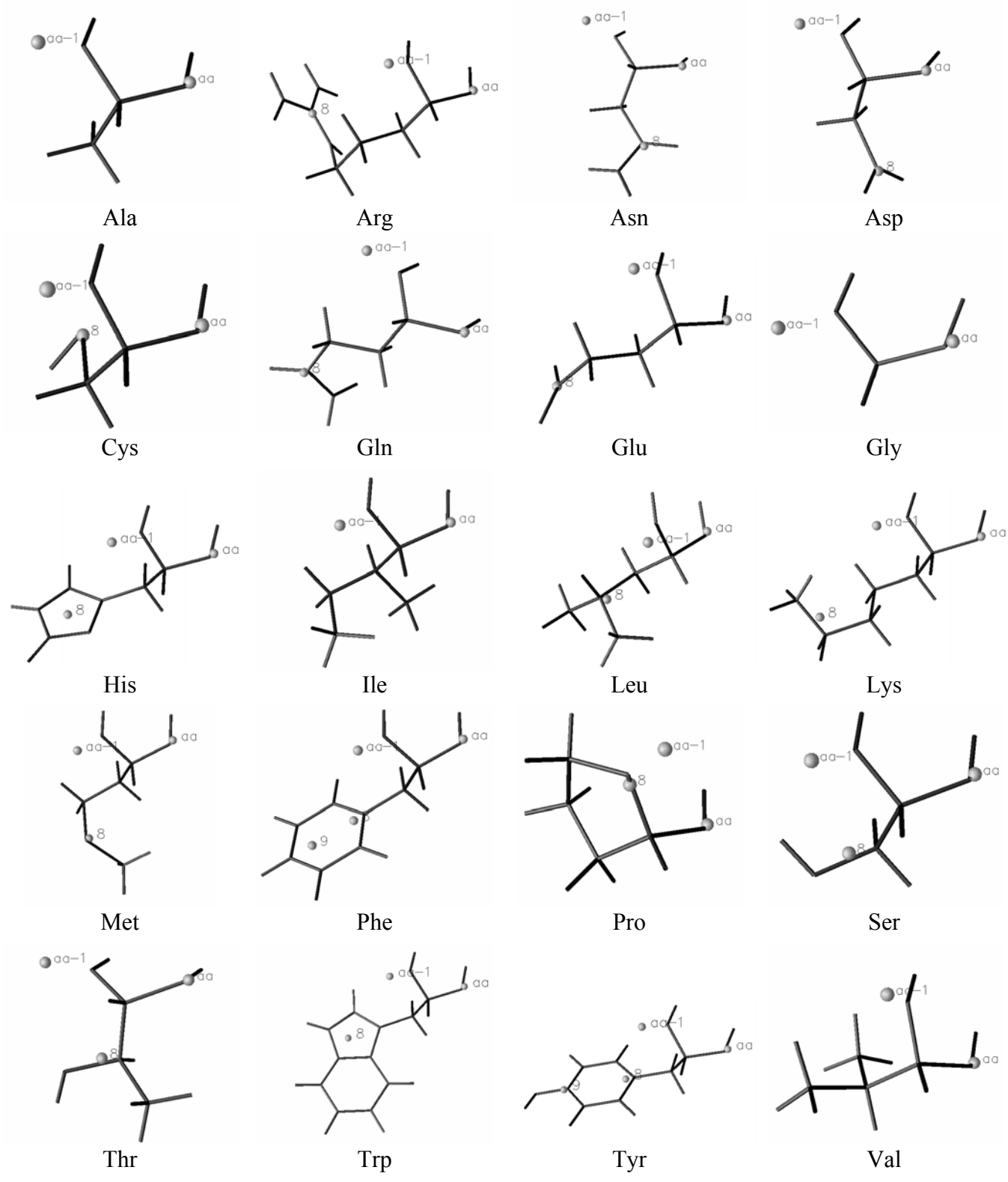


Figure III.12. CG model for each of the 20 AA residues as established at t = 1.4 bohr^2 from the hierarchical merging/clustering algorithm applied to the PASA ED distribution function. CG points are numbered as in Table III.VII. Figures were generated using DataExplorer [47].

Table III.V. CG charges (in e^-) for the AA residues obtained through a charge fitting algorithm using unsmoothed all-atom Amber MEP grids. CG locations were generated at *t* = 1.4 $bohr^2$ using a hierarchical merging/clustering algorithm applied to the all-atom Amber MEP function. *g* and *t* stand for *gauche* and *trans*, respectively (see [37] for details). *rmsdV* and *rmsdμ* are given in kcal/mol and D, respectively. Point numbers refer to Figure III.11.

	Conformation	Point 16	Point 17	Point 32	Point 33	Point 34	Point 35	Point 36	*rmsdV*	*rmsdμ*
Ala		C	O	Cβ						
		0.234	-0.236	-0.000(4)					1.77	0.66
Arg		C	O	$NH-NH_2$	NH_2-NH	NH_2-NH_2				
	g-,t,g-,g-								1.96	0.99
	g-,t,g-,t								1.78	0.85
	g-,t,g+,t	0.310	-0.199	0.281	0.312	0.284			1.72	0.21
	g-,t,t,t								1.67	0.30
Asn		C	O	$H\delta_{tr}$	Oδ	$H\delta_{cis}$				
	t,Nt								2.02	4.81
	t, Og-	0.371	-0.212	0.019	-0.178	0.002			2.55	3.42
	t, Og+								2.41	2.65
Asp		C	O	Oδ1	Oδ2					
	t,g+	-0.019	-0.205	-0.417	-0.358				1.58	0.64
Cys		C	O	Sγ						
	g-								1.64	0.77
	g+	0.398	-0.289	-0.103					1.92	1.46
	t								1.80	1.14
Gln		C	O	H_{tr}	Cγ	Oε	H_{cis}			
	g-,t,Nt								1.78	0.83
	g-,t,Og-	0.293	-0.289	0.167	0.001	-0.259	0.083		1.66	0.67
	g-,t,Og+								1.69	0.33
Glu		C	O	Oε1	Oε2					
	g-,t,g-	0.187	-0.270	-0.457	-0.456				1.51	0.29
	g-,t,g+								1.55	0.26
Gly		C	O							
		0.205	-0.205						1.32	0.34
His		C	O	Hε	Nδ					
	g-,Ng-	0.198	-0.188	0.178	-0.184				1.54	0.23
	t,Ng+								1.58	0.39
Ile		C	O	Cβ	Cδ1					
	g-,g-	0.266	-0.283	0.019					1.47	0.51
	g-,t	0.266	-0.283	0.019					1.46	0.42
	g+,t	0.266	-0.283		0.020				1.42	0.53
Ile		C	O	Cβ	Cδ1					
	g-,g-								1.44	0.49
	g-,t	0.226	-0.280	0.068	-0.012				1.44	0.51
	g+,t								1.41	0.61
Leu		C	O	Cγ	Cδ1	Cδ2				

	g-,t	0.219	-0.245	0.062	-0.030	-0.011			1.36	0.40
	t,g+								1.36	0.58
Lys		C	O	Nς						
	g-,g-,t,g-								1.68	0.95
	g-,g-,t,g+	0.367	-0.239	0.875					1.64	1.26
	g-,t,t,g-								1.63	1.25
	g-,t,t,g+								1.68	0.92
Met		C	O	Sδ						
	g-,g-,g-								1.89	1.87
	g-,g-,t								2.15	2.06
	g-,t,g-	0.283	-0.232	-0.059					1.79	1.46
	g-,t-g+								1.80	1.55
	g-,t,t								2.02	1.67
Phe		C	O	6-ring	Hδ2	Hε1	Hς	Hε2		
	g-,g-	0.263	-0.234	-0.163	0.033	0.047	0.031	0.027	1.40	0.34
	t,g+								1.40	0.30
Phe		C	O	6-ring						
	g-,g-	0.222	-0.219	-0.004					1.53	0.14
	t,g+								1.54	0.12
Pro		O	C							
	g+	-0.163	0.161						1.76	1.72
Ser		C	O	Oγ	Hγ					
	g-	0.304	-0.275	-0.173	0.153				1.50	0.45
	g+								1.65	0.62
Thr		C	O	Oγ	Hγ					
	g-	0.279	-0.247	-0.154	0.120				1.51	0.56
	g+								1.56	0.39
Trp		C	O	5-ring	6-ring	Hε1	HH			
	g-,g-								1.51	0.31
	g-,t								1.54	0.27
	t,g-	0.270	-0.210	-0.136	-0.098	0.146	0.028		1.58	0.37
	t,g+								1.54	0.43
	t,t								1.52	0.21
Tyr		C	O	6-ring	OH	HH	Hδ*	Hε*		
	g-,g-	0.267	-0.234	-0.110	-0.129	0.156	0.023	0.036	1.46	0.26
	t,g+								1.48	0.18
Val		C	O	Cβ						
	g-	0.092	-0.052	-0.051					1.62	0.66
	t								1.69	0.81

* Hδ and Hε stand on the opposite side of the O-H bond direction.

Table III.VI. Distances (in Å) observed between selected peaks and pits observed in all-atom Amber MEP function smoothed at t = 1.4 bohr^2, and their closest atom. '--' means that the peak/pit under consideration was not observed in the MEP grid of the considered rotamer.

	No. of rot.	Cβ	Cγ	Cδ1	Cδ2	Sγ (Cys), Sδ (Met), Nς (Lys)		
Ala	1	0.587						
Ile	3	0.846, 0.548, --		--, --, 0.418				
Leu	2		2.645, --	0.746, --	--, 0.796			
Val	2	0.355, 0.349						
Cys	3						0.358, 0.695, 0.450	
Met	5						0.201, 0.274, 0.098, 0.100, 0.105	
Lys	4						0.510, 0.380, 0.357, 0.408	
		O	H	Oδ1 (Asp), Oε1 (Glu)	Oδ2 (Asp), Oε2 (Glu)		Cς	
						Point 32	Point 33	Point 34
Ser	2	0.659, 0.796	1.100, 1.209					
Thr	2	0.873, 0.559	0.975, 0.932					
Tyr	2	0.634, 0.627	0.907, 0.910					
Asp	1			0.330	0.363			
Glu	2			0.342, 0.337	0.331, 0.342			
Arg	4					2.282, 2.256, 2.268, 2.215	1.902, 1.964, 1.989, 1.950	2.030, 2.036, 2.036, 2.047

Table III.VII. CG charges (in e^-) for the AA residues obtained from Equation (11) applied to fragments generated at t = 1.4 bohr^2 using a hierarchical merging/clustering algorithm applied to PASA ED distribution functions. *rmsdV* and *rmsdμ* are given in kcal/mol and D, respectively. They correspond to the mean value calculated per rotamer structure. 'BAK' stands for the backbone c.o.m. Point numbers refer to Figure III.12.

	BAK_{AA-2}	BAK_{AA-1}	BAK_{AA}	Point 8	Point 9	BAK_{AA+1}	*rmsdV*	*rmsdμ*
Ala	0.001	0.055	-0.058			0.001	5.19	4.27
Arg	0.001	0.020	0.081	0.899		0.001	18.60	14.68
Asn	0.001	0.002	0.022	-0.023		0.001	21.57	24.55
Asp	0.001	-0.077	-0.129	-0.793		0.001	5.70	6.78
Cys	0.001	0.038	-0.013	-0.023		0.001	30.68	21.62
Gln	0.001	0.205	-0.217	0.015		0.001	30.85	29.16
Glu	0.001	0.120	-0.194	-0.925		0.001	30.20	30.73
Gly	0.001	0.001	0.001			0.001	5.24	4.55
His	0.001	-0.100	0.062	0.039		0.001	28.98	29.95
Ile	0.001	0.123	-0.122			0.001	28.90	23.01
Leu	0.001	-0.023	-0.056	0.079		0.001	34.32	31.84
Lys	0.001	0.012	0.091	0.899		0.001	19.56	13.56
Met	0.001	0.110	-0.037	-0.073		0.001	20.87	18.31
Phe	0.001	0.030	-0.030	0.051	-0.051	0.001	29.73	24.40
Pro	0.001	0.072	-0.172	0.101		0.001	5.23	5.76
Ser	0.001	0.136	-0.169	0.033		0.001	38.75	38.52
Thr	0.001	-0.025	-0.063	0.090		0.001	37.38	34.14
Trp	0.001	0.050	0.018	-0.022		0.001	21.47	17.99
Tyr	0.001	-0.060	0.024	0.013	0.025	0.001	30.44	19.14
Val	0.001	0.029	-0.029			0.001	38.00	99.58

Table III.VIII. Atom content of the side chain ED-based fragments as obtained using a hierarchical merging/clustering algorithm, at t = 1.4 bohr^2. H atoms are not reported for clarity. Distances d between local ED maxima and closest side chain c.o.m. are given in Å.

	Fragment content	*d*
Arg	Cγ-Cδ-Nε-Cς-$(NH2)_2$	1.182
Asn	Cβ-Cγ-Oδ1-Nδ2	0.383
Asp	Cβ-Cγ-Oδ1-Oδ2	0.426
Cys	Cβ-Sγ	0.590
Gln	Cγ-Cδ-Oε1-Nε2	0.917
Glu	Cγ-Cδ-Oε1-Oε2	0.879
His	Cγ-Nδ1-Cε1-Nε2-Cδ2	0.618
Leu	Cγ-Cδ1-Cδ2	0.115
Lys	Cδ-Cε-Nς	1.187
Met	Cγ-Sδ-Cε	0.857
Phe	Cγ-Cδ1-Cδ2	0.483
	Cε1-Cε2-Cς	1.140
Ser	Cβ-Oγ	0.070
Thr	Cβ-Oγ1-Cγ2	0.474
Trp	Cγ-Cδ1-Nε1-Cε2-Cς2-CH2-Cς3-Cε3-Cδ2	1.035
Tyr	Cγ-Cδ1-Cδ2	1.213
	Cε1-Cς-OH-Cε2	1.436

Application to 12-Residue β-Hairpin HP7

The structure of peptide HP7 was retrieved from the PDB [38] (PDB code 2EVQ). The primary structure of that peptide is Lys-Thr-Trp-Asn-Pro-Ala-Thr-Gly-Lys-Trp-Thr-Glu (Figure III.13). It has a global net charge of +1.004 when summing over the atom charges given in reference [21]. The structure is interesting to consider as a reference structure because a fragment-based description, as well as corresponding point charges, have been provided [10]. In that representation, each pseudo-atom is defined as the geometric center of the heavy atoms of a protein fragment.

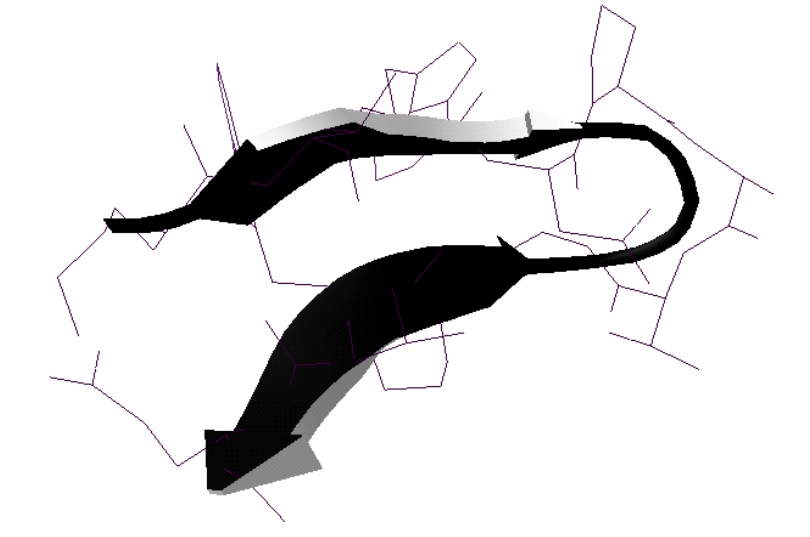


Figure III.13. 3D conformation and secondary structure of the 12-residue β-hairpin peptide HP7 (PDB code 2EVQ). Figure was generated using SwissPDBViewer [48].

The decompositions as obtained from PASA ED and all-atom Amber MEP functions smoothed at t = 1.4 bohr^2 are displayed in Figure III.14, together with the Basdevant's CG model which is composed of 28 grains. As already mentioned above, the MEP-based results are highly dependent on the conformation of the peptide, and a MEP-based CG description obtained at t = 1.4 bohr^2 now consists in only 22 points. This is well below the expected number of peaks and pits, *i.e.*, 44 as will be seen later, that would be obtained if all AA residues were considered as isolated. Figure III.14 illustrates the high diversity of the various CG models. In Table III.IX, we present the charges associated with the CG representations obtained from the application of the hierarchical merging/clustering algorithm to PASA ED distribution and all-atom Amber MEP functions, smoothed at t = 1.4 bohr^2, compared to the effective charges reported in the literature [10]. Our charges were obtained using the program QFIT [32] *vs.* unsmoothed all-atom Amber MEP grids. The major point to mention is the very bad approximation brought by the model built on MEP CG points whose charges were simply calculated using Equation (11). Indeed, *rmsdV* and *rmsdμ* values are equal to 33.04 kcal/mol and 43.13 D, respectively. The use of a simple approximation such as Equation (11) provides

better results when applied to ED-based fragments, with *rmsdV* = 12.78 kcal/mol and *rmsdμ* = 16.04 D. For that last model, we can also note that the charges obtained for the side chain peaks are identical to the values reported in Table III.VII for Gly7-AA-Gly7 structures. Thus, a change in the primary and secondary structures of a protein does seem to affect the backbone peaks only. When the charge fitting procedure is applied, both the 23-point ED- and 22-point MEP-based CG models provide similar quality approximations of the all-atom unsmoothed Amber MEP grid, with, respectively, *rmsdV* = 5.45 and 4.62 kcal/mol, and *rmsdμ* = 1.04 and 1.96 D.

In structure HP7, the two end residues are positively and negatively charged, respectively. These end charges prevent the regular carbon and oxygen CG MEP motif to appear on the residue backbones. Indeed, in Figure III.14 and Table III.IX, one can observe that these two point charges are missing for most of the residues, except for Trp3 (points 5 and 6 in Table III.IX). Thus, in order to build the backbone CG model of peptide HP7, two charges were generated for each residue backbone except for the first and the last ones. The two charges were located at distances of 0.828 and 0.623 Å, respectively from the C and O atoms, along the C=O axes. Each point was assigned a charge depending upon the AA residue type, as given in Table III.V. For the two Lys residues, a charge of 0.875 e^- was assigned to their Nς atom. For Ala, a charge of -0.000(4) e^- was set at the Cβ atom location. For the Thr, Trp, Asn, and Glu residues, a MEP-based hierarchical merging/clustering procedure was first carried out for each isolated residue, with coordinates as given in the PDB structure. This provided the location of the CG points, whose coordinates are reported in Table III.X. Then, charges were assigned to those points according to the values reported in Table III.V. This was achieved under the assumption of charge transferability between pentadecapeptide models and a protein structure. To strictly confirm this concept, a larger set of applications is however required. For Glu, a mean charge of -0.457 e^- was given to each of the CG points located close to the Oε atoms. The end charges located on N_{Lys1} and OXT_{Glu12} were calculated as a sum over a unit charge and the corresponding C and O charges of Lys1 and Glu12, respectively. For example, the charge located on N_{Lys1} was set equal to q = 1.127 e^- = +1.000 -0.239 + 0.367 e^-. Finally, it is recalled that there is no side chain CG point for Pro and Gly. There remained a 44-point CG model for the 12-residue peptide HP7 (Figure III.15), with a total charge of 0.999 e^-. For that particular model, and with respect to the unsmoothed all-atom Amber MEP grid, the calculated *rmsdV* and *rmsdμ* values are equal to 7.34 kcal/mol and 8.89 D (Table III.X). In comparison, the model proposed by Basdevant *et al.* [10] does not perform correctly (Table III.IX), with *rmsdV* = 37.74 kcal/mol and *rmsdμ* = 23.06 D; but this is most probably due to the use of a different set of atom charges, and a different parametrization of the charge fitting algorithm. An optimization of the Basdevant's model *vs*. our unsmoothed all-atom Amber MEP grid led to *rmsdV* = 5.45 kcal/mol and *rmsdμ* = 1.57 D (Table III.IX), while an

optimization of our 44-CG model *vs*. the same Amber MEP grid led to the charges reported in Table III.X, with *rmsdV* = 2.80 kcal/mol and *rmsdμ* = 1.11 D. The major changes brought to our model charges occurred at the level of the C atoms; indeed, the absolute differences between the model charges and their corresponding optimized values are higher than 0.30 e^- at residues 3, 7, and 9-11. Other drastic changes occurred, for example, at the level of $O\gamma_{Thr7}$, going from a charge value of -0.154 to 0.546 e^-, and for $H\gamma tr_{Asn4}$, with a charge difference of -0.31 e^-. There is also an important charge re-distribution between the two Oε atoms of Glu12. In comparison, larger charge differences are observed between the original Basdevant's model and the corresponding fitted charges; most of them, *i.e*., 17 over 28, are higher than 0.30 e^- in absolute value. Backbone CG are among the points that are characterized by the largest differences, *i.e*., Thr2, Asn4, Pro5, Lys9, Trp10, Thr11, and Glu12.

In conclusion, among the two models that can be easily built for HP7, *i.e*., ED-based CG with charges assigned using Equation (11) and MEP-based CG model as described in Table III.V, the last one is slightly better. It is however no doubt that an optimization of the charges would drastically improve the quality of the models, but this requires an additional step that can be time-consuming for large structures. In the present work, such an optimization stage was carried out on a single rigid conformation, while the initial (non optimized) model charges implicitly involved information relative to several AA conformations (but no information relative to various secondary structures). Our present feeling is that the use of Equation (11) in combination with MEP-based CG has to be rejected at this point.

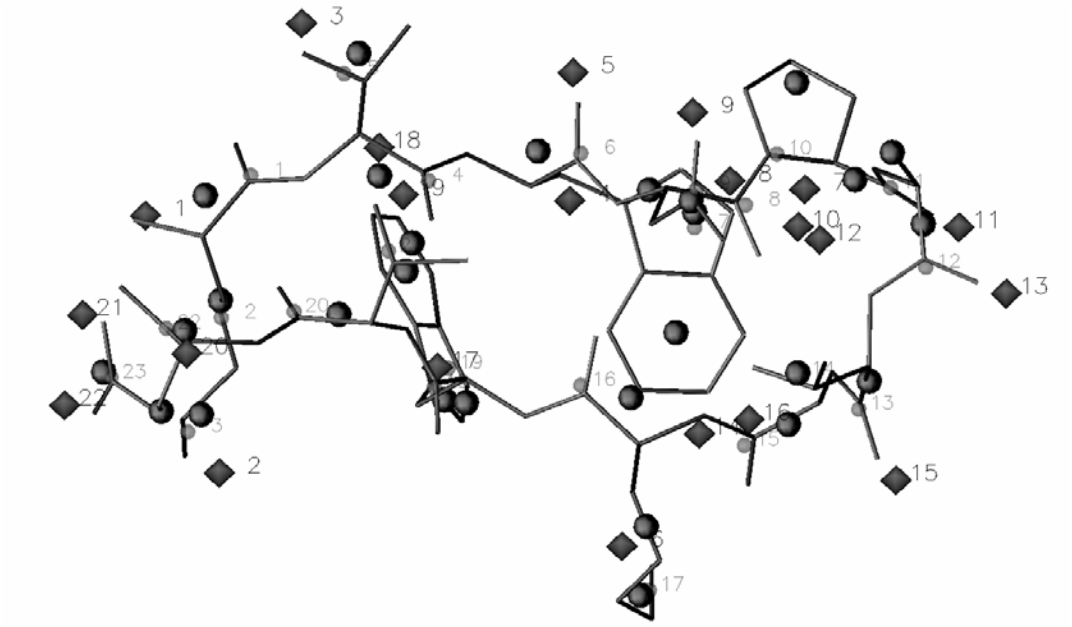


Figure III.14. 3D structure of the 12-residue peptide HP7 (sticks) superimposed on the 22 all-atom Amber MEP point charges at t = 1.4 $bohr^2$ (black diamonds), the 23 PASA ED peaks at t = 1.4 $bohr^2$ (small grey spheres), and the 28-point Basdevant's model (large black spheres). CG points are numbered as in Table III.IX. Figure was generated using DataExplorer [47].

Table III.IX. CG charges (in e^-) for the 12-residue peptide HP7 obtained through a charge fitting algorithm using unsmoothed all-atom Amber MEP grids. CG locations were generated at t = 1.4 bohr^2 using a hierarchical merging/clustering algorithm applied to the PASA ED and all-atom Amber MEP functions. 'BAK' and 'SCH' stand for backbone and side chain, respectively. *rmsdV* and *rmsdμ* are given in kcal/mol and D, respectively. Charges obtained by Basdevant *et al.* [10] are reported for comparison. Point numbers (#) refer to Figure III.14.

Closest c.o.m.		#	[10]	[10] (Fitted)	#	ED-based CG (Eq. 11)	ED-based CG (Fitted)	Closest atom		#	MEP-based CG (Eq. 11)	MEP-based CG (Fitted)
BAK	Lys1	1	1.342	1.033	1	0.952	0.685	N	Lys1	1	1.731	0.797
SCH	Lys1	2	-0.673	-0.412	2	0.053	-0.104	Hς1	Lys1	2	1.841	0.624
SCH	Lys1	3	1.332	0.717	3	0.899	0.660	Hς1	Lys9	3	-0.602	-0.110
BAK	Thr2	4	0.353	-0.452	4	-0.059	-0.374	Oγ1	Thr2	4	0.728	-0.203
SCH	Thr2	5	-0.353	-0.170	5	0.090	-0.168	C	Trp3	5	-0.495	-0.105
BAK	Trp3	6	-0.090	0.235	6	0.019	-0.271	O	Trp3	6	1.061	0.989
SCH	Trp3	7	-0.073	0.061	7	-0.022	0.033	Hε1	Trp3	7	-0.168	0.145
SCH	Trp3	8	0.163	0.101				C	Asn4	8	0.679	0.614
BAK	Asn4	9	0.329	-0.767	8	0.093	-0.317	Oδ1	Asn4	9	-0.527	-0.319
SCH	Asn4	10	-0.329	-0.155	9	-0.023	-0.128	Hδ2	Asn4	10	0.492	0.049
BAK	Asn4				10	0.101	1.020	O	Pro5	11	-0.435	-0.252
BAK	Pro5	11	-0.375	0.270	11	-0.118	-0.265	H	Thr7	12	1.019	0.495
SCH	Pro5	12	0.375	0.298				O	Ala6	13	-0.555	-0.225
BAK	Ala6	13	-0.189	-0.570	12	-0.082	-0.131	Hγ1	Thr7	14	0.031	0.196
SCH	Ala6	14	0.189	0.367				O	Thr7	15	-0.552	-0.172
BAK	Thr7	15	-0.778	-0.435	13	-0.063	-0.641	C	Gly8	16	-0.186	-0.262
SCH	Thr7	16	0.778	0.429	14	0.090	0.470	C	Trp10	17	1.289	0.002
BAK	Gly8	17	0.000	-0.277	15	0.012	-0.151	HH2	Trp10	18	-0.543	0.021
BAK	Lys9	18	0.687	1.290	16	0.095	0.895	Oγ1	Thr11	19	-0.602	0.139
SCH	Lys9	19	-0.718	-0.342	17	0.899	0.837	OXT	Glu12	20	-1.554	-0.852
SCH	Lys9	20	1.031	0.865				Oε1	Glu12	21	-0.824	-0.396
BAK	Trp10	21	0.365	-0.947	18	-0.008	-0.772	Oε2	Glu12	22	-0.824	-0.170
SCH	Trp10	22	0.093	0.538	19	-0.022	0.192					
SCH	Trp10	23	-0.458	-0.307								
BAK	Thr11	24	0.413	1.402	20	0.056	1.329					
SCH	Thr11	25	-0.413	0.023	21	0.090	-0.089					
BAK	Glu12	26	-1.194	-1.708	22	-1.123	-1.457					
SCH	Glu12	27	0.044	0.262	23	-0.925	-0.248					
SCH	Glu12	28	-0.850	-0.343								
Total charge			1.001	1.004		1.004	1.004				1.004	1.004
rmsdV			37.74	5.45		12.78	5.45				33.04	4.62
rmsdμ			23.06	1.57		16.04	1.04				43.13	1.96

Table III.X. 44-point CG model for the 12-residue peptide HP7 built from charges (in e^-) reported in Table III.V (see text for details). *rmsdV* and *rmsdμ* are given in kcal/mol and D, respectively. Coordinates X, Y, and Z are in Å. Point numbers (#) refer to Figure III.15.

#	X	Y	Z	CG location	Residue	Model charges	Optimized charges
1	-3.128	1.192	-0.973	O	Thr2	-0.247	-0.120
2	-3.626	3.719	-1.710	C	Thr2	0.279	0.579
3	-0.213	4.749	-0.653	O	Trp3	-0.210	-0.254
4	-0.350	2.484	-2.079	C	Trp3	0.270	-0.049
5	3.751	0.422	-1.629	O	Asn4	-0.212	-0.125
6	2.691	2.833	-1.094	C	Asn4	0.371	0.426
7	7.470	1.805	-2.513	O	Pro5	-0.163	-0.262
8	5.439	2.844	-1.112	C	Pro5	0.161	0.083
9	8.762	-0.558	0.290	O	Ala6	-0.236	-0.261
10	6.338	0.584	0.320	C	Ala6	0.234	0.308
11	6.116	-3.794	-2.484	O	Thr7	-0.247	-0.247
12	5.388	-1.445	-1.422	C	Thr7	0.279	-0.071
13	2.889	-3.106	-5.576	O	Gly8	-0.205	-0.337
14	3.347	-1.626	-3.384	C	Gly8	0.205	0.326
15	0.230	0.059	-1.888	O	Lys9	-0.239	-0.350
16	-0.359	-2.236	-3.134	C	Lys9	0.367	0.769
17	-2.974	-3.551	-0.908	O	Trp10	-0.210	-0.270
18	-3.164	-0.934	-1.453	C	Trp10	0.270	-0.067
19	-6.482	0.356	-0.257	O	Thr11	-0.247	-0.096
20	-5.445	-1.607	1.247	C	Thr11	0.279	0.696
21	-9.782	1.706	-1.460	N	Lys1	1.127	1.018
22	-8.558	-2.226	-4.238	Nς	Lys1	0.875	0.807
23	-5.926	5.187	0.580	Oγ	Thr2	-0.154	-0.225
24	-5.998	3.255	2.149	Hγ	Thr2	0.120	0.164
25	1.393	3.415	-4.851	5-ring	Trp3	-0.136	-0.054
26	1.396	0.189	-6.098	6-ring	Trp3	-0.098	-0.248
27	4.205	3.764	-5.295	Hε1	Trp3	0.146	0.162
28	3.190	-1.977	-7.552	HH	Trp3	0.028	0.033
29	2.628	2.645	2.624	Oγ	Asn4	-0.178	-0.208
30	2.929	-1.559	1.218	$H\gamma_{tr}$	Asn4	0.019	-0.294
31	4.253	-0.374	3.533	$H\gamma_{cis}$	Asn4	0.002	0.039
32	6.571	1.953	2.094	Cβ	Ala6	-0.000(4)	0.103
33	2.929	-3.312	-0.689	Hγ	Thr7	0.120	-0.007
34	3.603	-1.682	0.868	Oγ	Thr7	-0.154	0.546
35	0.908	-4.071	-6.309	Nς	Lys9	0.875	0.937
36	-3.204	-0.409	-5.747	5-ring	Trp10	-0.136	-0.036
37	-4.409	2.858	-5.518	6-ring	Trp10	-0.098	0.017
38	-2.424	0.132	-8.616	Hε1	Trp10	0.146	0.113
39	-5.904	4.299	-2.815	HH	Trp10	0.028	-0.189
40	-3.676	1.425	1.520	Oγ	Thr11	-0.154	-0.131
41	-5.535	0.735	2.365	Hγ	Thr11	0.120	-0.062
42	-7.699	-2.837	3.942	OXT	Glu12	-1.083	-1.036
43	-10.052	-0.302	-0.588	Oε	Glu12	-0.457	-0.774
44	-10.406	-1.697	-2.343	Oε	Glu12	-0.457	-0.347
					Total charge	0.999	1.004
					rmsdV	7.34	2.80
					rmsdμ	8.89	1.11

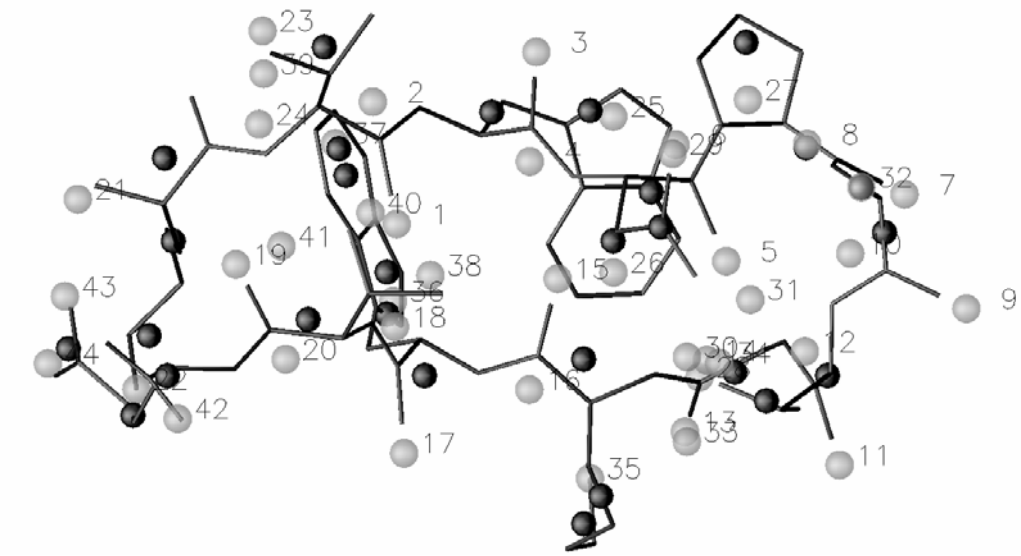


Figure III.15. 3D structure of the 12-residue peptide HP7 (sticks) superimposed on the 44-point model built from local maxima and minima obtained for pentadecapeptides using the hierarchical merging/clustering algorithm applied to the all-atom Amber MEP function smoothed at t = 1.4 bohr^2 (grey spheres) and the 28-point Basdevant's model (black spheres). CG points are numbered as in Table III.X. Figure was generated using DataExplorer [47].

Conclusion

Following a previously described method [20,23,29] for the hierarchical merging/clustering decomposition of a molecular structure, particularly a protein structure, based on a promolecular electron density (ED) distribution function, we present an original application to molecular electrostatic potential (MEP) functions. The approaches allow to reduce the number of representative points of a molecule and assign them point charge values. The decomposition of the protein structure is achieved by following the trajectories of the atoms in progressively smoothed molecular ED or MEP functions. The present work is especially focused on the use of the all-atom Amber MEP function [21], but is readily applicable to other charge sets that are available in the literature.

Two approaches were proposed to study the electrostatic properties of a molecular model. First, for the ED-based results, the Amber atomic charges were used to calculate fragment charges. This was achieved by summing over the charges of the atoms that belong to a fragment. Second, for the MEP-based coarse grain (CG) points, a charge fitting algorithm was used to assign charges from the all-atom unsmoothed MEP. For each model, each of the 20 natural amino acid (AA) residues were studied, with the following specific protonation states: Lys(+1), Arg(+1), Hisε, Glu(-1), and Asp(-1). To generate CG models that avoid too many interaction effects, we selected, for all ED-based and MEP-based calculations, extended β-strand conformations for the molecular structures. These structures consisted

in a set of pentadecapeptide Gly_7-AA-Gly_7, with various rotamers for each of the 20 AA (except Gly, Ala, Asp, and Pro).

The ED-based calculations were all achieved using ideal Gaussian-type promolecular ED distributions, without any random noise. When working with such ED distribution functions, a very interesting situation occurs at a smoothing degree t around 1.4 $bohr^2$, where the protein structure is clearly partitioned into backbone and side chain fragments. One observes one fragment for each residue backbone, mainly composed of –(C=O)-N-Cα or a derivative, and one fragment for each residue side chain, except for Gly, Ala, Ile, Pro, and Val (no fragment at all), and Tyr (two fragments). These observations are consistent with several descriptions already proposed in the literature, such as the globbic description levels of protein structures at a crystallographic resolution of about 3 Å [39] and the CG model proposed by Basdevant *et al*. [10]. Results showed to be independent on the AA residue conformation. On the contrary, the use of MEP functions provided very different decomposition results, which are hardly interpretable in terms of molecular fragments composed of chemically linked atoms, and are very sensitive to the molecular conformation. A detailed analysis was carried out at the smoothing level of 1.4 $bohr^2$, like for the ED-based results, a value beyond which there was no more drastic changes in the merging/clustering decomposition results.

Finally, the particular case of a 12-residue peptide HP7 (PDB code: 2EVQ) was studied. This structure was selected as it is deeply detailed in the literature [10] and was thus an interesting reference case. A 44-point CG model was built and evaluated in terms of its ability to reproduce all-atom MEP and corresponding dipole moment. We chose to design a CG model that already involves some simplifications for non cyclic C-H residues, sulfur-containing residues, and Lysine, with side chain CG charges placed at selected atom locations.

Further developments might include strategies to directly design CG representations for all 20 AA residues from their atomic coordinates. It is shown that without an optimization stage, our model is of a similar quality than the previously published CG model [10]; after optimization, the CG distribution is shown to provide a really better representation of the MEP and dipole moment.

Another extension of the present work resides in the evaluation of backbone charges as a function of the residue location along the protein sequence. Indeed, the CG charges of Gly_{15} models and the optimization results of the 44-point model of HP7 showed that the local charge separations observed along each C=O axis is far from being a constant.

Other perspectives to the present work are numerous. First of all, an extension to various molecular systems is needed to validate the transferability of our model. Second, the effect of the point charge set can be studied by considering other all-atom force fields such as, for example, in [40]. In this last work, a semi-empirical quantum mechanical procedure (FCPAC) was used to calculate the

partial atomic charges of amino acids from 494 high-resolution protein structures. Each AA was either considered as the center of a tripeptide with the PDB geometry (free) or the center of 13 to 16 AA clusters (buried). A more general parametrization, applicable to organic molecules, peptides, and proteins, has also been presented by Arnautova *et al*. [41] in the so-called ECEPP-05 force field (FF). The partial atomic charge of multiple configurations of small molecules were obtained by fitting to the MEP calculated with the HF/6-31G* quantum mechanical approach. Other sets of atomic charges are also available. For example, Matta and Bader [42] reported charges of isolated amino acids determined through the quantum theory of atoms-in-molecules (QTAIM) and showed their transferability properties. We can also mention databases of transferable parameters to evaluate atom charges of protein structures, as, for example, designed by Lecomte *et al.* [43,44] and already used in a previous work [20]. Another set of atomic charges in the Amber-type FF family designed for proteins can be found in [45]. In that new generation united-atom force field, all hydrogen atoms bonded to aliphatic carbons in all AA are united with C except those on Cα. Polar and aromatic H are represented explicitly. Charges were obtained as in [21]. In that family of FF, we can also cite the Gromos charge sets implemented in the program GROMACS [46]. Coarser descriptions are also available, such as the one proposed by Gabb *et al*. [17] who reported protein docking studies where electrostatic complementarity was evaluated by Fourier correlation.

Finally, a resolution dependency of the CG model could be studied, with the expected behavior that, at lower smoothing levels, the efficiency of the model is expected to be better since the number of CG points and their charges would be closer to the initial all-atom MEP function.

Acknowledgments

The authors thank Cl. Lecomte and the members of his research group for fruitful discussions. The 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] Tozzini, V. *Curr. Opin. Struct. Biol*. **2005**, *15*, 144-150.

[2] Nielsen, S. O.; Lopez, C. F.; Srinivas, G.; Klein, M. L. *J. Phys.: Condens. Matter* **2004**, *16*, R481-R512.

[3] Neri, M.; Anselmi, C.; Cascella, M.; Maritan, A.; Carloni, P. *Phys. Rev. Lett*. **2005**, *95*, 218102/1-218102/4.

[4] Colombo, G.; Micheletti, C. *Theor. Chem. Acc*. **2006**, *116*, 75-86.

[5] Vizcarra, C. L.; Mayo, S. L. *Curr. Opin. Chem. Biol*. **2005**, *9*, 622-626.

[6] Stigter, D.; Alonso, D. O. V.; Dill, K. A. *Proc. Natl. Acad. Sci. USA* **1991**, *88*, 4176-4180.

[7] Kumar, S.; Wolfson, H. J.; Nussinov, R. *IBM. J. Res. & Dev*. **2001**, *45*, 499-512.

[8] Skepö, M.; Linse, P.; Arnebrant, T. *J. Phys. Chem*. B **2006**, *110*, 12141-12148.

[9] Curcó, D.; Nussinov, R.; Alemán, C. *J. Phys. Chem. B* **2007**, *111*, 10538-10549.

[10] Basdevant, N.; Borgis, D.; Ha-Duong, T. *J. Phys. Chem. B* **2007**, 111, 9390-9399.

[11] Pizzitutti, F.; Marchi, M.; Borgis, D. *J. Chem. Theory Comput*. **2007**, *3*, 1867-1876.

[12] Hinsen, K., Petrescu, A.-J., Dellerue, S.; Bellissent-Funel, M.-C.; Kneller, G. R. Chem. Phys. 2000, 261, 25-37.

[13] Tama, F.; Gadea, F. X.; Marques, O.; Sanejouand, Y.-H. *Proteins* **2000**, *41*, 1-7.

[14] Schuyler, A. D.; Chirikjian, G. S. *J. Mol. Graph. Model***. 2004**, *22*, 183-193.

[15] Rader, A. J.; Chennubhotla, Ch.; Yang, L.-W.; Bahar, I. In *Normal Mode Analysis. Theory and Applications to Biological and Chemical Systems*; Cui, Q.; Bahar, I., Eds.; CRC Press: Boca Raton, FL, 2006, pp 41-64.

[16] Heyden, A.; Truhlar, D. G. *J. Chem. Theory Comput*. **2008**, *4*, 217-221.

[17] Gabb, H. A.; Jackson, R. M.; Sternberg M. J. E. *J. Mol. Biol*. **1997**, *272*, 106-120.

[18] Imai, K.; Mitaku, S. *Chem-Bio. Inform. J*. **2003**, *3*, 194-200.

[19] Golubkov, P. A.; Ren, P. Y. *J. Chem. Phys*. **2006**, *125*, 064103/1-064103/11.

[20] Leherte, L; Guillot, B.; Vercauteren, D.; Pichon-Pesme, V.; Jelsch, Ch.; Lagoutte, A.; Lecomte, Cl. In *The Quantum Theory of Atoms in Molecules - From Solid State to DNA and Drug Design*; Matta, C. F.; Boyd, R. J., Eds.; Wiley-VCH: Weinheim, 2007, pp 285-315.

[21] Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G. M.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J. M.; Kollman, P. *J. Comput. Chem*. **2003**, *24*, 1999-2012.

[22] Leung, Y.; Zhang, J.-S.; Xu, Z.-B. *IEEE T. Pattern Anal*. **2000**, *22*, 1396-1410.

[23] Leherte, L.; Dury, L.; Vercauteren, D. P. *J. Phys. Chem. A* **2003**, *107*, 9875-9886.

[24] Gilbert, D. G. *Phylodendron, for Drawing Phylogenetic Trees*; Indiana University: Bloomington, IN, 1996; http://iubio.bio.indiana.edu/treeapp/

[25] Dury, L. *DENDRO*; Facultés Universitaires Notre-Dame de la Paix: Namur, Belgium, 2002.

[26] Amat, L.; Carbó-Dorca, R. *J. Chem. Inf. Comput. Sci*. **2000**, *40*, 1188-1198.

[27] Amat, L.; Carbó-Dorca, R. *J. Comput. Chem.* **1997**, *18*, 2023-2039.

[28] Kostrowicki, J.; Piela, L.; Cherayil, B. J.; Scheraga, H. A. *J. Phys. Chem*. **1991**, *95*, 4113-4119.

[29] Leherte, L. *Acta Crystallogr. D* **2004**, *60*, 1254-1265.

[30] Abramowitz, M.; Stegun, I. A., Eds. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*; Dover Publications: New York, NY, **1970**.

[31] Hart, R. K.; Pappu, R. V.; Ponder, J. W. *J. Comput. Chem*. **2000**, *21*, 531-552.

[32] Borodin, O.; Smith, G. D. *Force Field Fitting Toolkit*; University of Utah: Salt lake City, UT; http://www.eng.utah.edu/~gdsmith/fff.html

[33] Becue, A.; Meurice, N.; Leherte, L.; Vercauteren, D. P. *Acta Crystallogr. D* **2003**, *59*, 2150-2162.

[34] Singh, U. C.; Kollman, P. A. *J. Comput. Chem*. **1984**, *5*, 129-145.

[35] Eisenmenger, F.; Hansmann, U. H. E.; Hayryan, S.; Hu, C.-K. *Comp. Phys. Comm*. **2006**, *174*, 422-429; http://www.smmp05.net/

[36] Nemethy, G.; Gibson, K. D.; Palmer, K. A.; Yoon, C. N.; Paterlini, G.; Zagari, A.; Rumsey, S.; Scheraga, H. A. *J. Phys. Chem.* **1992**, *96*, 6472-6484.

[37] Simms, A. M.; Toofanny, R. D.; Kehl, C.; Benson, N. C.; Daggett, V. *Prot. Eng. Design & Selection* **2008**, *21*, 369-377; http://www.dynameomics.org/

[38] Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic Acids Res.* **2000**, *28*, 235-242; http://www.rcsb.org/pdb

[39] Guo, D.-Y.; Blessing, R. H.; Langs, D. A.; Smith, G. D. *Acta Cryst. D* **1999**, *55*, 230-237.

[40] Thomas, A.; Milon, A.; Brasseur, R. *Proteins* **2004**, *56*, 102-109.

[41] Arnautova, Y.A.; Jagielska, A.; Scheraga, H. A. *J. Phys. Chem. B* **2006**, *10*, 5025-5044.

[42] Matta, C. F.; Bader, R. F. W. *Proteins* **2003**, *52*, 360-399.

[43] Pichon-Pesme, V.; Lecomte, C., Lachekar, H. *J. Phys. Chem*. **1995**, *99*, 6242-6250.

[44] Jelsch, C.; Pichon-Pesme, V.; Lecomte, C., Aubry, A. *Acta Crystallogr. D* **1998**, *54*, 1306-1318.

[45] Yang, L.; Tan, C.-H.; Hsieh, M.-J.; Wang, J. M.; Duan, Y.; Cieplak, P.; Caldwell, J.; Kollman, P. A.; Luo, R. *J. Phys. Chem. B* **2006**, *110*,13166-13176.

[46] van der Spoel, D.; Lindahl, E.; Hess, B.; van Buuren, A. R.; Apol, E.; Meulenhoff, P. J.; Tieleman, D. P.; Sijbers, A. L. T. M.; Feenstra, K. A.; van Drunen, R.; Berendsen, H. J. C.; *Gromacs User Manual version 3.3*, 2005; http://www.gromacs.org/

[47] *OpenDX, The Open Source Software Project Based on IBM's Visualization Data Explorer*; Visualization and Imagery Solutions, Inc.; http://www.opendx.org/

[48] Guex, N.; Peitsch, M. C. *Electrophoresis* **1997**, *18*, 2714-2723; http://us.expasy.org/spdbv/

IV. Refinement of the Amber-Based CG Model

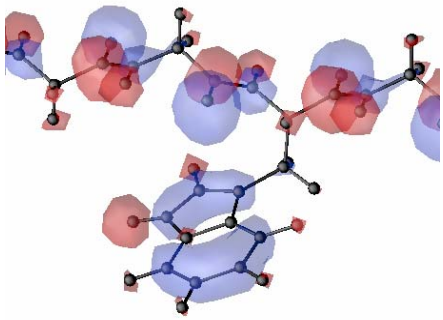
Following the previous publication regarding the determination of CG charges from smoothed Amber MEP (see Section III and Appendix I), a refined approach is presented in this Section. First, rather than working at a single smoothing value t = 1.4 bohr^2, we have carried out a set of charge fitting calculations at various t in order to select the most adapted smoothing degree at which the protein CG model should be established. This value t corresponds to the CG model that allows the best fit with the corresponding all-atom MEP. Second, we have chosen to determine the location of the CG backbone points through the use of a superposition algorithm of amino acid CG templates rather than using approximate geometric relationships.

Selection of the Smoothing Degree

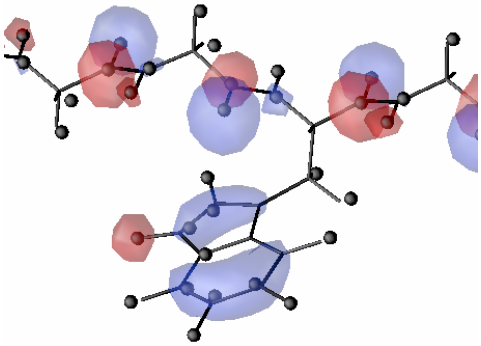
As illustrated in Figure IV.1 for residue Trp located at the center of a pentadecapeptide chain, the CG description of an AA is dependent on the smoothing value t. At t = 0.05 bohr^2, peaks and pits observed in the MEP are closely located on the atoms of the molecular structure. Starting at t = 0.3 bohr^2, the CPs begin to move away from the atomic centers and their number decreases, first for the CPs located near the H atoms (illustrated at t = 0.5 and 1.0 bohr^2), and finally, for the CPs of the rings (illustrated at t = 2.0 and 2.5 bohr^2). At t = 2.5 bohr^2, it is even difficult to visually and unambiguously assign a point either to the backbone or to the side chain (see arrow on Figure IV.1). At such a high smoothing level, the side chain conformation certainly strongly affects the CG representation of the AA residue.

To select the optimal smoothing degree, we have used the charge fitting algorithm QFIT [bor] and applied it, with the same conditions as reported in Section III, to each set of peaks and pits obtained for the β-Gly$_{15}$ structure at various smoothing levels. The resulting Minimal Objective Function (MOF) values are displayed in Figure IV.2. The MOF function is built on the *rmsdV* and *rmsdμ* values as defined in Section III. The best fit is obtained at t = 1.35 bohr^2, with *MOF* = 1.462.

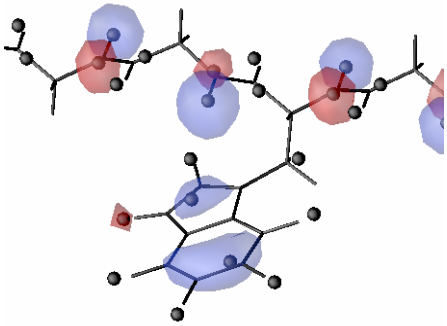
The major structural difference between the models obtained at t = 1.4 bohr^2, discussed in Section III, and 1.35 bohr^2, lies in the presence of one extra point in the latter case, as clearly observable by comparing Figures III.2 (top) and IV.3.



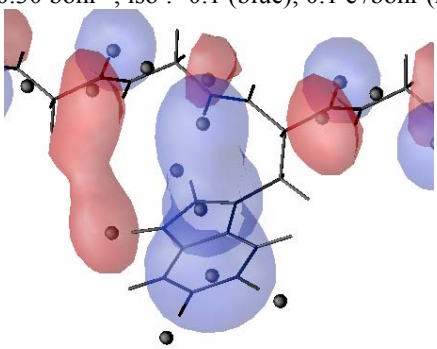
t = 0.05 bohr^2 ; iso : -0.1 (blue), 0.1 e$^-$/bohr (red)



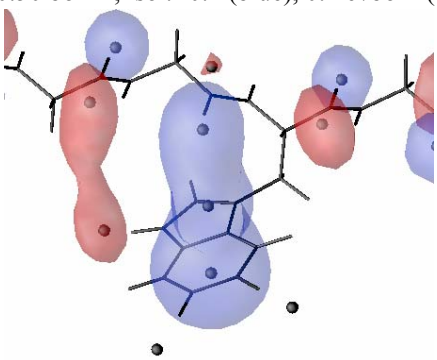
t = 0.30 bohr^2 ; iso : -0.1 (blue), 0.1 e$^-$/bohr (red)



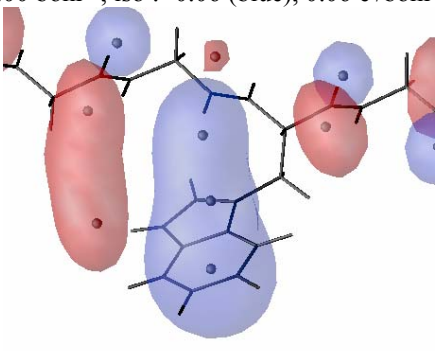
t = 0.50 bohr^2 ; iso : -0.1 (blue), 0.1 e$^-$/bohr (red)



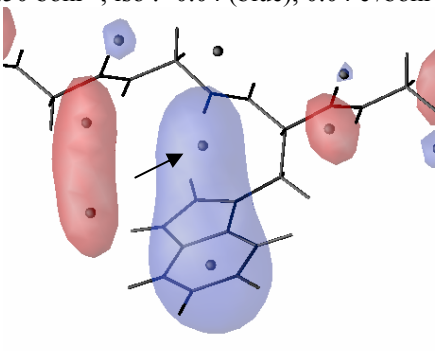
t = 1.00 bohr^2 ; iso : -0.06 (blue), 0.06 e$^-$/bohr (red)



t = 1.50 bohr^2 ; iso : -0.04 (blue), 0.04 e$^-$/bohr (red)



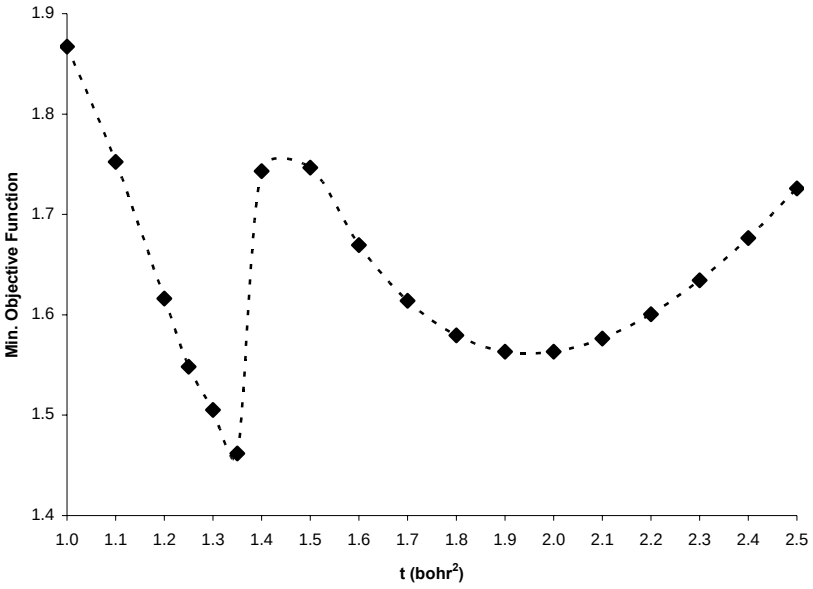
t = 2.00 bohr^2 ; iso : -0.03 (blue), 0.03 e$^-$/bohr (red)



t = 2.50 bohr^2 ; iso : -0.03 (blue), 0.03 e$^-$/bohr (red)

Figure IV.1. Amber MEP iso-contours of Gly_7-Trp-Gly_7 in the *g*-,*g*- conformation, smoothed at various values of *t*. Local maxima and minima (black spheres) were obtained using the hierarchical merging/clustering algorithm applied to the all-atom Amber MEP function. Figures were generated using DataExplorer [odx].

The loss of that extra point involves a steep rise in the MOF value, followed by a slower decrease, observed up to t = 1.9 bohr^2. Between t = 1.5 to 1.9 bohr^2, the better fit is due only to a more adequate arrangement of peaks and pits, as their number is constant, *i.e.*, equal to 30 (Figure IV.2).



t	MOF	#
0.05	0.00	105
0.50	0.82	60
0.80	1.93	47
0.90	1.96	46
1.00	1.87	44
1.10	1.75	33
1.20	1.62	32
1.25	1.55	32
1.30	1.51	32
1.35	1.46	32
1.40	1.74	31
1.50	1.75	30
1.60	1.67	30
1.70	1.61	30
1.80	1.58	30
1.90	1.56	30
2.00	1.56	30
2.50	1.73	30

Figure IV.2. Minimal objective function (MOF) for the charge fitting of ß-Gly$_{15}$ MEP CGs from unsmoothed all-atom Amber MEP, as a function of the smoothing degree t. # stands for the number of local minima and maxima observed in the MEPs.

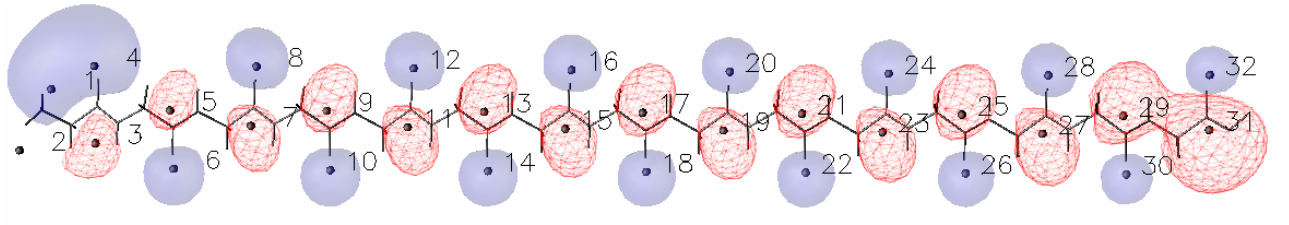


Figure IV.3. Amber MEP iso-contours (blue: -0.03 ; red: 0.03 e$^-$/bohr) of ß-Gly$_{15}$ smoothed at t = 1.35 bohr^2. Local maxima and minima at t = 1.35 bohr^2 were obtained using the hierarchical merging/clustering algorithm applied to the all-atom Amber MEP function. CG points are numbered as in Table IV.I. Figure was generated using DataExplorer [odx].

For comparison with results obtained at t = 1.4 bohr^2 (Table III.II), fitted CG charges of structure β-Gly$_{15}$ are reported in Table IV.I. Positive and negative charges located near the C and O atoms of the central residue Gly8 are now equal to ± 0.208 e$^-$, rather than ± 0.205 e$^-$ for the model established at t = 1.4 bohr^2 (Table III.II), and are separated by a distance of 2.59 Å. *rmsdV* and

rmsdμ values are also slightly lower than the values obtained at t = 1.4 bohr^2, with values of 1.21 kcal/mol and 0.28 D, respectively.

As described in Section III, CG descriptions and charges were established for each of the AA residues. Charge values obtained for the CG descriptions at t = 1.35 bohr^2 are reported in Table IV.II and corresponding CG structures are shown in Figure IV.4. Let us note that charges are now given with a precision of four digits because, as further explained in Section VI, corrections of that order of magnitude may be considered. In the present case, two charge sets are presented for a unique CG representation of Asn. The first one was established from the MEPs of the three selected rotamers, as in Section III. The second one was obtained by considering only the MEPs of the two rotamers that are described by a similar CG motif, *i.e.*, by excluding conformation *t,Nt* that presents, its in side chain, a single CG located close to atom Oδ. The resulting *rmsdV* and *rmsdμ* values are largely improved, especially for *rmsdμ* that evolves from 2.3-2.7 to 0.18 D. For Phe, twi CG models are given. For each of them, point 33 is located exactly at the same position in the 6-membered ring.

Table IV.I. CG charges $q_{0.0}$ (in e$^-$) of ß-Gly$_{15}$ fitted from the all-atom Amber MEP grid smoothed at t = 0.0 bohr^2, using the program QFIT. Local maxima and minima at t = 1.35 bohr^2 were obtained using the hierarchical merging/clustering algorithm applied to the all-atom Amber MEP function. For each point, the distance *vs.* the closest atom, *d*, is given in Å. *rmsdV* and *rmsdμ* are given in kcal/mol and D, respectively. Point numbers (#) refer to Figure IV.3.

#	Closest atom		*d*	$q_{0.0}$	#	Closest atom		*d*	$q_{0.0}$
1	H	Gly1	0.955	-0.1123	17	C	Gly8	0.788	0.2078
2	H	Gly1	1.154	0.0580	18	O	Gly8	0.605	-0.2075
3	Hα	Gly1	1.125	0.2067	19	C	Gly9	0.789	0.2101
4	O	Gly1	0.583	-0.2226	20	O	Gly9	0.606	-0.2078
5	C	Gly2	0.760	0.2235	21	C	Gly10	0.789	0.2055
6	O	Gly2	0.589	-0.2361	22	O	Gly10	0.605	-0.2070
7	C	Gly3	0.799	0.2247	23	C	Gly11	0.789	0.2127
8	O	Gly3	0.607	-0.2112	24	O	Gly11	0.606	-0.2059
9	C	Gly4	0.786	0.2007	25	C	Gly12	0.788	0.1982
10	O	Gly4	0.603	-0.2074	26	O	Gly12	0.605	-0.2084
11	C	Gly5	0.792	0.2079	27	C	Gly13	0.790	0.2235
12	O	Gly5	0.606	-0.2043	28	O	Gly13	0.608	-0.2077
13	C	Gly6	0.788	0.2062	29	C	Gly14	0.783	0.2168
14	O	Gly6	0.605	-0.2060	30	O	Gly14	0.614	-0.2163
15	C	Gly7	0.790	0.2081	31	C	Gly15	0.590	0.2756
16	O	Gly7	0.605	-0.2069	32	O	Gly15	0.738	-0.2034
								rmsdV	1.21
								rmsdμ	0.28

Table IV.II. CG charges (in e^-) for the AA residues obtained through a charge fitting algorithm using unsmoothed all-atom Amber MEP grids. CG locations were generated at t = 1.35 bohr^2 using a hierarchical merging/clustering algorithm applied to the all-atom Amber MEP function. *g* and *t* stand for *gauche* and *trans*, respectively (see [sim08] for details). *rmsdV* and *rmsdμ* are given in kcal/mol and D, respectively. Point numbers refer to Figure IV.4.

	Conformation	Point 17	Point 18	Point 33	Point 34	Point 35	Point 36	Point 37	*rmsdV*	*rmsdμ*
Ala		C	O	Cβ						
		0.2349	-0.2399	0.0020					1.29	0.45
Arg		C	O	NH-NH_2	NH_2-NH	NH_2-NH_2				
	g-,t,g-,g-								1.92	0.91
	g-,t,g-,t								1.61	0.92
	g-,t,g+,t	0.3141	-0.2022	0.2801	0.3162	0.2813			1.57	0.26
	g-,t,t,t								1.58	0.26
Asn		C	O	$H\delta_{tr}$	Oδ	$H\delta_{cis}$				
	t,Nt								1.91	3.09
	t, Og-	0.3887	-0.2202	0.0204	-0.1869	0.00077			2.44	2.28
	t, Og+								2.38	2.71
Asn		C	O	$H\delta_{tr}$	Oδ	$H\delta_{cis}$				
	t, Og-	0.2614	-0.2003	0.1034	0.2316	0.0689			1.45	0.18
	t, Og+								1.45	0.18
Asp		C	O	Oδ1	Oδ2					
	t,g+	-0.0180	-0.2044	-0.4175	-0.3593				1.47	0.54
Cys		C	O	Sγ						
	g-								1.56	0.65
	g+	0.4033	-0.2949	-0.1025					1.84	1.52
	t								1.71	1.25
Gln		C	O	H_{tr}	Cγ	Oε	H_{cis}			
	g-,t,Nt								1.70	0.84
	g-,t,Og-	0.2986	-0.2939	0.1679	0.0013	-0.2615	0.0837		1.56	0.68
	g-,t,Og+								1.60	0.31
Glu		C	O	Oε1	Oε2					
	g-,t,g-	0.1915	-0.2734	-0.4579	-0.4583				1.40	0.28
	g-,t,g+								1.44	0.25
Gly		C	O							
		0.2078	-0.2075						1.21	0.28
His		C	O	Hε	Nδ					
	g-,Ng-	0.1998	-0.1905	0.1790	-0.1845				1.44	0.26
	t,Ng+								1.48	0.30
Ile		C	O	Cβ	Cδ1					
	g-,g-								1.33	0.53
	g-,t								1.34	0.55
	g+,t	0.2292	-0.2839	0.0681	-0.0118				1.30	0.49

Leu		C	O	Cγ	Cδ1	Cδ2				
	g-,t	0.2219	-0.2479	0.0586	-0.0280	-0.0096			1.25	0.37
	t,g+								1.25	0.53
Lys		C	O	Nς						
	g-,g-,t,g-								1.58	0.99
	g-,g-,t,g+	0.3627	-0.2390	0.8726					1.54	1.34
	g-,t,t,g-								1.54	1.11
	g-,t,t,g+								1.59	0.82
Met		C	O	Sδ						
	g-,g-,g-								1.82	1.80
	g-,g-,t								2.08	2.13
	g-,t,g-	0.2866	-0.2358	-0.0589					1.71	1.40
	g-,t-g+								1.73	1.51
	g-,t,t								1.94	1.79
Phe		C	O	6-ring	Hδ2	Hε1	Hς	Hε2		
	g-,g-	0.2655	-0.2371	-0.1659	0.0333	0.0482	0.0315	0.0270	1.29	0.22
	t,g+								1.30	0.35
Phe		C	O	6-ring						
	g-,g-	0.2243	-0.2220	-0.0034					1.43	0.12
	t,g+								1.45	0.03
Pro		O	C							
	g+	-0.1628	0.1598						1.70	1.69
Ser		C	O	Oγ	Hγ					
	g-	0.3096	-0.2795	-0.1765	0.1546				1.40	0.37
	g+								1.55	0.72
Thr		C	O	Oγ	Hγ					
	g-	0.2834	-0.2507	-0.1572	0.1229				1.35	0.47
	g+								1.41	0.45
Trp		C	O	5-ring	6-ring	Hε1	HH			
	g-,g-								1.42	0.22
	g-,t								1.45	0.18
	t,g-	0.2743	-0.2139	-0.1387	-0.0996	0.1487	0.0294		1.49	0.47
	t,g+								1.42	0.31
	t,t								1.42	0.20
Tyr		C	O	6-ring	OH	HH	Hδ*	Hε*		
	g-,g-	0.2702	-0.2371	-0.1116	-0.1318	0.1591	0.0238	0.0378	1.35	0.18
	t,g+								1.38	0.17
Val		C	O	Cβ						
	g-	0.0924	-0.0513	-0.0514					1.53	0.60
	t								1.60	0.76

* Hγ and Hδ stand on the opposite side of the O-H bond direction.

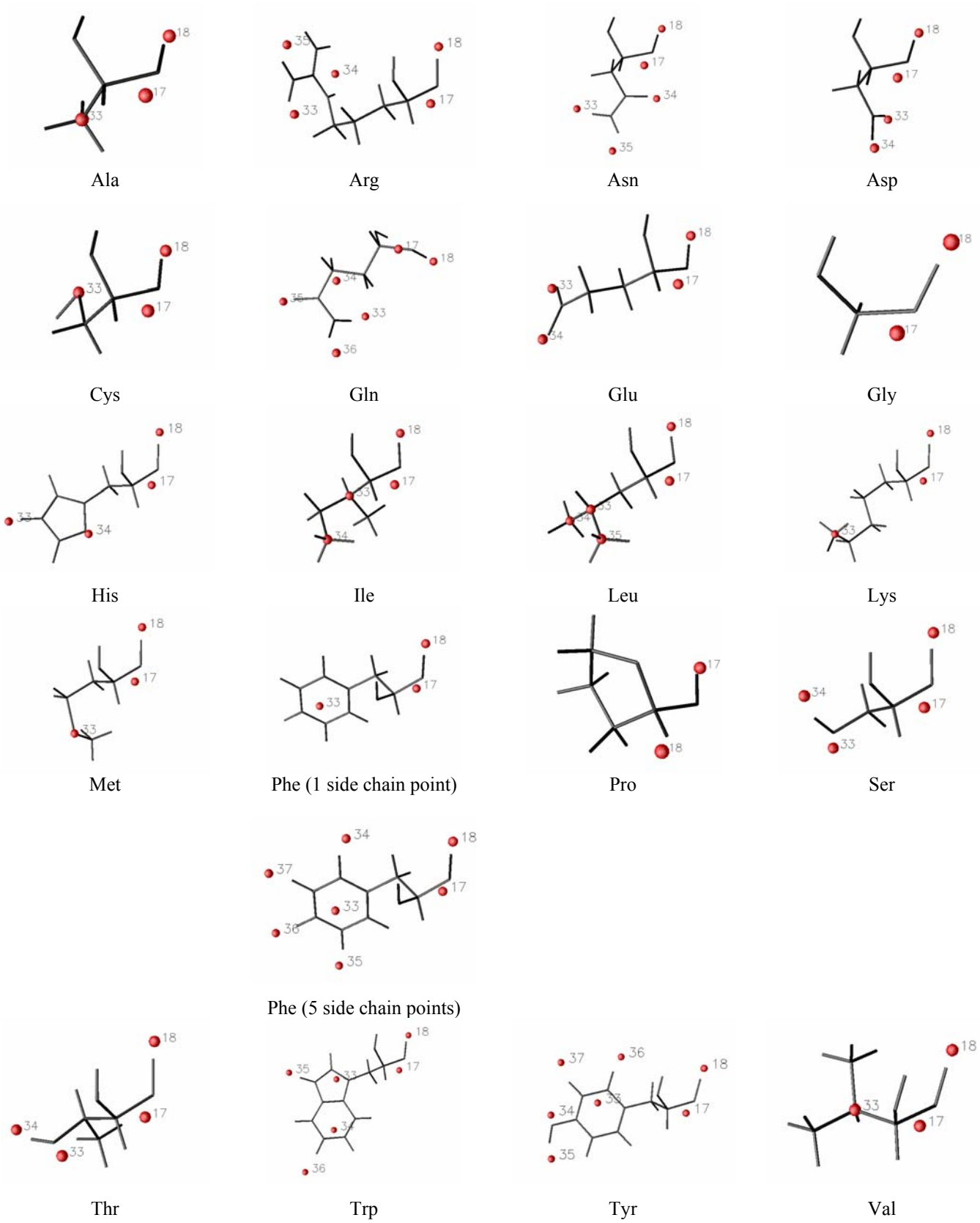


Figure IV.4. CG model for each of the 20 AA residues as established at t = 1.35 bohr^2 from the hierarchical merging/clustering algorithm applied to the all-atom Amber MEP function. CG points are numbered as in Table IV.II. Figures were generated using DataExplorer [odx].

Application to 12-Residue β-Hairpin HP7

As reported in Section III, the determination of the location of the side chain CG points was achieved by applying our merging/clustering procedure to the isolated AA in their PDB conformation. We here selected a different approach to generate the backbone CG points. Rather than applying a simple geometric relationship, such as one based on distances with respect to C and O atoms, we have selected the 3-atom motif, $(C, O)_{Gly8}N_{Gly9}$, observed in β-Gly_{15} and the corresponding two-point motif obtained at t = 1.35 $bohr^2$ (points 17 and 18 in Table IV.1). Through the use of the program QUATFIT [hei90], the CG coordinates of each AA backbone of HP7 were determined by first superimposing the three atoms $(C, O)_{Gly8}$ and N_{Gly9} of β-Gly_{15} on the corresponding atoms of each HP7 AA residues, and second, by applying the resulting transformation matrix to the CG coordinates. The backbone CG points of the two end residues were obtained by adding unitary positive and negative charges on the N_{Lys1} and OXT_{Glu12} atoms, respectively. As also reported in Section III, the generations of backbone and side chain CGs were achieved separately, as the side chain descriptions may be dependent on the AA conformation. The side chain CG points were obtained by determining the MEP local minima and maxima for each AA residue separately. This procedure led to a 48-point model for the HP7 structure (Table IV.III. and Figure IV.5). For residue Asn4, the best results were obtained with the charges fitted on the two-rotamer model (Table IV.II). By comparison with our model, the fitted charges on the Basdevant's model led to *rmsdV* = 5.45 kcal/mol and *rmsdμ* = 1.57 D. Therefore, a non fitted MEP-CG model is as good as a c.o.m.-based model, but is clearly better when a fitting is applied.

In comparison with the results obtained at t = 1.4 $bohr^2$ (Section III), where *rmsdV* = 7.34 and *rmsdμ* = 8.89 D, the refined model is characterized by *rmsdV* and *rmsdμ* values equal to 4.63 and 5.51 D, respectively. The backbone CG description involves 26 points rather than 22, while the number of side chain CGs is left unchanged, *i.e.*, 22, as illustrated by comparing Figures III.11 and IV.4. Besides slight changes in the CG charge values, the backbone description may thus be important in the quality of CG models, most probably in the modeling of the end charges. As additionally shown in Table IV.III., while the optimized charges may differ significantly from the model charges, their sign is rather well preserved. There are only four charge inversions, that occur at CG numbers 12, 35, 38, and 39. This however remains hardly interpretable except for the fact that those four points are located at the level of the AA residues that form the bend in the peptide structure.

Table IV.III. 48-point CG model for the 12-residue peptide HP7 built from charges (in e^-) reported in Table IV.II. *rmsdV* and *rmsdμ* are given in kcal/mol and D, respectively. Coordinates X, Y, and Z are in Å. Point numbers (#) refer to Figure IV.5.

#	X	Y	Z	CG location	Residue	Model charges	Optimized charges
1	-9.426	1.989	-1.495	N	Lys1	1.0000	0.9325
2	-6.954	2.039	-1.366	C	Lys1	0.3627	0.7429
3	-7.533	4.340	-2.400	O	Lys1	-0.2390	-0.3331
4	-3.654	3.692	-1.452	C	Thr2	0.2834	0.4115
5	-3.105	1.208	-0.971	O	Thr2	-0.2507	-0.3780
6	-0.385	2.696	-2.245	C	Trp3	0.2743	0.3149
7	-0.175	4.731	-0.658	O	Trp3	-0.2139	-0.2939
8	2.679	2.681	-0.884	C	Asn4	0.2614	0.6359
9	3.771	0.453	-1.619	O	Asn4	-0.2003	-0.4756
10	5.340	2.722	-1.332	C	Pro5	0.1598	0.2694
11	7.463	1.783	-2.478	O	Pro5	-0.1628	-0.2639
12	6.509	0.742	0.186	C	Ala6	0.2349	-0.0429
13	8.735	-0.576	0.264	O	Ala6	-0.2399	-0.2083
14	5.640	-1.498	-1.356	C	Thr7	0.2834	0.1017
15	6.118	-3.767	-2.508	O	Thr7	-0.2507	-0.2716
16	3.426	-1.498	-3.599	C	Gly8	0.2078	0.3987
17	2.864	-3.087	-5.564	O	Gly8	-0.2075	-0.3746
18	-0.170	-2.289	-2.938	C	Lys9	0.3627	0.7023
19	0.185	0.046	-1.878	O	Lys9	-0.2390	-0.4284
20	-3.236	-1.077	-1.677	C	Trp10	0.2743	0.1738
21	-3.009	-3.531	-0.885	O	Trp10	-0.2139	-0.2796
22	-5.287	-1.380	1.274	C	Thr11	0.2834	0.6033
23	-6.510	0.344	-0.220	O	Thr11	-0.2507	-0.6760
24	-7.933	-2.461	2.676	C	Glu12	0.1915	0.2863
25	-9.551	-0.877	3.931	O	Glu12	-0.2734	-0.3335
26	-7.612	-2.864	3.802	OXT	Glu12	-1.0000	-0.9941
27	-8.558	-2.226	-4.238	Nζ	Lys1	0.8726	0.8128
28	-6.457	5.442	0.656	Oγ	Thr2	-0.1572	-0.2336
29	-5.325	3.369	1.929	Hγ	Thr2	0.1229	0.1483
30	1.351	3.482	-4.734	5-ring	Trp3	-0.1387	-0.0641
31	4.303	3.708	-5.153	6-ring	Trp3	-0.0996	-0.2723
32	1.381	0.234	-6.043	Hε1	Trp3	0.1487	0.1616
33	3.108	-2.064	-7.441	HH	Trp3	0.0294	0.0348
34	3.030	-1.276	0.781	Oδ	Asn4	-0.2316	-0.2572
35	2.747	2.707	2.505	$H\delta_{tr}$	Asn4	0.1034	-0.3258
36	4.113	-0.614	3.448	$H\delta_{cis}$	Asn4	0.0689	0.0560
37	6.571	1.953	2.094	Cβ	Ala6	0.0020	0.1046
38	3.238	-1.211	0.619	Oγ	Thr7	-0.1572	0.6244
39	3.398	-3.723	-0.308	Hγ	Thr7	0.1229	-0.0247
40	0.908	-4.071	-6.309	Nζ	Lys9	0.8726	0.9688
41	-3.195	-0.449	-5.845	5-ring	Trp10	-0.1387	-0.0295
42	-2.527	0.091	-8.710	6-ring	Trp10	-0.0996	-0.1100
43	-4.405	2.817	-5.480	Hε1	Trp10	0.1487	0.1265
44	-4.794	5.686	-6.833	HH	Trp10	0.0294	0.0237
45	-3.784	2.096	1.550	Oγ	Thr11	-0.1572	-0.1544
46	-5.381	0.117	2.406	Hγ	Thr11	0.1229	0.0617
47	-10.069	-0.268	-0.658	Oε1	Glu12	-0.4581	-0.4016
48	-10.420	-1.793	-2.459	Oε2	Glu12	-0.4581	-0.4656
					Total charge	0.9862	1.0039
					rmsdV	4.63	1.56
					rmsdμ	5.51	0.21

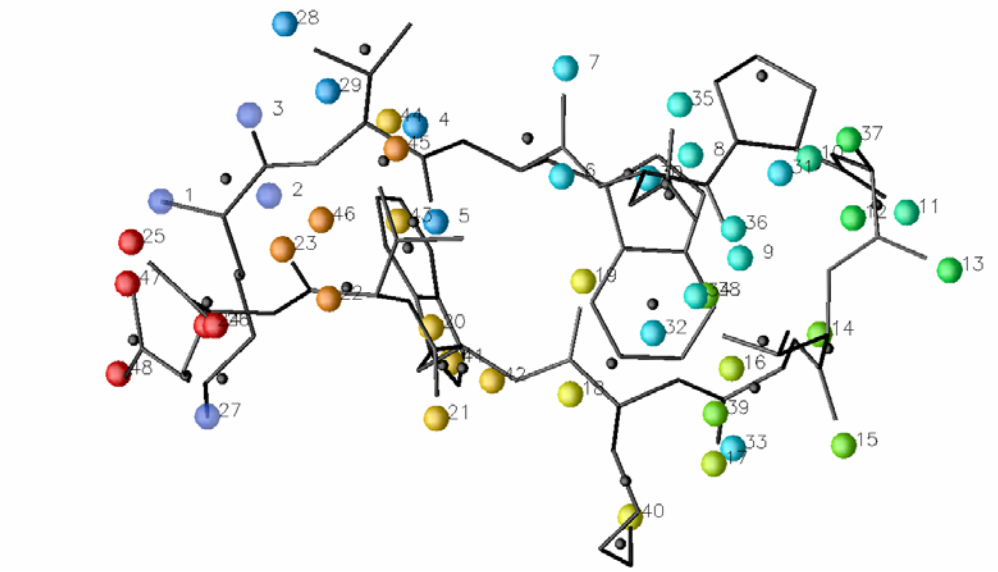
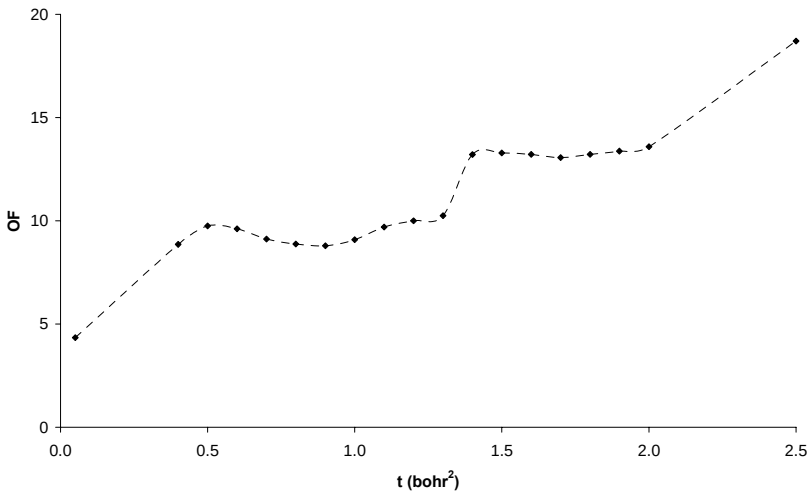


Figure IV.5. 3D structure of the 12-residue peptide HP7 (sticks) superimposed on the 48 Amber CG point charges at t = 1.35 bohr^2 (color-coded by AA), and the 28-point Basdevant's model (black spheres). CG points are numbered as in Table IV.III. Figure was generated using DataExplorer [odx].

In addition to the refinement of the MEP-based CG description of a protein, we have also deepened the study of the ED-based CG models, *i.e.*, models built from ED peaks in smoothed PASA ED distribution functions. In Figure IV.6, we report the evolution of the Objective Function (OF) calculated between the MEP generated by the peak charges obtained using Equation 11 of Section III considering the FF Amber charges [dua03], and the all-atom Amber MEP. There is thus no charge fitting. The results clearly show that t = 1.4 bohr^2 is not the best smoothing degree to select, regardless of its ability to neatly partition a protein structure into backbone and side chain fragments. Rather, the lowest OF value was obtained at t = 0.9 bohr^2.



t	OF	#
0.05	4.34	101
0.50	9.75	87
0.80	8.87	64
0.90	8.79	55
1.00	9.08	33
1.10	9.70	31
1.20	10.00	30
1.30	10.25	25
1.40	13.21	23
1.50	13.28	23
1.60	13.22	22
1.70	13.06	20
1.80	13.21	19
1.90	13.37	19
2.00	13.59	19
2.50	18.70	17

Figure IV.6. Objective function (OF) for the ED-based CG models of HP7 *vs.* unsmoothed all-atom Amber MEP, as a function of the smoothing degree t. # stands for the number of local maxima observed in the PASA EDs.

At t = 0.9 bohr^2, there is a close connection between the ED peaks and protein atoms. Particularly, the backbone of the residues is systematically represented by two atoms, N and O, rather than being described by a single ED maximum located close to the backbone c.o.m. of the residue like at t = 1.4 bohr^2 (Table IV.IV).

Table IV.IV. Description of the 55 ED peaks obtained for the 12-residue peptide HP7 using a hierarchical merging/clustering algorithm applied to the PASA ED distribution function, at t = 0.9 bohr^2. For each point, the distance *vs*. the closest atom or c.o.m., *d*, is given in Å, and its charge, *q*, in e$^-$. *rmsdV* and *rmsdμ* are given in kcal/mol and D, respectively. Point number '#' refer to Figure IV.7. 'SCH' stands for the c.o.m. of a residue side chain.

#	Closest atom/c.o.m.		*d*	*q*	#	Closest atom/c.o.m.		*d*	*q*
1	N	Lys1	0.352	0.8150	29	O	Thr7	0.552	0.0080
2	Cα	Lys1	0.308	0.0900	30	Oγ	Thr7	0.340	0.0900
3	O	Lys1	0.548	0.1620	31	N	Gly8	0.415	-0.0710
4	Cβ	Lys1	0.225	-0.0180	32	O	Gly8	0.517	0.0720
5	Cγ	Lys1	0.227	0.0530	33	N	Lys9	0.394	-0.0950
6	Cδ	Lys1	0.226	0.0940	34	O	Lys9	0.530	0.1620
7	Nς	Lys1	0.282	0.8050	35	Cβ	Lys9	0.232	-0.0180
8	N	Thr2	0.397	-0.0970	36	Cγ	Lys9	0.212	0.0530
9	O	Thr2	0.535	0.0080	37	Cδ	Lys9	0.213	0.0940
10	Oγ	Thr2	0.362	0.0900	38	Nς	Lys9	0.291	0.8050
11	N	Trp3	0.409	-0.0990	39	N	Trp10	0.378	-0.0990
12	O	Trp3	0.547	0.0890	40	O	Trp10	0.550	0.0890
13	Cβ	Trp3	0.227	0.0320	41	Cβ	Trp10	0.223	0.0320
14	Nε1	Trp3	0.357	0.1530	42	Nε1	Trp10	0.357	0.1530
15	Cε3	Trp3	0.288	-0.0310	43	Cε3	Trp10	0.288	-0.0310
16	Cς2	Trp3	0.305	-0.0850	44	Cς2	Trp10	0.307	-0.0850
17	Cς3	Trp3	0.289	-0.0450	45	Cς3	Trp10	0.290	-0.0450
18	CH2	Trp3	0.293	-0.0140	46	CH2	Trp10	0.295	-0.0140
19	N	Asn4	0.351	-0.0700	47	N	Thr11	0.368	-0.0970
20	O	Asn4	0.515	0.0930	48	O	Thr11	0.532	0.0080
21	Oδ1	Asn4	0.581	-0.0230	49	Oγ	Thr11	0.356	0.0900
22	N	Pro5	0.027	0.0130	50	N	Glu12	0.395	-0.0190
23	O	Pro5	0.550	-0.1010	51	C	Glu12	0.506	-0.1230
24	Cβ	Pro5	0.420	0.0350	52	Cβ	Glu12	0.231	0.0670
25	SCH	Pro5	0.217	0.0530	53	Cδ	Glu12	0.565	-0.1010
26	N	Ala6	0.360	-0.0170	54	Cδ	Glu12	0.496	-0.8240
27	O	Ala6	0.543	0.0150	55	C	Glu12	0.532	-1.0000
28	N	Thr7	0.379	-0.0970					
							rmsdV		8.66
							rmsdμ		6.96

The three Thr residues are all described by three ED peaks located close to the O, N, and Oγ1 atoms. Also, both Trp residues of HP7 adopt a unique motif made of 8 points, located close to N and O for the backbone, and close to Cβ, Nε1, Cε3, Cζ2, Cζ3, and CH2 for the side chain (Table IV.IV and Figure IV.7). As already mentioned, the charges q that are reported in Table V.IV were calculated using Equation 11 of Section III. That 55 ED-point model is less efficient in approximating the all-atom unsmoothed Amber MEP, with $rmsdV$ = 8.66 kcal/mol and $rmsd\mu$ = 6.96 D, than the 48-point model built from the peaks and pits observed in the corresponding MEP smoothed at t = 1.35 bohr^2 (Table IV.III).

It will be further shown, in Section VI, that t = 0.9 bohr^2 is not a universal value to consider in the design of an ED-based electrostatic CG model, *i.e.*, that it is not valid for any protein structure.

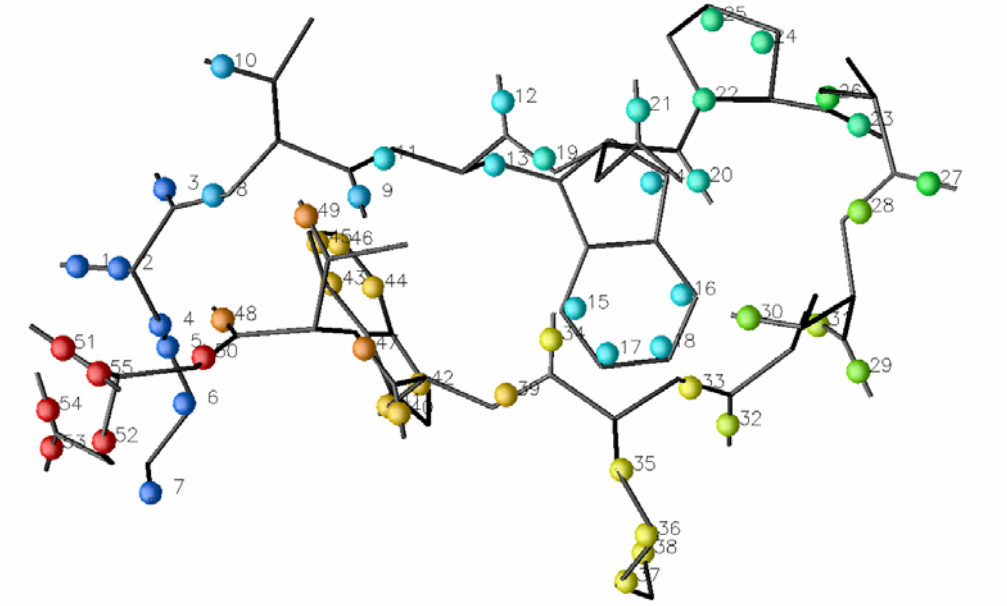


Figure IV.7. 3D structure of the 12-residue peptide HP7 (sticks) superimposed on the 55 ED peaks at t = 0.9 bohr^2 (color-coded by AA). CG points are numbered as in Table IV.IV. Figure was generated using DataExplorer [odx].

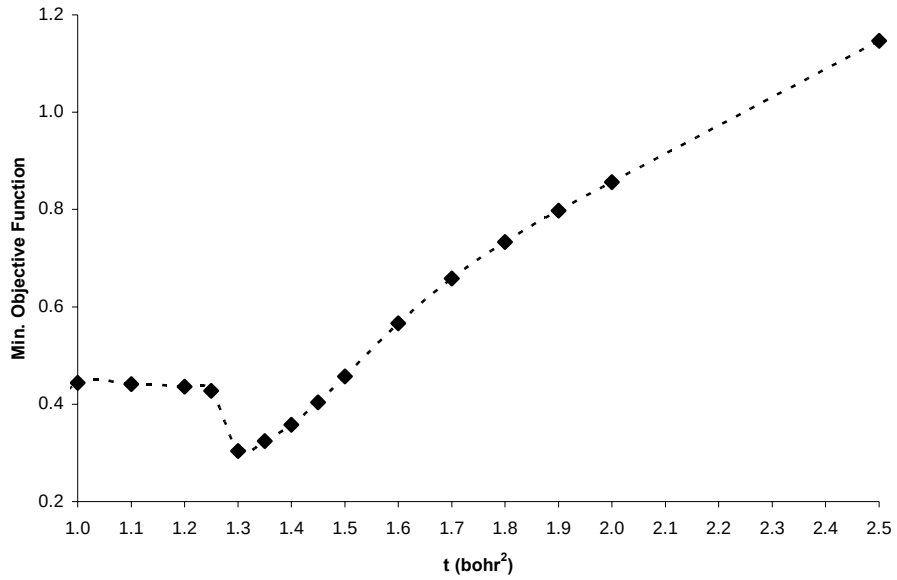
V. Extension to Other Force Fields – Application to the Gromos43A1 Set of Charges

In this Section, we present the results obtained using the set of charges implemented in the FF Gromos43A1 (Appendix II). All procedures were identical to those applied in Section IV, for the all-atom Amber FF. The presentation of the Section is thus very similar to that previous Section IV.

Selection of the Smoothing Degree

To select the optimal smoothing degree, we have used the charge fitting algorithm QFIT [bor] and applied it to each set of peaks and pits obtained for the β-Gly_{15} structure at various smoothing levels. The resulting Minimal Objective Function (MOF) is displayed in Figure V.1. The best fit is obtained at t = 1.3 $bohr^2$, with MOF = 0.304. As shown by the values reported in Figure V.1, the MOF values are well below the corresponding values obtained with the FF Amber (Figure IV.2). This appears to be due to the fact that Gromos43A1 is a CG-type FF itself, as detailed in Appendix II. Indeed, most of the atoms in alkyl groups, for instance, have a nul electric charge. The model obtained for β-Gly15 at t = 1.3 $bohr^2$ contains 32 CGs (Figure V.2), like for the Amber FF at t = 1.35 $bohr^2$ (Figure IV.3). In this sense, the application of a smoothing algorithm to the MEP function thus tends to level out differences between all-atom and united-atom FF. As shown in Figure V.1, above t = 1.3 $bohr^2$, the fit is less and less efficient due to a reduction in the number of CGs and a progressive change in their location with respect to the original structure.

As described in the previous Sections, CG descriptions and charges were established for each of the AA residues. Charge values obtained for the CG descriptions at t = 1.3 $bohr^2$ are reported in Table V.II and corresponding CG structures are shown in Figure V.3. Two charge sets are again presented for the CG representation of Asn. The first one was established from the MEPs of three rotamers. The second one was obtained by considering the two rotamers that are described by a common CG motif.



t	MOF	#
0.05	0.14	59
0.50	0.20	59
0.80	0.33	45
0.90	0.34	45
1.00	0.44	43
1.10	0.44	43
1.20	0.44	43
1.25	0.43	42
1.30	0.30	32
1.35	0.32	31
1.40	0.36	31
1.50	0.46	31
1.60	0.57	31
1.70	0.66	31
1.80	0.73	31
1.90	0.80	31
2.00	0.86	31
2.50	1.15	31

Figure V.1. Minimal objective function (MOF) for the charge fitting of ß-Gly_{15} MEP CGs from unsmoothed Gromos43A1 MEP, as a function of the smoothing degree *t*. # stands for the number of local minima and maxima observed in the MEPs.

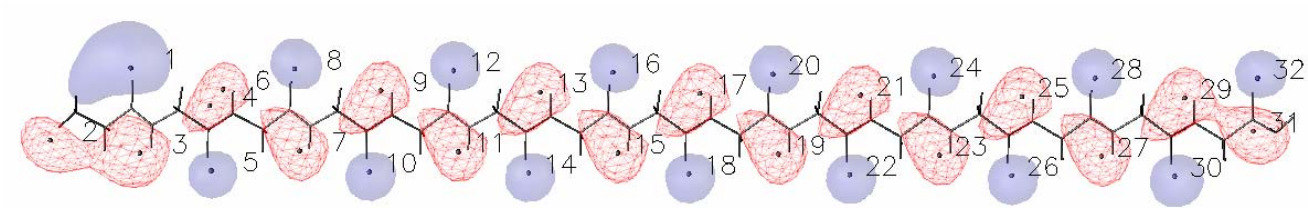


Figure V.2. Gromos43A1 MEP iso-contours (blue: -0.03 ; red: 0.03 e^-/bohr) of ß-Gly_{15} smoothed at t = 1.3 bohr^2. Local maxima and minima at t = 1.3 bohr^2 were obtained using the hierarchical merging/clustering algorithm applied to the Gromos43A1 MEP function. CG points are numbered as in Table V.I. Figure was generated using DataExplorer [odx].

Table V.I. CG charges $q_{0.0}$ (in e^-) of ß-Gly_{15} fitted from the Gromos43A1 MEP grids smoothed at t = 0.0 $bohr^2$, respectively, using the program QFIT. Local maxima and minima at t = 1.3 $bohr^2$ were obtained using the hierarchical merging/clustering algorithm applied to the Gromos43A1 MEP function. For each point, the distance *vs.* the closest atom, *d*, is given in Å. *rmsdV* and *rmsdμ* are given in kcal/mol and D, respectively. Point numbers (#) refer to Figure V.2.

#	Closest atom		*d*	$q_{0.0}$
1	O	Gly1	0.525	-0.2467
2	H	Gly1	0.678	0.0856
3	H	Gly2	0.552	0.1697
4	C	Gly2	1.028	0.0166
5	O	Gly2	0.559	-0.1834
6	H	Gly3	0.468	0.1859
7	H	Gly4	0.484	0.1775
8	O	Gly3	0.572	-0.1842
9	H	Gly5	0.490	0.1820
10	O	Gly4	0.569	-0.1807
11	H	Gly6	0.489	0.1801
12	O	Gly5	0.571	-0.1807
13	H	Gly7	0.492	0.1805
14	O	Gly6	0.570	-0.1805
15	H	Gly8	0.491	0.1804
16	O	Gly7	0.571	-0.1806
17	H	Gly9	0.491	0.1809
18	O	Gly8	0.571	-0.1806
19	H	Gly10	0.492	0.1804
20	O	Gly9	0.571	-0.1808
21	H	Gly11	0.490	0.1806
22	O	Gly10	0.571	-0.1802
23	H	Gly12	0.494	0.1804
24	O	Gly11	0.571	-0.1808
25	H	Gly13	0.492	0.1796
26	O	Gly12	0.571	-0.1798
27	H	Gly14	0.495	0.1795
28	O	Gly13	0.571	-0.1769
29	H	Gly15	0.457	0.1615
30	O	Gly14	0.580	-0.1809
31	C	Gly15	0.542	0.1646
32	O	Gly15	0.665	-0.1690
rmsdV				0.55
rmsdμ				0.11

Table V.II. CG charges (in e^-) for the AA residues obtained through a charge fitting algorithm using unsmoothed Gromos43A1 MEP grids. CG locations were generated at t = 1.3 bohr^2 using a hierarchical merging/clustering algorithm applied to the Gromos43A1 MEP function. *g* and *t* stand for *gauche* and *trans*, respectively (see [sim08] for details). *rmsdV* and *rmsdμ* are given in kcal/mol and D, respectively. Point numbers refer to Figure V.3.

	Conformation	Point 17	Point 18	Point 33	Point 34	Point 35	Point 36	Point 37	Point 38	*rmsdV*	*rmsdμ*
Ala		H	O								
		0.1821	-0.1815							0.57	0.12
Arg		H	O	Cζ							
	g-,t,g-,g-									1.62	0.94
	g-,t,g-,t	0.1867	-0.1785	0.9836						1.83	0.96
	g-,t,g+,t									1.79	1.02
	g-,t,t,t									1.81	0.95
Asn		H	O	$H\delta_{tr}$	Oδ	$H\delta_{cis}$					
	t,Nt									1.90	3.81
	t, Og-	0.2362	-0.1560	0.0402	-0.1095	-0.0148				2.41	3.85
	t, Og+									2.38	3.08
Asn		H	O	$H\delta_{tr}$	Oδ	$H\delta_{cis}$					
	t, Og-	0.1773	-0.1696	0.1023	-0.1933	0.0847				1.13	0.31
	t, Og+									1.03	0.30
Asp		H	O	Oδ1	Oδ2						
	t,g+	0.1575	-0.1799	-0.4875	-0.4901					1.26	0.08
Cys		H	O	Sγ							
	g-									0.61	0.12
	g+	0.1993	-0.1933	-0.0036						0.68	0.59
	t									0.59	0.08
Cys		H	O								
	g-									0.61	0.12
	g+	0.1972	-0.1942							0.66	0.53
	t									0.59	0.12
Gln		H	O	H_{tr}	Oε	H_{cis}					
	g-,t,Nt									0.64	0.13
	g-,t,Og-	0.1918	-0.1801	0.1075	-0.2119	0.0933				0.61	0.13
	g-,t,Og+									0.60	0.15
Glu		H	O	Oε1	Oε2						
	g-,t,g-	0.1800	-0.1818	-0.4948	-0.4993					0.98	0.09
	g-,t,g+									1.00	0.12
Gly		H	O								
		0.1809	-0.1806							0.55	0.11
His		H	O	Nδ	Hε						
	g-,Ng-	0.1769	-0.1731	-0.2648	0.2617					0.66	0.08
	t,Ng+									0.64	0.16
Ile		H	O								
	g-,g-	0.1828	-0.1812							0.55	0.11

	g-,t									0.55	0.11
	g+,t									0.55	0.11
Leu		H	O								
	g-,t	0.1837	-0.1817							0.55	0.12
	t,g+									0.55	0.12
Lys		H	O	Nζ							
	g-,g-,t,g-									0.69	0.28
	g-,g-,t,g+	0.1796	-0.1811	1.0014						0.69	0.39
	g-,t,t,g-									0.70	0.25
	g-,t,t,g+									0.70	0.27
Met		H	O								
	g-,g-,g-									0.55	0.11
	g-,g-,t									0.55	0.11
	g-,t,g-	0.1834	-0.1817							0.55	0.11
	g-,t-g+									0.55	0.11
	g-,t,t									0.55	0.11
Phe		H	O	6-ring	Hε1	Hζ	Hε2				
	g-,g-	0.1981	-0.1788	-0.1102	0.0370	0.0167	0.0414			0.68	0.19
	t,g+									0.68	0.29
Phe		H	O	6-ring							
	g-,g-	0.1812	-0.1886	0.0083						0.88	0.31
	t,g+									0.90	0.44
Pro		H	O								
	g+	0.1872	-0.1843							0.63	0.16
Ser		H	O	Oγ	Hγ						
	g-	0.1963	-0.1761	-0.1636	0.1472					0.68	0.34
	g+									0.74	0.61
Thr		H	O	Oγ	Hγ						
	g-	0.2073	-0.1755	-0.1727	0.1435					0.64	0.30
	g+									0.64	0.38
Trp		H	O	5-ring	6-ring	Hε1	HH	Hζ3	Hε3		
	g-,g-									0.71	0.24
	g-,t									0.75	0.15
	t,g-	0.1933	-0.1795	-0.1301	-0.1051	0.1444	0.0358	0.0409	0.0023	0.71	0.28
	t,g+									0.75	0.24
	t,t									0.74	0.16
Tyr		H	O	6-ring	OH	HH	Hδ*				
	g-,g-	0.2064	-0.1986	-0.0118	-0.1848	0.1550	0.0374			0.75	0.17
	t,g+									0.76	0.16
Val		H	O								
	g-	0.1831	-0.1815							0.55	0.11
	t									0.56	0.11

* Hδ stands on the opposite side of the O-H bond direction.

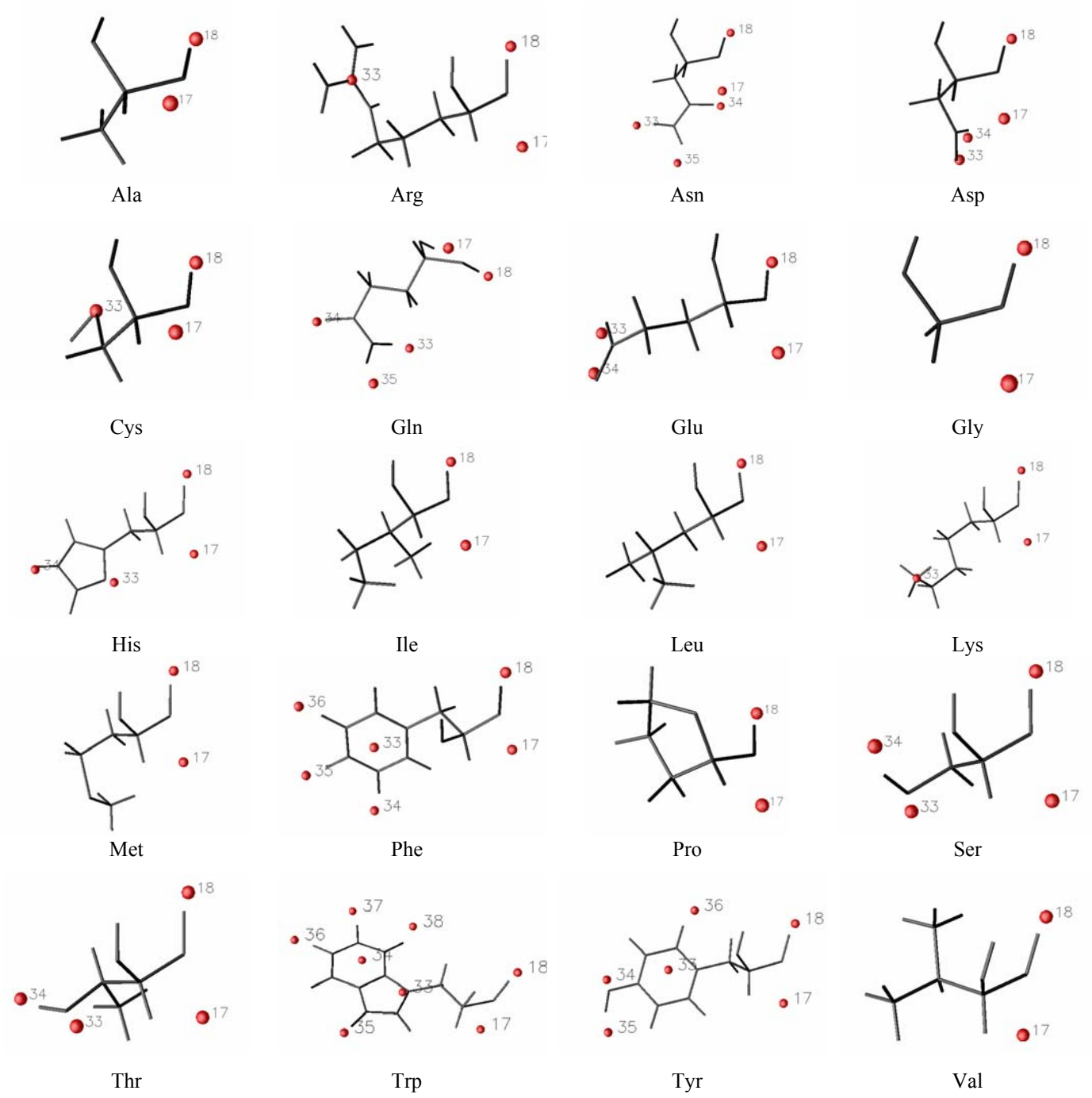


Figure V.3. CG model for each of the 20 AA residues as established at t = 1.3 bohr^2 from the hierarchical merging/clustering algorithm applied to the Gromos43A1 MEP function. CG points are numbered as in Table V.II. Figures were generated using DataExplorer [odx].

On the whole, the two-site CG description of each AA backbone differs from the one obtained with the all-atom Amber FF. Rather than being located along the C=O axis of a residue, it is displaced such as the positive charge is closer to the H atom the neighboring residue (Table V.I). Regarding side chain descriptions, alkyl chains such as Ala, Ile, Leu, and Val do not involve any

CG. This is also observed for Met. For Cys, that contains a sulfur atom like in Met, two models were tested; one involving a CG located on Sγ, and the other without any side chain CG. As shown by the *rmsdV* and *rmsdμ* values, the two models led to a very similar fitting quality. For Phe, the two models described in Table V.II include point 33 that is identically located in the 6-membered ring. The CG description of Arg differs from the Amber-based representation in that there is only one positive charge, initially located in the neighborhood of the atom Cς. We have simplified the Arg CG model by fixing that CG point exactly on Cς. Conversely, the side chain of Trp is described by 6 CGs with Gromos43A1, rather than 4 CG points with Amber.

A comparison of our two protein MEP-based CG models, generated from Amber and gromos43A1 sets of charges, with existing ones is reported in Table V.III. AA residues are listed according to their properties defined in the MARTINI FF [mon08], *i.e.*, hydrophobic residues, mainly classified as apolar, polar residues with or without H-bond forming characteristics, and charged side chains. Parallely, we report a description of the Basdevant's model [bas07] and our Amber- and Gromos43A1-based models. For residue Gly, there is a backbone CG representation only that consists, like for all other residues, either of one polar bead, one center, or two CGs with opposite charges, in the frames of the MARTINI FF, the Basdevant's model, and our own CG models, respectively.

The number of residue CGs in each model is variable. The number of grains in the Basdevant model is strongly dependent on the size of the side chains, but does not exceed two. For MARTINI, it is higher than two only for ring-shaped side chains, *i.e.*, Phe, His, Trp, and Tyr. In the case of our MEP-based CG representations, there might be only one CG for a residue as large as Phe, and there are up to three CG for Asn, a small residue. For all small hydrophobic residues, the MARTINI CG representations involve only one apolar grain. Parallely, in our MEP-based models, all CGs are located on C atoms, with small charge values $|q| < 0.07$ e^- using the Amber FF. With Gromos443A1, there is not any side chain CG, except for Pro, a particular case where there is only one CG on the backbone. For Phe, our models may involve a large number of points, *i.e.*, up to 4 and 5 for Gromos43A1 and Amber, respectively. The total charge brought by the side chain of Phe stays low, with $|q| < 0.03$ e^-. Sulfur-containing residues, that are hydrophobic and do not form any H-bond, are however characterized by a strong dipole. In MARTINI, they are thus represented by one CG with the intermediate apolar/polar state. In the Amber-based model, there is also only one CG, located on the atom Sγ, but with Gromos43A1, Cys (and Met) can be represented without any side chain CG. It is interesting to note that, with the Amber-based CG model, Cys differs largely

from Met through the strong charge separation brought by its backbone. Indeed, charge values q = 0.4033 and -0.2949 for Cys, and q = 0.2866 and -0.2358 e^- for Met (Table IV.II).

Table V.III. Descriptions of protein side chain CG models, in terms of the number and property of the CGs, as defined in the MARTINI FF [mon08], in the Basdevant's model [bas07], and as obtained from a hierarchical merging/clustering of MEP functions smoothed at t = 1.35 bohr^2 using Amber and t = 1.30 bohr^2 using Gromos43A1.

		MARTINI	Basdevant	Amber	Gromos43A1
	Gly	-	-	-	-
Small hydrophic residues					
	Ala	-	1	1 (on C)	-
	Ile	1 apolar	1	2 (on C)	-
	Leu	1 apolar	1	3 (on C)	-
	Pro	1 apolar	1	-	(a)
	Val	1 apolar	1	1 (on C) $\lvert q\rvert < 0.07$ e^-	-
Large hydrophobic residue					
	Phe	3 apolar	2	1 or 5 $\lvert q\rvert < 0.03$ e^-	1 or 4 $\lvert q\rvert < 0.02$ e^-
Sulfur-containing residues					
	Cys	1 apolar/polar	1	1 (on S)	0 or 1 (on S)
	Met	1 apolar/polar	2	1 (on S)	-
Polar amide-containing residues with H-bond property					
	Asn	1 polar	1	3	3
	Gln	1 polar	2	4	3
Small hydrophilic residues with OH group					
	Ser	1 polar	1	2	2
	Thr	1 polar	1	2	2
Ring-shape hydrophobic residue with H-bond property					
	His	1 apolar, 2 polar	2	2	2
	Trp	3 apolar, 1 polar	2	4	6
	Tyr	2 apolar, 1 polar	2	5	4
Charged residues					
	Arg	1 apolar/polar, 1 charged	2	3	1
	Asp	1 charged	1	2	2
	Glu	1 charged	2	2	2
	Lys	1 apolar, 1 charged	2	1	1

[a]There is one CG on the backbone and one on the side chain.

Regarding Asn and Gln, our MEP-based models provide a finer description of the side chains, with three grains located at the vicinity of the O and H atoms. They are thus somewhat closer to an all-atom representation. Within MARTINI, these side chains are represented using one grain characterized by a polar type with hydrogen bonding donor and acceptor character. For all residues containing an O-H group, *i.e.*, Ser, Thr, and Tyr, our models include at least two opposite charges located at the neighborhood of O and H; they correspond to one polar group in MARTINI. The side chain of His and Trp contain hydrophobic rings, but also hydrogen bonding properties. In the frame of our MEP-based models, they are represented by CGs with a strong dipole occurring between Hε

and Nδ in His, and Hε1 and the rings in Trp. The polarity property in MARTINI is thus expressed as a charge separation in our models. Finally, regarding the residues that are explicitely charges in the MARTINI FF, we observe a finer description of the negative Asp and Glu residues in our models, with two separate negative charges close to the O of the carboxyl group. The Amber-based CG model of Arg is rather interesting and original as it involves three positive charges almost symmetrically spread around the atom Cς. This could be seen as a description that is more consistent with a charge delocalization.

Application to 12-Residue β-Hairpin HP7

As reported in Section IV, the determination of the location of the side chain CG points was achieved by applying our merging/clustering procedure to the isolated AA in their PDB conformation while the backbone points were determined through a superimposition algorithm, QUATFIT [hei90]. The backbone CG description of HP7, generated for all residues but the last one, was then completed by adding unitary positive and negative charges on the N_{Lys1} and OXT_{Glu12} atoms, respectively. This procedure led to a 49-point model (Table V.IV. and Figure V.4), with 24 and 25 CGs describing the protein backbone and the side chains, respectively. Due to a change in the dipolar description of the AA backbones, there are two CGs less in the Gromos43A1 description than in the Amber one. There are however three extra CG needed to describe the side chain structure of HP7, *vs*. the Amber case, due to a different description of residues Trp3, Ala6, and Trp10 (Tables IV.III and V.IV). As for Amber, the best results were also obtained with the charges fitted on the two-rotamer model of residue Asn (Table V.II). By comparison with our model, the fitted charges on the 28-point Basdevant's c.o.m. model led, in the present case, to *rmsdV* = 5.98 kcal/mol and *rmsdμ* = 1.66 D, while the raw 49-CG model led to *rmsdV* = 2.70 and *rmsdμ* = 0.26 D. One will however notice, from Table V.IV, that a charge fitting slightly improves the 49-point MEP representation, but also slightly alters the dipolar value, with *rmsdμ* = 0.56 D. There is no big difference between the statistical quality of the two sets of charges reported in Table V.IV even if, locally, some charges are strongly altered. For instance, charge number 22 in Table V.IV, and as for Amber, the charge that is located in the neighborhood of atom Oγ of Thr7, and that adopts a positive value of 0.1501 e^- after fitting.

Table V.IV. 49-point CG model for the 12-residue peptide HP7 built from charges (in e^-) reported in Table V.II. *rmsdV* and *rmsdμ* are given in kcal/mol and D, respectively. Coordinates X, Y, and Z are in Å. Point numbers (#) refer to Figure V.4.

#	X	Y	Z	CG location	Residue	Model charges	Optimized charges
1	-7.425	4.328	-2.377	H	Lys1	0.1796	0.6554
2	-3.232	4.602	-2.326	O	Lys1	-0.1811	-0.2363
3	-3.022	1.264	-1.019	H	Thr2	0.2073	0.2090
4	0.122	1.466	-2.278	O	Thr2	-0.1755	-0.4208
5	-0.087	4.663	-0.656	H	Trp3	0.1933	0.2548
6	2.859	3.966	-1.180	O	Trp3	-0.1795	-0.1894
7	3.839	0.540	-1.628	H	Asn4	0.1773	0.0882
8	4.975	2.973	-0.077	O	Asn4	-0.1696	0.0351
9	7.450	1.749	-2.373	H	Pro5	0.1872	0.2057
10	5.218	0.418	0.169	O	Pro5	-0.1843	-0.1970
11	8.651	-0.633	0.219	H	Ala6	0.1821	0.1403
12	4.875	-0.443	-1.624	O	Ala6	-0.1815	-0.1921
13	6.085	-3.687	-2.578	H	Thr7	0.2073	0.1160
14	2.885	-1.233	-2.413	O	Thr7	-0.1755	-0.1867
15	2.775	-3.058	-5.503	H	Gly8	0.1809	0.2007
16	-1.176	-2.906	-3.553	O	Gly8	-0.1806	-0.2357
17	0.079	0.014	-1.879	H	Lys9	0.1796	0.1974
18	-3.516	0.084	-1.091	O	Lys9	-0.1811	-0.3598
19	-3.076	-3.470	-0.821	H	Trp10	0.1933	0.7349
20	-5.653	-2.444	1.985	O	Trp10	-0.1795	-0.2261
21	-6.575	0.298	-0.143	H	Thr11	0.2073	0.3602
22	-9.426	1.989	-1.495	O	Thr11	-0.1755	-0.8977
23	-7.612	-2.864	3.802	N	Lys1	1.0000	1.2064
24	-8.558	-2.226	-4.238	OXT	Glu12	-1.0000	-1.0357
25	-5.881	5.232	0.580	Nζ	Lys1	1.0014	0.9102
26	-6.116	3.282	2.049	Oγ	Thr2	-0.1727	-0.2059
27	0.620	3.090	-4.787	Hγ	Thr2	0.1435	0.1294
28	1.151	0.596	-5.848	5-ring	Trp3	-0.1301	-0.1108
29	3.959	3.675	-5.274	6-ring	Trp3	-0.1051	-0.1644
30	3.026	-1.702	-7.405	Hε1	Trp3	0.1444	0.1524
31	-0.370	-2.013	-6.602	HH	Trp3	0.0358	0.0632
32	-1.949	0.072	-5.140	Hζ3	Trp3	0.0409	0.0563
33	3.148	-1.608	1.341	Hε3	Trp3	0.0023	0.1149
34	2.602	2.555	2.469	$H\delta_{tr}$	Asn4	0.1023	-0.1091
35	4.212	-0.247	3.512	Oδ	Asn4	-0.1933	-0.2017
36	3.580	-1.748	0.915	$H\delta_{cis}$	Asn4	0.0847	0.0825
37	2.878	-3.179	-0.714	Oγ	Thr7	-0.1727	0.1501
38	0.908	-4.071	-6.309	Hγ	Thr7	0.1435	0.1011
39	-3.550	-0.276	-5.083	Nζ	Lys9	1.0014	0.9857
40	-4.321	2.339	-5.305	5-ring	Trp10	-0.1301	-0.1352
41	-2.506	0.118	-8.346	6-ring	Trp10	-0.1051	-0.1586
42	-4.715	5.368	-6.775	Hε1	Trp10	0.1444	0.1249
43	-5.748	4.544	-3.567	HH	Trp10	0.0358	0.0434
44	-5.029	1.324	-2.165	Hζ3	Trp10	0.0409	0.0297
45	-3.601	1.423	1.674	Hε3	Trp10	0.0023	-0.0859
46	-5.633	0.600	2.311	Oγ	Thr11	-0.1727	-0.1623
47	-10.070	-0.652	-0.934	Hγ	Thr11	0.1435	0.1173
48	-10.277	-1.550	-2.002	Oε1	Glu12	-0.4971	-0.6338
49	-7.425	4.328	-2.377	Oε2	Glu12	-0.4971	-0.3198
					Total charge	1.0206	1.0000
					rmsdV	2.70	1.73
					rmsdμ	0.26	0.56

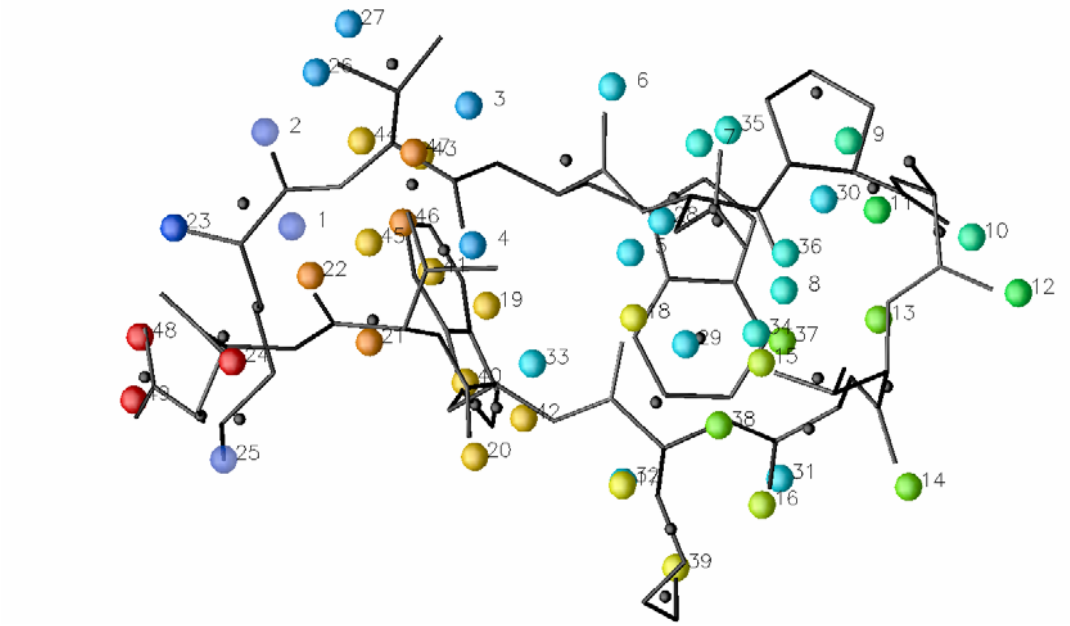
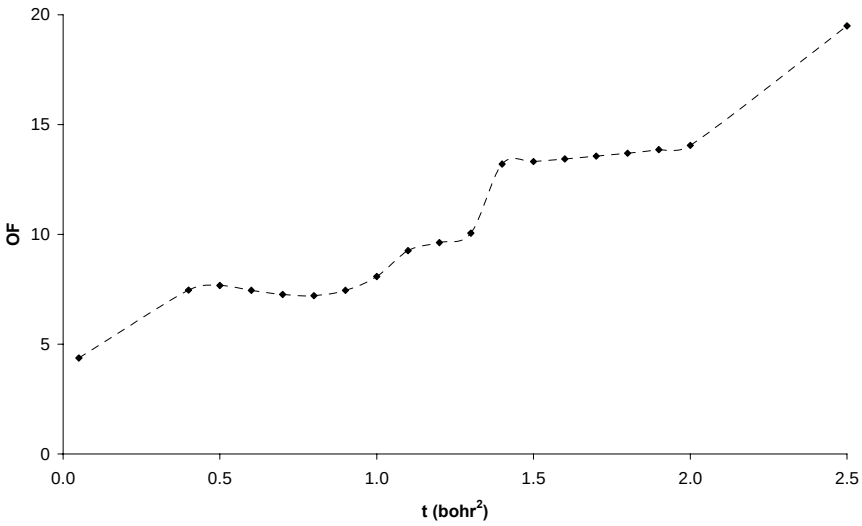


Figure V.4. 3D structure of the 12-residue peptide HP7 (sticks) superimposed on the 49 Gromos43A1 CG point charges at t = 1.3 bohr^2 (color-coded by AA), and the 28-point Basdevant's model (black spheres). CG points are numbered as in Table V.IV. Figure was generated using DataExplorer [odx].

In a further stage, we have deepened the study of CG models generated through the analysis of smoothed PASA ED distribution functions. In Figure V.5, we report the evolution of the OF calculated from the peak charges obtained using Equation 11 of Section III and the Gromos43A1 charges (Appendix II).



t	OF	#
0.05	4.37	101
0.50	7.67	87
0.80	7.21	64
0.90	7.45	55
1.00	8.08	33
1.10	9.26	31
1.20	9.63	30
1.30	10.05	32
1.40	13.20	23
1.50	13.32	23
1.60	13.43	22
1.70	13.56	20
1.80	13.70	19
1.90	13.86	19
2.00	14.05	19
2.50	19.49	17

Figure V.5. Objective function (OF) for the ED-based CG models of HP7 *vs*. unsmoothed Gromos43A1 MEP, as a function of the smoothing degree t. # stands for the number of local maxima observed in the PASA EDs.

Figure V.5 shows that t = 1.4 bohr^2 is not the best smoothing degree to select, as already concluded in Section IV. Besides the ED-based models obtained at very low values of t, *i.e.*, models similar to the original all-atom structure, the lowest OF value was obtained at t = 0.8 bohr^2, and the resulting ED-peak model consists of 64 points rather than 23 at t = 1.4 bohr^2. The 64 points observed at t = 0.8 bohr^2 are described in Table V.V and Figure V.6.

Table V.V. Description of the 64 ED peaks obtained for the 12-residue peptide HP7 using a hierarchical merging/clustering algorithm applied to the PASA ED distribution function at t = 0.8 bohr^2. For each point, the distance *vs.* the closest atom or c.o.m., d, is given in Å, and its charge, q, in e$^-$. Point numbers '#' refer to Figure V.6. 'SCH' stands for the c.o.m. of a residue side chain.

#	Closest atom		d	q	#	Closest atom		d	q
1	N	Lys1	0.235	0.8150	33	O	Thr7	0.417	0.0080
2	Cα	Lys1	0.247	0.0900	34	Oγ	Thr7	0.246	0.0900
3	O	Lys1	0.415	0.1620	35	N	Gly8	0.237	-0.0710
4	Cβ	Lys1	0.169	-0.0180	36	O	Gly8	0.393	0.0720
5	Cγ	Lys1	0.183	0.0530	37	N	Lys9	0.235	-0.1850
6	Cδ	Lys1	0.179	0.0940	38	Cα	Lys9	0.291	0.0900
7	Nς	Lys1	0.204	0.8050	39	O	Lys9	0.400	0.1620
8	N	Thr2	0.233	0.0100	40	Cβ	Lys9	0.178	-0.0180
9	Cα	Thr2	0.255	-0.1070	41	Cγ	Lys9	0.169	0.0530
10	O	Thr2	0.404	0.0080	42	Cδ	Lys9	0.168	0.0940
11	Oγ	Thr2	0.261	0.0900	43	Nς	Lys9	0.210	0.8050
12	N	Trp3	0.243	-0.0990	44	N	Trp10	0.229	-0.0990
13	O	Trp3	0.412	0.0890	45	O	Trp10	0.414	0.0890
14	Cβ	Trp3	0.180	0.0320	46	Cβ	Trp10	0.176	0.0320
15	Nε1	Trp3	0.277	0.0630	47	Nε1	Trp10	0.276	0.0630
16	SCH	Trp3	0.151	0.0900	48	SCH	Trp10	0.177	0.0900
17	Cε3	Trp3	0.228	-0.0310	49	Cε3	Trp10	0.227	-0.0310
18	Cς2	Trp3	0.239	-0.0850	50	Cς2	Trp10	0.239	-0.0850
19	Cς3	Trp3	0.231	-0.0450	51	Cς3	Trp10	0.232	-0.0450
20	CH2	Trp3	0.235	-0.0140	52	CH2	Trp10	0.236	-0.0140
21	N	Asn4	0.223	-0.0700	53	N	Thr11	0.226	0.0100
22	O	Asn4	0.390	0.0930	54	Cα	Thr11	0.263	-0.1070
23	Cβ	Asn4	0.211	-0.0080	55	O	Thr11	0.401	0.0080
24	Oδ1	Asn4	0.440	-0.0150	56	Oγ	Thr11	0.256	0.0900
25	N	Pro5	0.020	0.0130	57	N	Glu12	0.222	-0.1160
26	O	Pro5	0.415	-0.1010	58	Cα	Glu12	0.281	0.0970
27	Cβ	Pro5	0.290	0.0350	59	O	Glu12	0.447	-0.1230
28	Cγ	Pro5	0.298	0.0530	60	Cβ	Glu12	0.176	0.0670
29	N	Ala6	0.225	-0.0170	61	Cγ	Glu12	0.176	-0.0420
30	O	Ala6	0.411	0.0150	62	Oε2	Glu12	0.446	-0.0590
31	N	Thr7	0.230	0.0100	63	Oε1	Glu12	0.436	-0.8240
32	Cα	Thr7	0.227	-0.1070	64	OXT	Glu12	0.447	-1.0000
							rmsdV		7.10
							rmsdμ		6.07

The mean distance between the ED peaks and their closest atom or c.o.m. is equal to 0.270 Å, a value that is expectedly lower than the corresponding average distance calculated for ED peaks obtained at a higher t = 0.9 bohr^2, *i.e.*, 0.372 Å (Table IV.IV). The charges q that are reported in Table V.V were calculated using Equation 11 of Section III. That 64 ED-point model is less efficient than the 49-point model built from the peaks and pits observed in the MEP smoothed at t = 1.3 bohr^2 (Table V.III), with *rmsdV* = 7.10 kcal/mol and *rmsdμ* = 6.07 D.

As already mentioned for the Amber FF (Section IV) it will be shown in the next Section that t = 0.8 bohr^2 is not a universal value to consider in the design of an ED-based electrostatic CG model and is not valid for all protein structures.

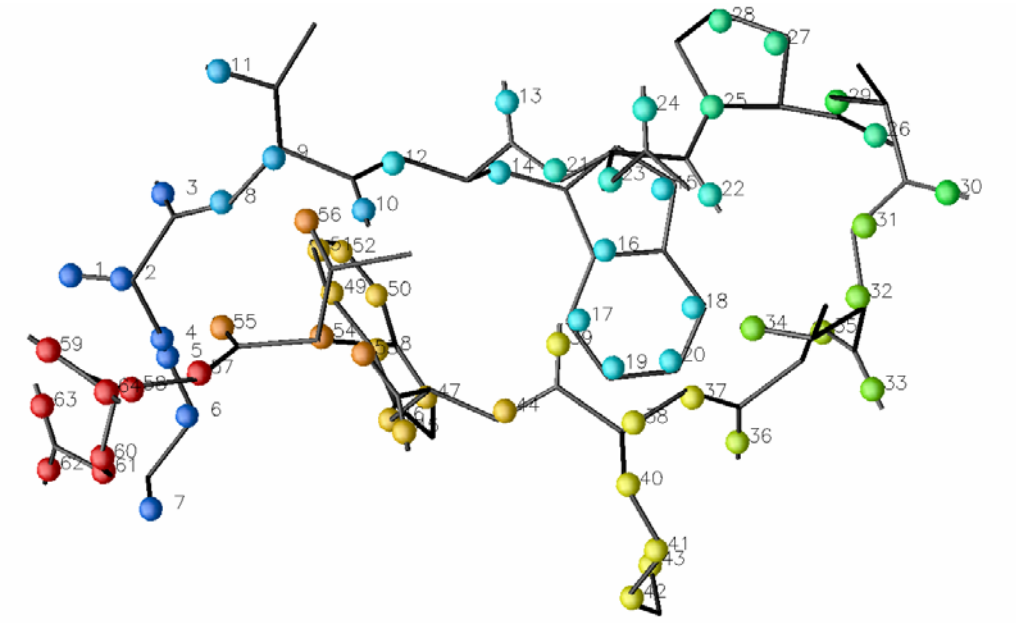


Figure V.6. 3D structure of the 12-residue peptide HP7 (sticks) superimposed on the 64 ED peaks at t = 0.8 bohr^2 (color-coded by AA). CG points are numbered as in Table V.IV. Figure was generated using DataExplorer [odx].

VI. Automation of the CG Generation Procedure – Application to the Potassium Ion Channel KcsA

In a further work to study systems that are larger than polypeptides, an automation stage was carried out to avoid the lengthy "manual" generation of the AA CGs. The resulting automated procedure was fully based on the application of a superimposition algorithm of CG motif templates onto the AA structures of the large protein under study in this Section VI. As already mentioned, we used the program QUATFIT [hei90] to superimpose a limited set of atoms from the template on the studied structure, and then used the resulting transformation matrix to generate the CG coordinates. The program, written in Fortran90, simply consists in the generation of reference and fitted molecular files, using the PDB coordinates and a table of template coordinates, respectively. A call to the program QUATFIT was implemented as:

```
call system('./quatfit.sh')
```

where "quatfit.sh" is a script file calling the executable "quatfit":

```
./quatfit -r refmol -f fitmol -p pairs
```

wherein "refmol" and "fitmol" stand for the reference PDB atom and the template atom files, respectively. "pairs" is a file that contains the mapping between the atoms of the two coordinate files. Examples are provided in Table VI.1.

The templates that were selected in this study are described in Tables VI.II and VI.III for the two force fields Amber and Gromos43A1, respectively. Their size consisted of at least three atoms so as to generate unique superposition results. For rigid side chains, such as His, Phe, and Trp, more than three atoms were used to better fit the whole side chain plane. For Arg and Gln, within the frame of the Amber FF only, more than 3 atoms were used too to generate, at once, all CGs. For the AA residues that are not reported in Tables VI.II and VI.III, the CG coordinates were directly obtained from the side chain atom coordinates as specified in Section IV and V (Tables IV.II and

V.II). Thus, due to differences in the Amber- and Gromos43A1-based electrostatic CG models, templates presented in Tables VI.II and VI.III also differ.

Table VI.1. Examples of coordinate files needed to generate backbone CGs and charges (in e^-) for Ala in the frame of the Amber FF, using the program QUATFIT. The last column contains the point charge values. A value that differs from 0.0 is given only for CG points.

```
refmol:     3

            C       66.14800    24.07500    47.53400   0.0000
            O       65.32700    24.91600    47.90200   0.0000
            N       66.83700    24.17600    46.40100   0.0000
fitmol:     5

            C       22.57500    13.92300     2.13100   0.0000
            O       23.02100    13.16700     2.99300   0.0000
            N       23.31800    14.68800     1.34500   0.0000
            PT17    22.10169    14.44978     1.78587   0.2349
            PT18    23.28951    12.82598     3.41467  -0.2399
pairs:      3
            1   1   1.0
            2   2   1.0
            3   3   1.0
```

The protein that was selected to test our new procedure is the KcsA potassium channel (Figure VI.1), a transmembrane protein structure that is commonly used to model biological ion channels [gas06, boi07, nos07, war07, pic08]. It is formed by four identical chains, each chain containing two α-helices connected by a loop located in the channel region (Figure VI.1.b). The channel consists of a 15 Å long narrow gating pore opened towards the intracellular region, a larger cavity of about 10 Å, and the so-called selectivity filter, that is about 18 Å long, pointing to the extracellular 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. They build a structure that is similar to a water solvation shell around a K^+ ion. Their role is to remove the hydration shell from K^+ when it enters the selectivity filter. Potassium and other ion channels are known to switch between closed and open states [jia02]. We will however restrict our studies to the opened state. A closed configuration can be found in the PDB under the access code 1K4C, but only the Cα coordinates are available. Let us finally mention that Roderick MacKinnon and Peter Agre were awarded the Nobel Prize in 2003 for "discoveries concerning channels in cell membranes", and more specifically, R. MacKinnon, for "structural and mechanistic studies of ion channels".

In the present work, the 3D model of the entire protein was prepared according to the X-ray crystal structure of the KcsA K^+ channel (PDB access code 1BL8) by adding missing side chain

atoms using the program SwissPDBViewer [gue97]. The addition of H atoms and the design of the histidine residues into a Hisε configuration were then achieved with the program VEGA ZZ [ped04]. The three K^+ ions, labelled K401, K402, and K403, were deleted. After the addition of unitary charges on the N and OXT atoms of the end residues of each of the four monomers, the application of our automated procedure finally led to the generation of 1492 and 1176 CGs, in the frame of the Amber and Gromos43A1 FF, respectively, for an original structure of 5888 atoms. Those reduction ratio correspond to the 4:1 value reported by Bond and Sansom [bon06, bon07] who studied the interaction of membrane proteins with lipid molecules through MD. A visualization of the *rmsd* values obtained between the atoms of the templates and the coresponding atoms of the protein crystal structure, for each of the superimposition achieved using QUATFIT [hei90] during the CG generation, is presented in Figure VI.2.

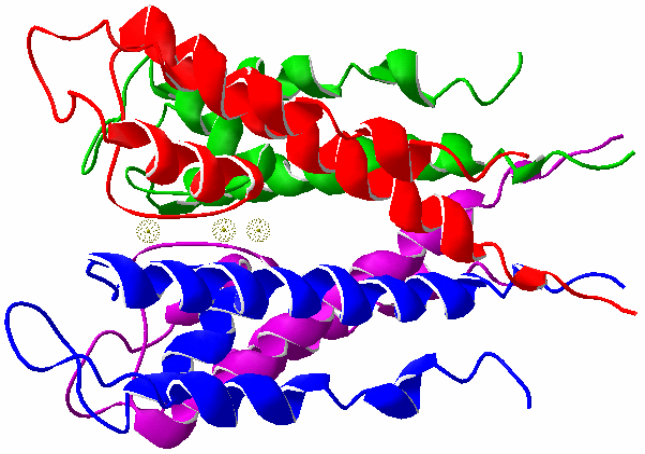


Figure VI.1.a. 3D conformation and secondary structure of the potassium channel KcsA (PDB code 1BL8). The four monomers are displayed using different colors. The three K^+ ions are displayed using dotted spheres. Figure was generated using SwissPDBViewer [gue97].

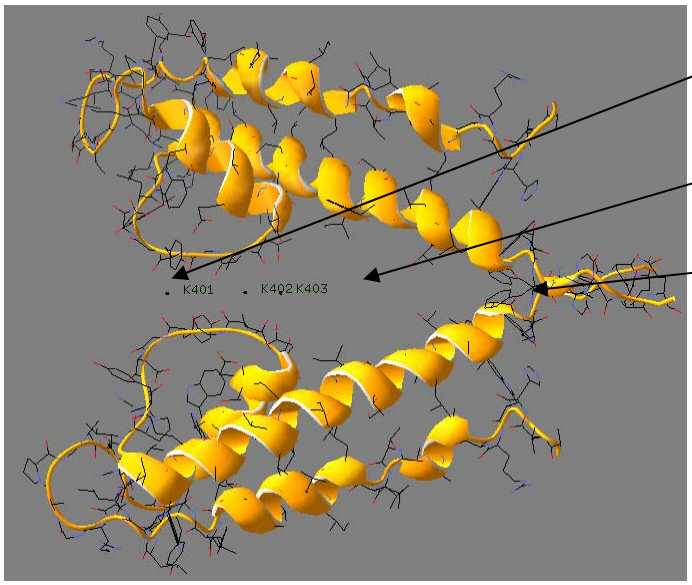


Figure VI.1.b. 3D conformation and secondary structure of the potassium channel KcsA (PDB code 1BL8). Two monomers, chains A and C, are displayed. Figure was rotated by 45° around the channel axis *vs.* Fig. VI.1.a. Figure was generated using SwissPDBViewer [gue97].

Table VI.II. Atom and CG template coordinates (in Å) and charges (in e^-) as used for the Amber-based CG generation.

	X	Y	Z	Charge		X	Y	Z	Charge
Backbone									
C	22.575	13.923	2.131						
O	23.021	13.167	2.993						
N	23.318	14.688	1.345						
PT17	22.102	14.450	1.786	q17[a]					
PT18	23.290	12.826	3.415	q18[a]					
Side Chain									
ARG					**PHE**				
Cδ	18.357	15.973	3.895		Cγ	18.926	14.982	2.956	
Nε	19.007	17.303	3.901		Cδ1	18.399	15.648	1.894	
Cς	18.693	18.288	4.755		Cε1	16.993	15.816	1.788	
NH1	17.715	18.110	5.653		Cς	16.173	15.310	2.748	
NH2	19.359	19.450	4.710		Cε2	16.700	14.643	3.810	
PT33	20.133	18.289	3.372	0.2807[b]	Cδ2	18.106	14.476	3.916	
PT34	17.392	16.471	5.163	0.3162	PT33	17.174	15.244	2.693	-0.0034
PT35	18.264	19.822	6.034	0.2807[b]	**SER**				
ASN					Cβ	20.443	14.916	2.987	
Cγ	21.154	16.268	3.071		Oγ	19.047	15.112	2.779	
Oδ1	22.355	16.412	3.224		Hγ	18.568	14.235	2.826	
Nδ2	20.298	17.275	2.917		PT33	18.900	15.724	2.684	-0.1765
PT33	18.533	16.790	3.100	0.1034	PT34	18.430	13.186	2.874	0.1546
PT34	22.869	16.602	3.382	-0.2316	**THR**				
PT35	20.271	19.098	2.484	0.0689	Cβ	20.358	14.870	2.971	
ASP					Oγ1	19.001	14.913	2.536	
Cγ	21.094	16.293	3.152		Hγ1	18.418	15.260	3.270	
Oδ1	20.959	17.047	2.164		PT33	18.997	14.681	1.742	-0.1572
Oδ2	21.670	16.583	4.223		PT34	18.141	15.562	4.127	0.1229
PT33	21.836	16.816	4.381	-0.3884[c]	**TRP**				
PT34	21.030	17.385	2.074	-0.3884[c]	Cγ	18.914	15.167	2.725	
GLN					Cδ1	17.884	14.443	3.185	
Cγ	18.930	15.114	2.699		Nε1	16.676	14.963	2.769	
Cδ	18.288	16.002	3.767		Cε2	16.950	16.087	1.998	
OE1	17.102	16.285	3.744		Cς2	16.065	16.958	1.352	
NE2	19.135	16.423	4.701		CH2	16.638	18.008	0.644	
PT33	20.182	15.158	5.396	0.1679	Cς3	18.021	18.142	0.611	
PT34	19.052	15.204	3.242	0.0013	Cε3	18.919	17.280	1.251	
PT35	16.582	16.448	3.815	-0.2615	Cδ2	18.322	16.220	1.965	
PT36	19.842	18.076	5.102	0.0837	PT33	18.085	14.932	2.668	-0.1387
GLU					PT34	17.376	17.628	0.881	-0.0996
Cδ	18.288	16.002	3.767		PT35	15.118	14.205	3.123	0.1487
OE1	17.754	17.063	3.377		PT36	15.122	19.569	-0.413	0.0294
OE2	18.345	15.599	4.949		**TYR**				
PT33	18.226	15.738	5.228	-0.4581[c]	Cς	16.164	15.242	2.786	
PT34	17.596	17.297	3.536	-0.4581[c]	OH	14.815	15.390	2.691	
HIS					HH	14.573	15.752	1.791	
Cγ	18.921	15.161	2.698		PT33	17.466	14.781	3.416	-0.1116
Nδ1	18.437	15.752	1.543		PT34	14.804	15.233	3.290	-0.1318
Cε1	17.116	15.812	1.626		PT35	14.604	16.119	0.988	0.1591
Nε2	16.744	15.267	2.820		PT36	19.079	13.422	5.428	0.0238
Cδ2	17.833	14.874	3.469		PT37	15.691	13.419	5.482	0.0378
PT33	15.258	15.134	3.239	0.1790					
PT34	18.481	15.611	1.258	-0.1845					

[a]values of q17 and q18 depend on the AA type (Table IV.II)
[b]mean value for point charges 33 and 35 (Table IV.II)
[c]mean value for point charges 33 and 34 (Table IV.II)

Table VI.III. Atom and CG template coordinates (in Å) and charges (in e^-) as used for the Gromos43A1-based CG generation.

	X	Y	Z	Charge		X	Y	Z	Charge
Backbone									
C	22.575	13.923	2.131						
O	23.021	13.167	2.993						
N	23.318	14.688	1.345						
PT17	22.102	14.450	1.786	q17[a]					
PT18	23.290	12.826	3.415	q18[a]					
Side Chain									
ASN					**SER**				
Cγ	21.154	16.268	3.071		Cβ	20.443	14.916	2.987	
Oδ1	22.355	16.412	3.224		Oγ	19.047	15.112	2.779	
Nδ2	20.298	17.275	2.917		Hγ	18.568	14.235	2.826	
PT33	18.471	17.151	3.053	0.1023	PT33	19.007	15.708	2.646	-0.1636
PT34	22.736	16.552	3.393	-0.1933	PT34	18.524	13.120	2.804	0.1472
PT35	20.361	19.068	2.506	0.0847	**THR**				
ASP					Cβ	20.358	14.870	2.971	
Cγ	21.094	16.293	3.152		Oγ1	19.001	14.913	2.536	
Oδ1	20.959	17.047	2.164		Hγ1	18.418	15.260	3.270	
Oδ2	21.670	16.583	4.223		PT33	19.019	14.657	1.736	-0.1727
PT33	21.108	17.162	2.514	-0.4888[b]	PT34	18.070	15.546	4.019	0.1435
PT34	21.621	16.812	3.960	-0.4888[b]	**TRP**				
GLN					Cγ	18.914	15.167	2.725	
Cδ	18.288	16.002	3.767		Cδ1	18.147	14.690	1.736	
Oε1	17.102	16.285	3.744		Nε1	16.840	15.119	1.858	
Nε2	19.135	16.423	4.701		Cε2	16.769	15.917	2.994	
PT33	20.260	15.383	5.718	0.1075	Cς2	15.668	16.584	3.544	
PT34	16.720	16.401	3.814	-0.2119	CH2	15.904	17.316	4.702	
PT35	19.762	18.085	5.232	0.0933	Cς3	17.183	17.352	5.245	
GLU					Cε3	18.294	16.690	4.708	
Cδ	18.288	16.002	3.767		Cδ2	18.038	15.953	3.534	
Oε1	17.754	17.063	3.377		PT33	18.651	15.285	2.695	-0.1301
Oε2	18.345	15.599	4.949		PT34	17.190	16.749	4.345	-0.1051
PT33	18.179	15.929	4.817	-0.4971[b]	PT35	15.681	14.642	1.003	0.1444
PT34	17.771	16.939	3.723	-0.4971[b]	PT36	14.283	18.275	5.502	0.0358
HIS					PT37	17.283	18.516	6.988	0.0409
Cγ	18.921	15.161	2.698		PT38	19.994	17.188	6.029	0.0023
Nδ1	18.437	15.752	1.543		**TYR**				
Cε1	17.116	15.812	1.626		Cς	16.164	15.242	2.786	
Nε2	16.744	15.267	2.820		OH	14.815	15.390	2.691	
Cδ2	17.833	14.874	3.469		HH	14.573	15.752	1.791	
PT33	18.742	15.741	1.243	-0.2648	PT33	17.620	14.986	3.187	-0.0118
PT34	15.772	15.205	3.031	0.2617	PT34	14.770	15.231	3.206	-0.1848
PHE					PT35	14.608	16.118	0.924	0.1550
Cγ	18.926	14.982	2.956		PT36	19.112	13.203	5.070	0.0374
Cδ1	18.399	15.648	1.894						
Cε1	16.993	15.816	1.788						
Cς	16.173	15.310	2.748						
Cε2	16.700	14.643	3.810						
Cδ2	18.106	14.476	3.916						
PT33	17.346	15.235	2.728	-0.1102					
PT34	15.969	16.862	0.293	0.0392[c]					
PT35	14.133	15.433	2.656	0.0167					
PT36	15.435	13.823	5.172	0.0392[c]					

[a]values of q17 and q18 depend on the AA type (Table V.II)
[b]mean value for point charges 33 and 34 (Table V.II)
[c]mean value for point charges 34 and 36 (Table V.II)
Additional remark: for Cys, the model with a CG located on Sγ was selected (Table V.II).

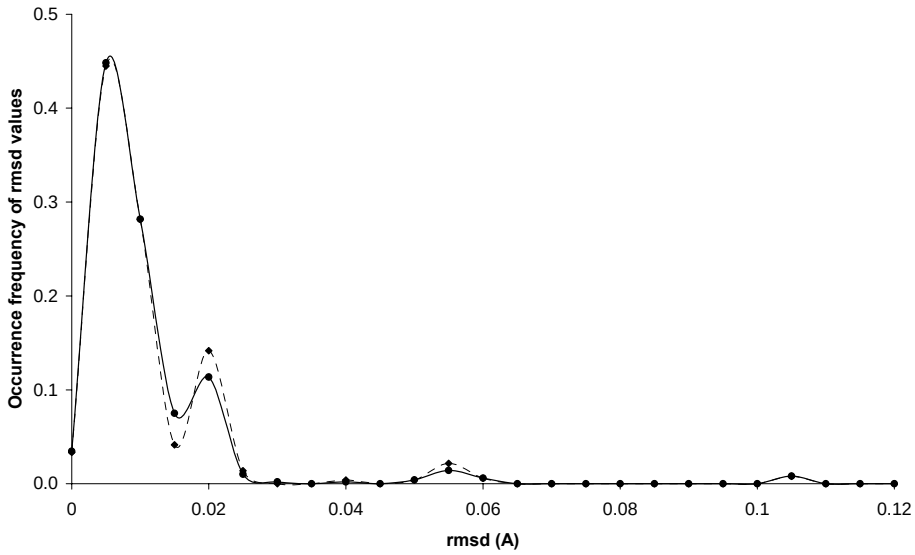


Figure VI.2. Occurrence frequency of the root mean square deviation (rmsd) values calculated between the atoms of the template motif and the atoms of the actual AA backbone or side chain residues, over all superimpositions achieved for the generation of the Amber-based (plain lines) and Gromos43A1-based (dashed lines) CGs of protein structure KcsA.

The largest *rmsd* values, *i.e.*, beyond 0.1 Å, correspond to a less efficient fit of the four end residues Gln119 required to design the Amber-based CG model, due to the terminal OXT atoms (Figure VI.3a). The lowest *rmsd* values, around 0.01 Å, characterize the superimpositions of the backbone templates, while all larger *rmsd* values, from 0.02 to 0.06 Å, characterize the superimpositions of the side chain templates. Particularly, *rmsd* values at about 0.05-0.06 Å originate from the superimpositions of the Tyr side chains.

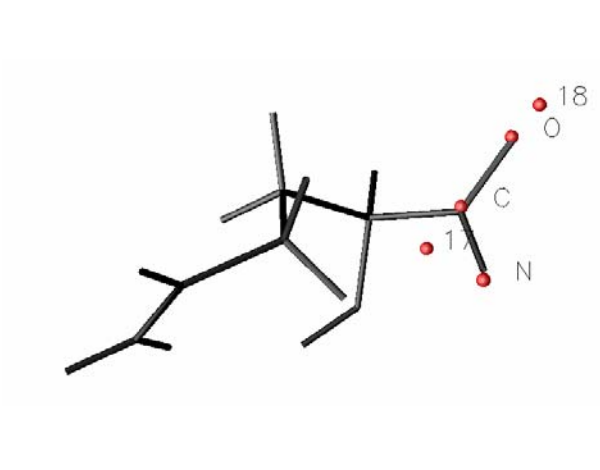


Figure VI.3a. Amber-based template motif of Gln backbone (red spheres) as superimposed on Gln119 of chain A of protein KcsA. The three atoms C, O, and N are used to generate the transformation matrix that is further applied to CGs numbered 17-18.

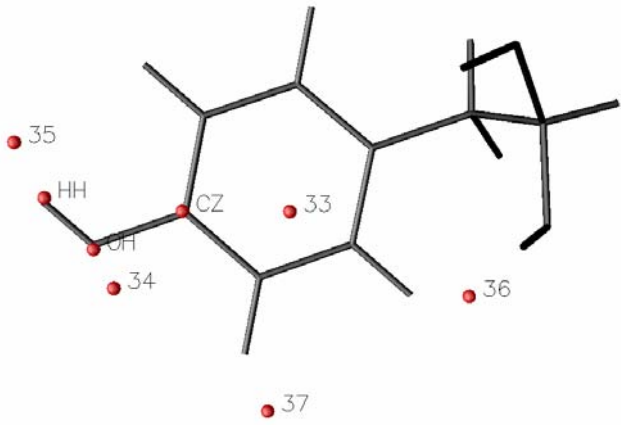


Figure VI.3b. Amber-based template motif of Tyr side chain (red spheres) as superimposed on Tyr82 of chain A of protein KcsA. The three atoms Cς, OH, and HH are used to generate the transformation matrix that is further applied to CGs numbered 33-37.

For example, Tyr82 of chain A that led to *rmsd* = 0.062 Å, is illustrated in Figure VI.3b where one can see that it nevertheless corresponds to a rather good superimposition of the three template atoms Cς, OH, and HH.

The resulting complete KcsA CG models are characterized by dipole moments and total charges that are reported in Table VI.IV, both for the Amber and Gromos43A1 FF. As shown, the total electric charges of the raw models, 2.9480 and 4.4332, are not strictly equal to the total charges of the all-atom representations, *i.e.*, 3.8240 and 4.0000, for Amber and Gromos43A1, respectively. Thus, as already applied for the hAr structure [leh07], the charge values of the raw Amber and Gromos43A1 models were corrected by adding a small charge amount to each of the CG ("Correction" in Table VI.IV) so as to reach a total molecular charge of 3.824 and 4.000 e^-, respectively. The differences between the raw CG models and the corresponding original electrostatic properties are given in Table VI.IV, with initial values of *rmsdV* = 11.63 and 6.50 kcal/mol, and *rmsdμ* = 310.31 and 180.14 D, respectively. We here recall that with the program QFIT [bor], *rmsdV* values were calculated by considering all grid points located at distances between 1.4 and 2.0 times the vdW radius of the atoms of the original protein structure.

Table VI.IV. Electrostatic properties of the original atom-based models and their CG versions.

	Raw CG model	Corrected CG model	ED-based model
Amber			
Total charge (all-atom)	3.8240		
μ all-atom (D)	(1237.70 ; 496.08 ; 141.52)[a]		
Number of CG points	1492	1492	494
Total charge (CG)	2.9480	3.8240	3.8240
Correction	-	0.0006	-
μ CG (D)	(951.33 ; 384.23 ; 99.39) [a]	(1255.17 ; 496.06 ; 183.58) [a]	(1241.50 ; 492.20 ; 185.91) [a]
rmsdV (kcal/mol)	11.63	7.58	17.72
rmsdμ (D)	310.31	45.53	44.72
Gromos43A1			
Total charge (all-atom)	4.0000		
μ all-atom (D)	(1293.20 ; 518.94 ; 148.06) [a]		
Number of CG points	1176	1176	494
Total charge (CG)	4.4332	4.0000	4.0000
Correction	-	-0.0004	-
μ CG (D)	(1452.86 ; 575.32 ; 209.55) [a]	(1302.46 ; 519.98 ; 167.65) [a]	(1312.24 ; 507.56 ; 203.04) [a]
rmsdV (kcal/mol)	6.50	3.57	18.38
rmsdμ (D)	180.14	21.69	59.29

[a]x, y, and z components of μ.

The charge correction drastically improves both the MEP and dipole moment values, with *rmsdV* = 7.58 and 3.57 kcal/mol, and *rmsdμ* = 45.53 and 21.69 D. As already concluded before, the ED-based models built from peaks observed in PASA ED distribution functions at t = 1.4 bohr^2 are less good approximations, especially for the Gromos43A1 model which consists of a limited set of unitary charges, located on Arg, Asp, Glu, Lys, and end residues.

Visualizations of 3D MEP iso-contours, generated from MEP maps built with a grid step of 0.5 Å (Figures VI.4 and VI.5), do not permit to clearly differentiate the all-atom and CG models (see for example Figures VI.4a and b, and Figures VI.5a and b), while ED-based models only reproduce global features of the all-atom unsmoothed MEP grids (Figures VI.4c and VI.5c). One will additionally notice that, for the ED-based models generated at t = 1.4 bohr^2, most of the charge values are close to +1, 0, or -1. These integer values are perfectly obtained for the Gromos43A1 CGs considering (i) the atom charge values reported in Appendix II and (ii) at t = 1.4 bohr^2, ED-based protein fragments are the result of a nice backbone/side chain decomposition procedure.

Finer and more quantitative comparisons were thus achieved. For that purpose, MEP profiles were also calculated using the original atom charges along the channel axis, defined by the Cartesian coordinates of ions K401 and K403 (Figures VI.6 and VI.7). As illustrated in Figures VI.6 and VI.7, the selective filter region is characterize by two MEP minima, followed by a large energy barrier which covers the hydrophobic cavity and narrow pore regions. The calculation of the corresponding MEP profiles using the Amber- and Gromos43A1-based CG models also generate similar behaviors, however displaced towards higher energy values. The introduction of a correction to the CG charges so as to preserve the initial total charges led to a slight increase or decrease of the MEP values depending on the correction sign, *i.e.*, positive or negative, respectively (Table VI.IV). Differences in electrostatic energy values are smaller for the Gromos43A1 FF, but for that last model (Figure VI.7), the barrier highest value now adopts a slightly positive value. The largest discrepancies are observed with the ED-based models for which the minima features are strongly modified. Particularly, the two-minima region has been leveled out and the energy barrier was drastically shifted toward positive energy values. Equivalent views are given as MEP iso-contours depicting the inner channel of KcsA (Figures VI.8 and VI.9), respectively for the Amber and Gromos43A1 FFs. The contours are displayed in a plane formed by K401, K403, and O_{Thr75A}. In these Figures, only chains A and C are displayed in order to clearly visualize the channel and cavity of the protein structure. The three potassium ions are also shown, but were not included in the calculations of the MEP grid values.

In a final attempt to assess the ED-based CG models, plots of axial MEP values established at various values of t with charges calculated using Equation 11 of Section III (Figures VI.10 and

VI.11) show that values of t = 0.9 and 0.8 bohr^2 determined previously for the Amber and Gromos43A1 FF, respectively, are not adequate to approximate the electrostatic properties of KcsA due to the lack of backbone dipoles. Rather, to reflect the two-minima region, a value lower than 0.5 bohr^2 should be needed. At such a value of t, the number of ED peaks is equal to 2495, a value that is too large to be very useful in the reduction of the calculation time of electrostatic properties. Let us note that the number of ED peaks is reduced to 2289 and 1350 when t = 0.7 and 0.75 bohr^2, respectively, but at such a smoothing degree, fine details of the MEP are lost, while they are preserved with the 1494 or the 1176 MEP CGs. It is nonetheless interesting to note that a value of t = 1.4 bohr^2 is not without any interest. Indeed, at smoothing degrees close to that value, *i.e.*, at t = 1.3 bohr^2, the CG-based dipole moment of the whole molecular structure of KcsA, while not being very good, is the best approximation of the all-atom dipole moment, both for the Amber and Gromos43A1 FFs (Table VI.V), with *rmsdμ* = 43.82 and 58.53 D, respectively.

Table VI.V. *rmsdV* (kcal/mol) and *rmsdμ* (D) values calculated between ED-based CG models obtained using a hierarchical merging/clustering algorithm, at various smoothing degrees t, and the corresponding all-atom MEP.

t	No. ED peaks	Amber		Gromos43A1	
		rmsdV	*rmsdμ*	*rmsdV*	*rmsdμ*
0.3	2900	8.06	41.80	6.94	41.87
0.8	1303	17.07	107.63	14.59	118.05
0.9	898	16.35	94.97	17.42	64.71
1.0	500	16.40	51.47	17.20	80.49
1.1	499	17.04	57.23	17.61	73.17
1.2	499	17.10	56.41	17.65	71.89
1.4	494	17.72	44.72	18.38	59.29
1.5	494	17.74	46.00	18.38	60.48
1.6	486	17.69	44.54	18.38	62.05
1.7	472	17.68	49.07	18.38	63.97
1.8	461	17.99	57.55	18.38	66.23
1.9	461	17.99	59.91	18.38	68.83
2.0	460	17.88	61.18	18.39	71.80

Finally, the automated procedure described in this Section was applied to the 12-residue β-hairpin peptide structure HP7, with the following results (Table VI.VI.). All structural details are not given as they are close to the CG models that were already presented in Sections IV and V.

Table VI.VI. Total molecular charge (e^-), *rmsdV* (kcal/mol) and *rmsdμ* (D) values calculated for MEP-based CG models obtained using the templates given in Tables VI.II and VI.III, respectively for the Amber and Gromos43A1 sets of charges. Corresponding values obtained from CG models built from a hierarchical merging/clustering algorithm applied to isolated AA are given in parentheses.

	Amber	Gromos43A1
Number of CG points	48	49
Total molecular charge (no correction)	0.9862	1.0227
rmsdV	4.81 (4.63)	2.90 (2.70)
rmsdμ	4.42 (5.51)	0.98 (0.26)

The results presented in Table VI.VI tend to show that a CG model built from a more time-consuming application of our hierarchical merging/clustering algorithm to each of the AA individually, is not a drastrically better model than the one that can be directly obtained from the use of a table of AA CG templates.

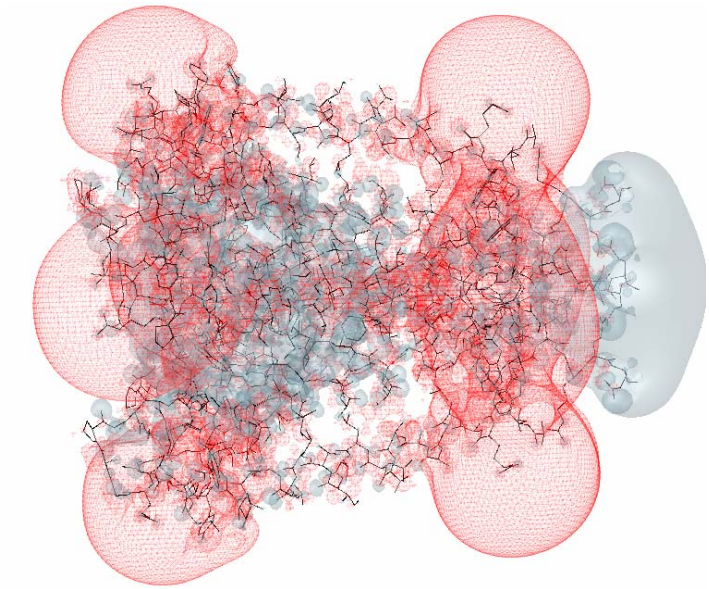


Figure VI.4a. All –atom Amber MEP iso-contours (blue: -0.1 e^-/bohr, red: 0.1 e^-/bohr) superimposed on the 3D structure of protein KcsA (sticks). Figure was generated using DataExplorer [odx].

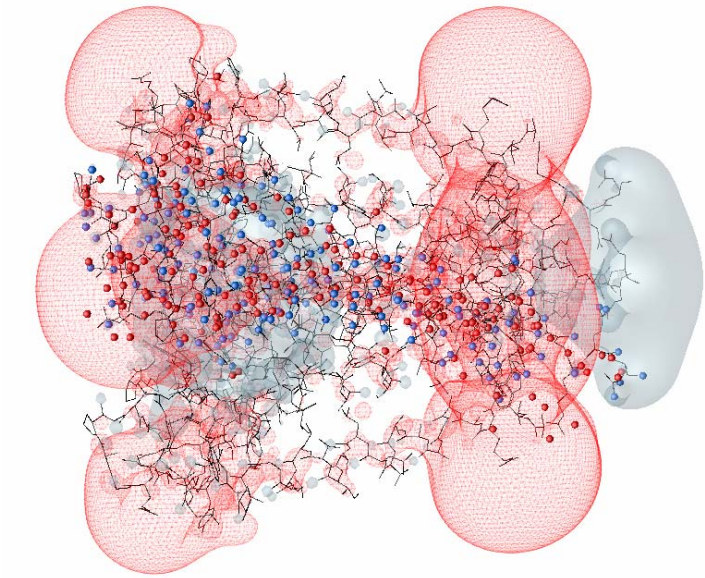


Figure VI.4b. Amber-based CG MEP iso-contours (blue: -0.1 e^-/bohr, red: 0.1 e^-/bohr) superimposed on the 3D structure of protein KcsA (sticks) and the Amber-based CG model built for one monomer of the structure (negative and positive CGs are displayed using blue and red spheres, respectively). Figure was generated using DataExplorer [odx].

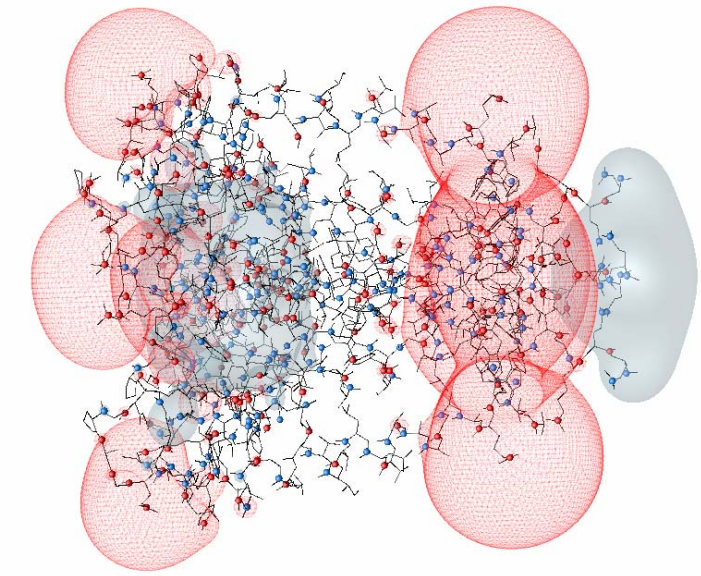


Figure VI.4c. ED-based MEP iso-contours (blue: -0.1 e^-/bohr, red: 0.1 e^-/bohr) superimposed on the 3D structure of protein KcsA (sticks) and the CG model built from the ED peaks obtained using a hierarchical merging/clustering algorithm, at t = 1.4 bohr^2. Negative and positive CGs are displayed using blue and red spheres, respectively. Charges were calculated using the Amber force field. Figure was generated using DataExplorer [odx].

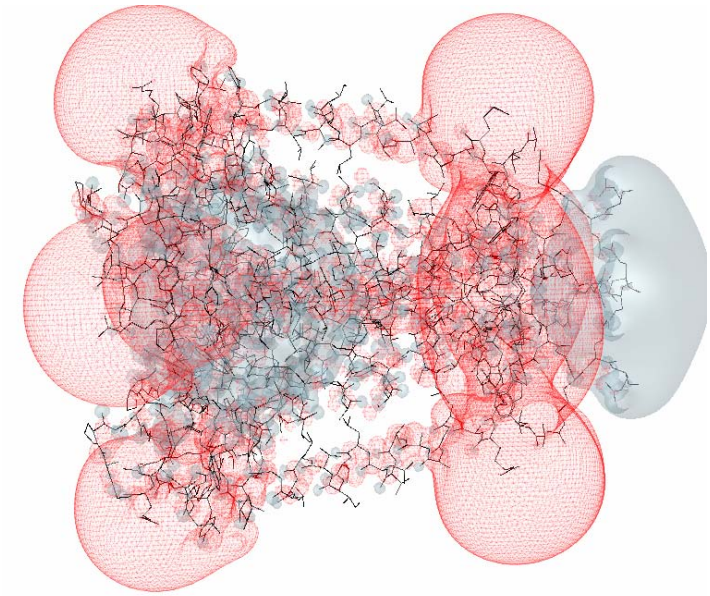


Figure VI.5a. Gromos43A1 MEP iso-contours (blue: -0.1 e^-/bohr, red: 0.1 e^-/bohr) superimposed on the 3D structure of protein KcsA (sticks). Figure was generated using DataExplorer [odx].

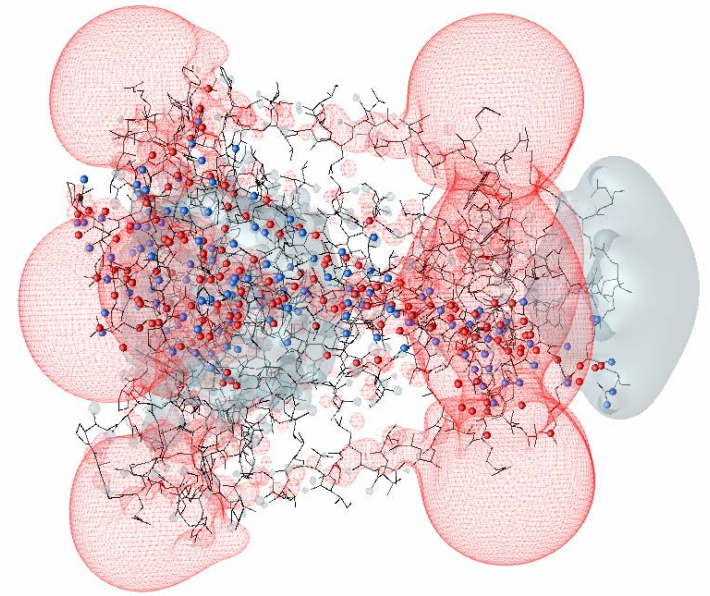


Figure VI.5b. Gromos43A1-based CG MEP iso-contours (blue: -0.1 e^-/bohr, red: 0.1 e^-/bohr) superimposed on the 3D structure of protein KcsA (sticks) and the Amber-based CG model built for one monomer of the structure (negative and positive CGs are displayed using blue and red spheres, respectively). Figure was generated using DataExplorer [odx].

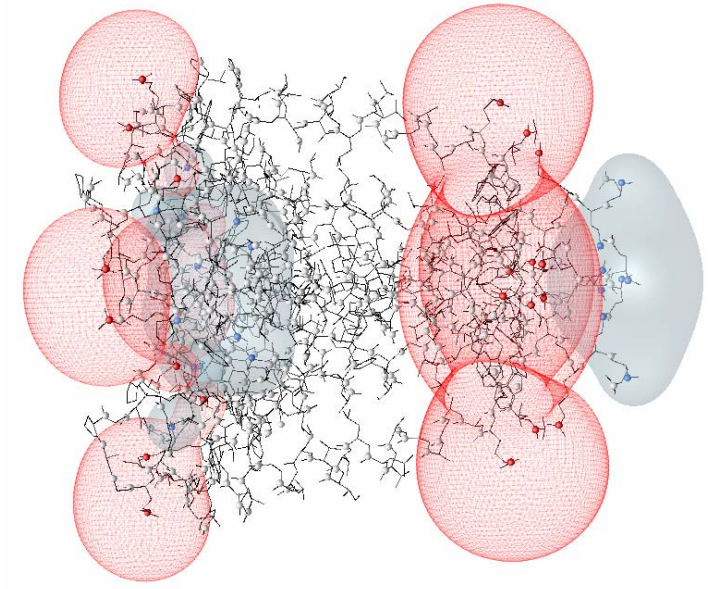


Figure VI.5c. ED-based MEP iso-contours (blue: -0.1 e^-/bohr, red: 0.1 e^-/bohr) superimposed on the 3D structure of protein KcsA (sticks) and the CG model built from the ED peaks obtained using a hierarchical merging/clustering algorithm, at t = 1.4 bohr^2. Negative, neutral, and positive CGs are displayed using blue, white, and red spheres, respectively. Charges were calculated using the Gromos43A1 force field. Figure was generated using DataExplorer [odx].

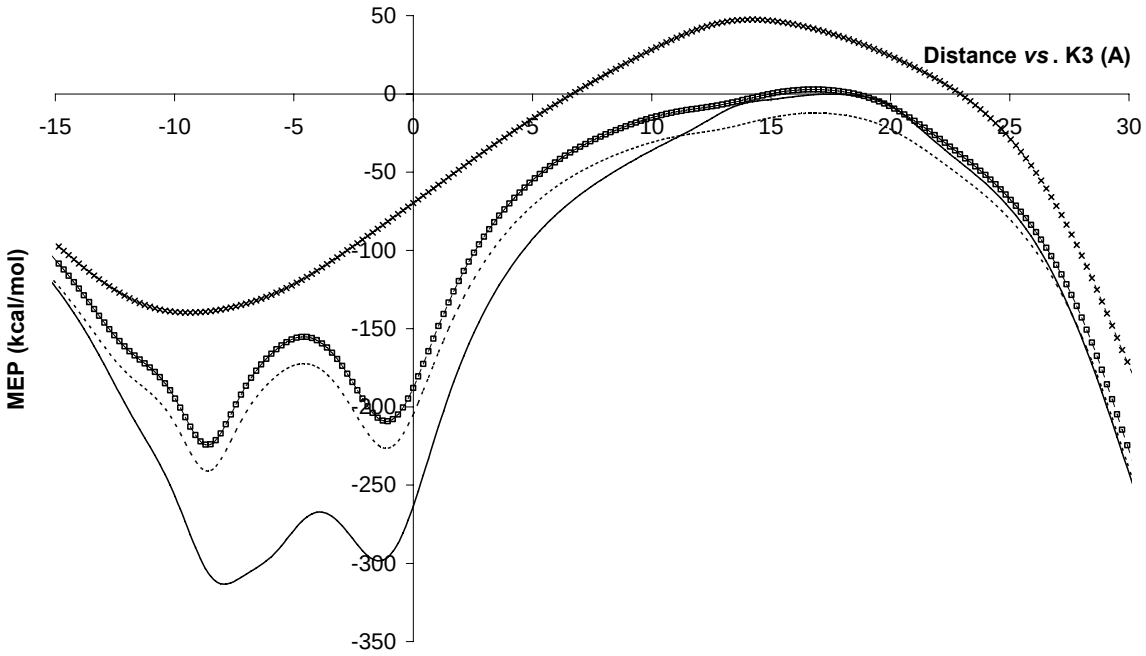


Figure VI.6. MEP along the central axis of the KcsA potassium channel calculated (a) using the all-atom Amber FF (plain line), (b) Amber-based CG model (dashed lines), (c) corrected Amber-based CG model (dashed lines & squares), and (d) the ED-based CG model with charges calculated at t = 1.4 bohr^2 using Amber (dashed lines & crosses).

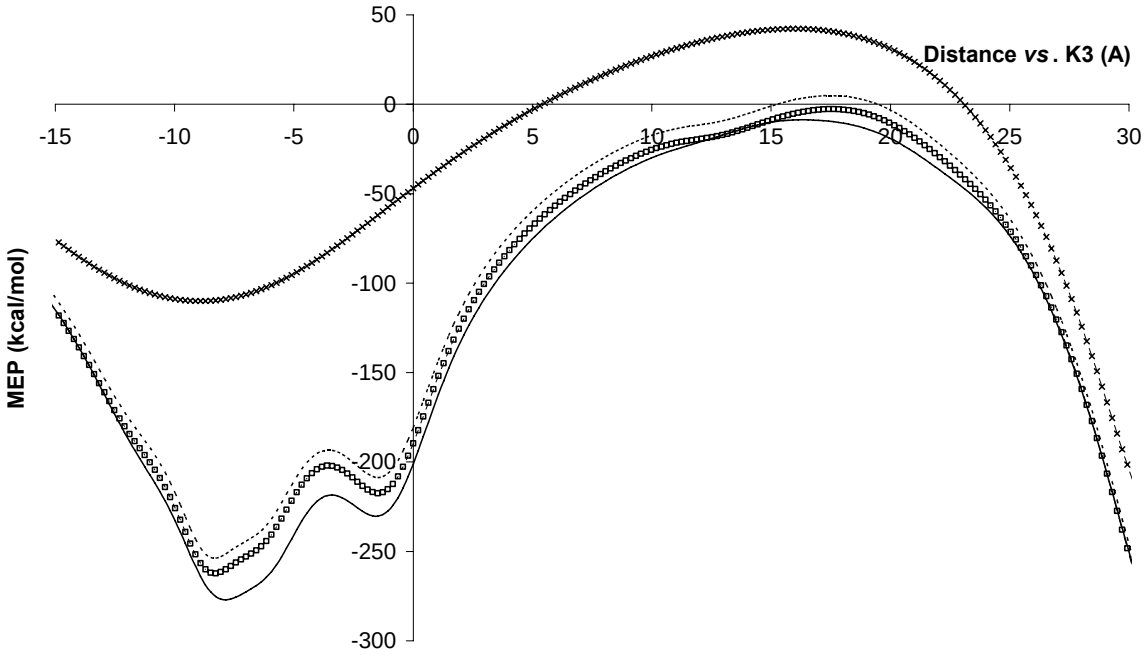


Figure VI.7. MEP along the central axis of the KcsA potassium channel calculated (a) using the Gromos43A1 FF, (b) Gromos43A1-based CG model (dashed lines), (c) corrected Gromos43A1-based CG model (dashed lines & squares), and (d) ED-based CG model with charges calculated at t = 1.4 bohr^2 using Gromos43A1 (dashed lines & crosses).

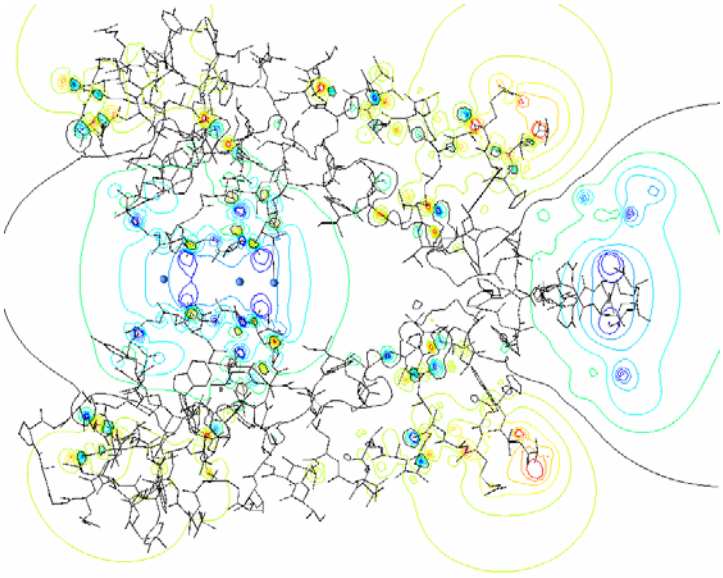
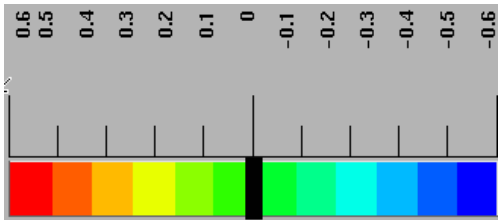


Figure VI.8a. All –atom Amber MEP iso-contours (-0.6 to 0.6 e^-/bohr) superimposed on the 3D structure of protein KcsA chains A and C (sticks) and the three K^+ ions (blue spheres). Ions K401 and K403 are separated by a distance of 10.62 Å. Figure was generated using DataExplorer [odx].



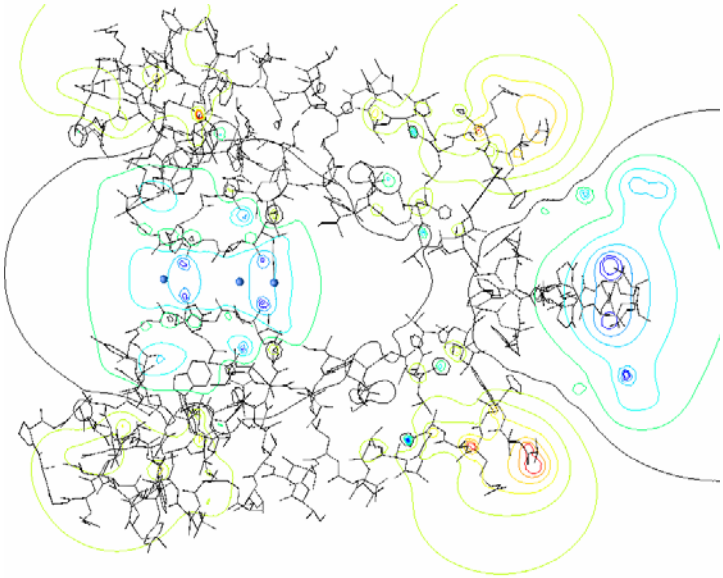


Figure VI.8b. Amber-based corrected CG MEP iso-contours (-0.6 e^-/bohr to 0.6 e^-/bohr) superimposed on the 3D structure of protein KcsA chains A and C (sticks) and the three K^+ ions (blue spheres). Ions K401 and K403 are separated by a distance of 10.62 Å. Figure was generated using DataExplorer [odx].

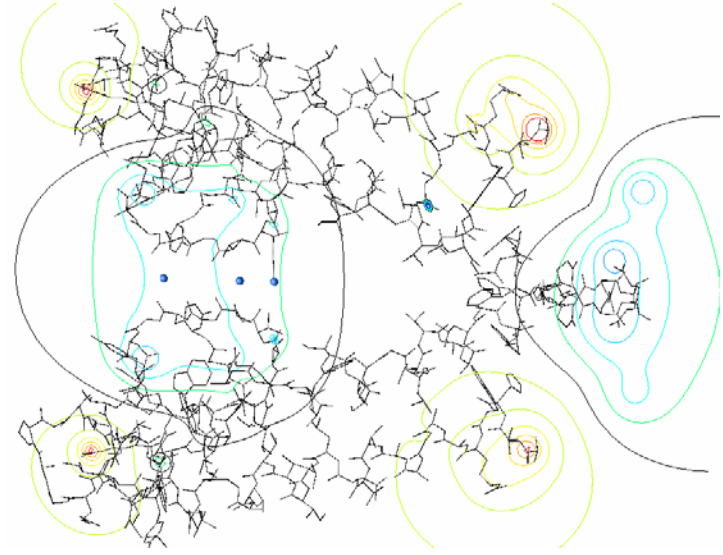


Figure VI.8c. ED-based MEP iso-contours (-0.6 e^-/bohr to 0.6 e^-/bohr) superimposed on the 3D structure of protein KcsA chains A and C (sticks) and the three K^+ ions (blue spheres). Peaks were obtained using a hierarchical merging/clustering algorithm applied to the PASA ED distribution function, at t = 1.4 bohr^2. Ions K401 and K403 are separated by a distance of 10.62 Å. Figure was generated using DataExplorer [odx].

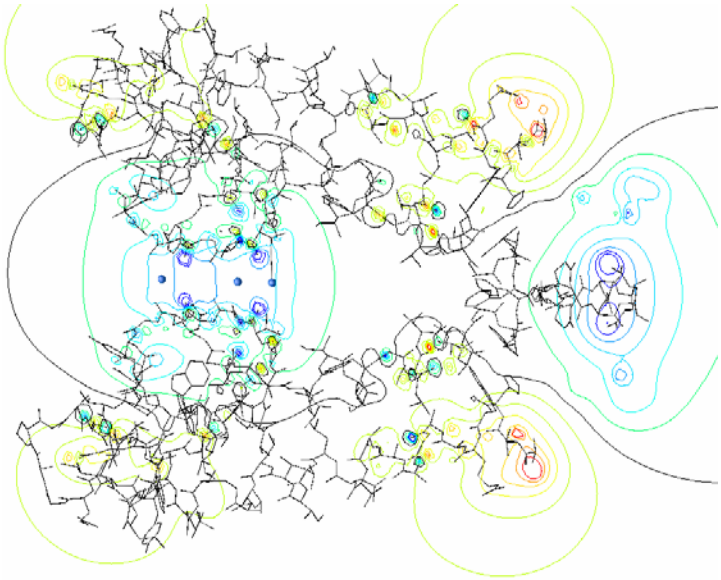
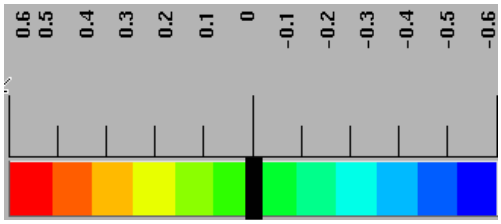


Figure VI.9a. Gromos43A1 MEP iso-contours (-0.6 e^-/bohr to 0.6 e^-/bohr) superimposed on the 3D structure of protein KcsA chains A and C (sticks) and the three K^+ ions (blue spheres). Ions K401 and K403 are separated by a distance of 10.62 Å. Figure was generated using DataExplorer [odx].



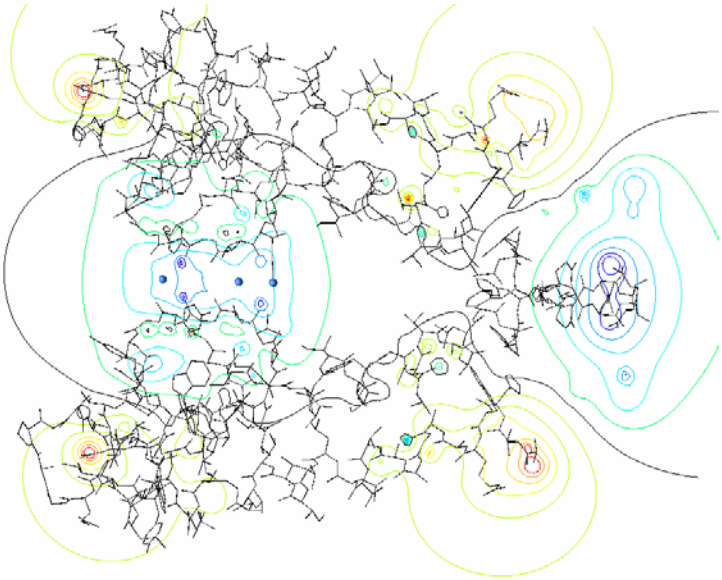


Figure VI.9b. Gromos43A1-based corrected CG MEP iso-contours (-0.6 e^-/bohr to 0.6 e^-/bohr) superimposed on the 3D structure of protein KcsA chains A and C (sticks) and the three K^+ ions (blue spheres). Ions K401 and K403 are separated by a distance of 10.62 Å. Figure was generated using DataExplorer [odx].

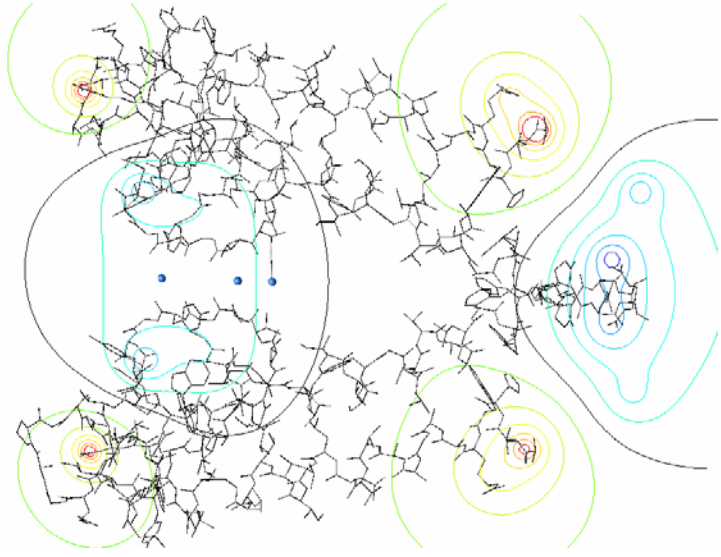


Figure VI.9c. ED-based MEP iso-contours (-0.5 e^-/bohr to 0.6 e^-/bohr) superimposed on the 3D structure of protein KcsA chains A and C (sticks) and the three K^+ ions (blue spheres). Peaks were obtained using a hierarchical merging/clustering algorithm applied to the PASA ED distribution function, at t = 1.4 bohr^2. Ions K401 and K403 are separated by a distance of 10.62 Å. Figure was generated using DataExplorer [odx].

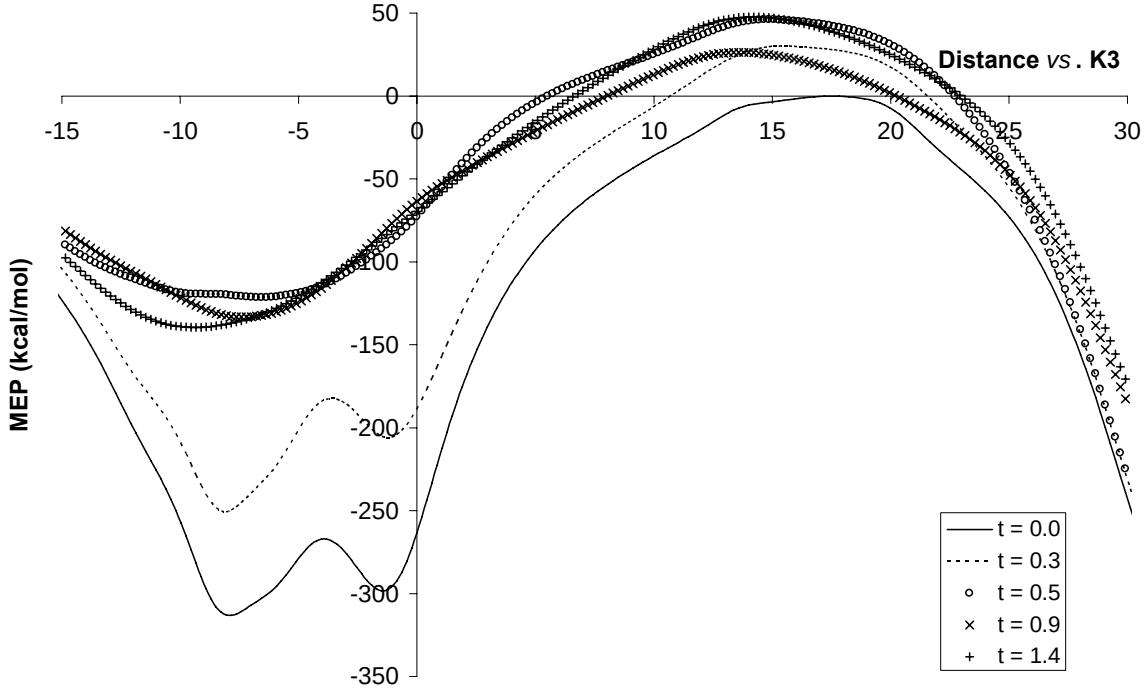


Figure VI.10. MEP along the central axis of the KcsA potassium channel calculated using the ED peak models obtained from a hierarchical merging/clustering algorithm, at various values of t. Charges were calculated using the Amber FF charges.

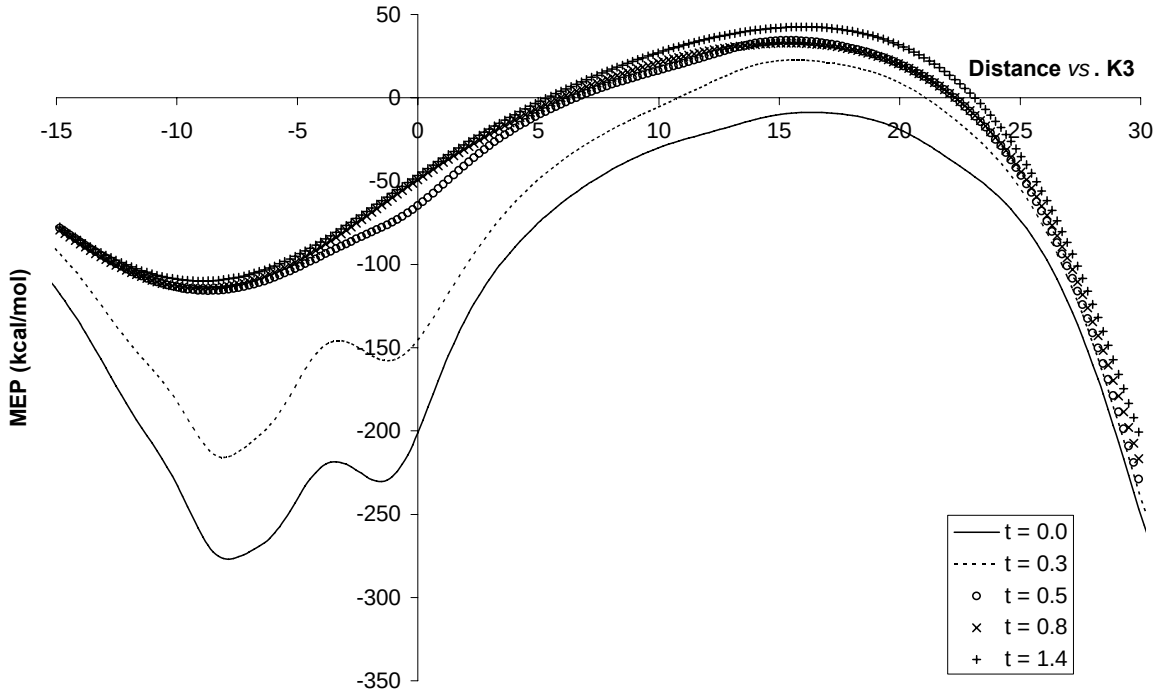


Figure VI.11. MEP along the central axis of the KcsA potassium channel calculated using the ED peak models obtained from a hierarchical merging/clustering algorithm, at various values of t. Charges were calculated using the Gromos43A1 FF charges.

VII. Conclusions and Perspectives

In this work, we applied a hierarchical merging/clustering algorithm to molecular scalar fields such as electron density (ED) distribution functions and molecular electrostatic potential (MEP) functions. Through the use of such an algorithm, the decomposition of a molecular structure, particularly a protein structure, was achieved by following the trajectories of its consitituting atoms in its progressively smoothed three-dimensional (3D) molecular field. A protein structure can thus be described by a limited set of representative points, which correspond to peaks (and pits for a MEP) of the considered 3D molecular property. The aim of such calculations consisted in the evaluation of electrostatic properties such as point charges and dipole moments of a protein using reduced, or so-called coarse grain (CG) descriptions.

Particularly, to model an ED distribution function, we selected the Promolecular Atomic Shell Approximation (PASA) description [ama97] which consists in the representation of the molecular ED as a summation over atom-centered Gaussian functions. In that framework, a protein is decomposed into fragments whose size gets progressively bigger with the smoothing degree. An interesting situation occurs at a smoothing degree t around 1.4 bohr^2, where the protein structure is clearly partitioned into backbone and side chain fragments. One observes one fragment for each residue backbone, mainly composed of –(C=O)-N-Cα or a derivative, and one fragment for each residue side chain, except for Gly, Ala, Ile, Pro, and Val (no fragment at all), and Tyr (two fragments). These observations are consistent with several descriptions already proposed in the literature, such as the globbic description levels of protein structures at a crystallographic resolution of about 3 Å [guo99] and the CG model proposed by Basdevant *et al*. [bas07]. Results showed to be independent on the AA residue conformation. Electric charges can be associated with each molecular fragment. They are located at the corresponding local ED maxima and are calculated as summations over the atomic charges involved in the fragments. A previous work about the electrostatic properties of the human Aldose reductase (hAr) protein [leh07] showed that such an ED-based reduced description of a protein, built on the ED peaks at t = 1.4 bohr^2, and its resulting electric charges, led to a better description of the MEP than coarser approximations such as the well-known CG description based on unitary charges placed on charged residues only. Some MEP features were also preserved with respect to the all-atom MEP.

The application of the hierarchical merging/clustering algorithm to 3D Coulomb MEP functions led to the location of peaks and pits. The calculation of charges associated with such topological features from corresponding molecular fragments was however shown to be physically irrelevant, and the calculation of point charges was achieved using a charge fitting algorithm *vs.* unsmoothed MEP functions. The present work was especially focused on the use of the all-atom Amber MEP function [dua03] and was extended to the Gromos43A1 set of charges, but is readily applicable to other charge sets that are available in the literature. As electric charges cannot be directly obtained from the atom content of the molecular fragments, it was necessary to design a method to easily assign a MEP-based CG description to a protein structure. The first stage was to define reduced descriptions to each of the 20 natural amino acid (AA) residues that were selected with the following specific protonation states: Lys(+1), Arg(+1), Hisε, Glu(-1), and Asp(-1). To generate CG models that avoid too many interaction effects, we selected, for all MEP-based calculations, extended pentadecapeptide Gly_7-AA-Gly_7 with various rotamers for each of the 20 AA (except Gly, Ala, Asp, and Pro). The second stage was to apply our merging/clustering algorithm to determine the CG locations of the central AA residue. Charges were then assigned to these AA CGs through a charge fitting algorithm, and were tabulated as reference values to be used for any CG model of a protein structure. Contrarily to ED fragments, MEP-based CG descriptions were shown to be sensitive to the molecular conformation. Detailed analyses were first carried out at the smoothing level of 1.4 $bohr^2$, like for the ED-based results, a value beyond which there were no more drastic changes in the merging/clustering decomposition results.

First applications were carried out for the particular case of a 12-residue peptide HP7 (PDB access code 2EVQ). This structure was selected as it is deeply detailed in the literature [bas07] and was thus an interesting reference case. Results showed that MEP-based CG descriptions led to electrostatic models of similar quality than the previously published CG model [bas07]; after a final optimization stage, the CG distribution was shown to even provide a better representation of the MEP and dipole moment of HP7.

Extended studies were achieved at various levels of smoothing, and showed that the optimal value of t is slightly dependent on the selected FF charges. For ED-based reduced descriptions, it appeared that the choice of t may also be dependent on the protein structure. Indeed, new applications were also achieved for the potassium ion channel KcsA, a tetrameric structure made of four 97-residue long monomers (PDB access code 1BL8). Such a system has gained a strong interest as an ion channel model in the scientific community since the resolution of its 3D crystallographic structure. In the case of ED-based CG descriptions, values of t = 0.9 and 0.8 $bohr^2$ for a small peptide such as HP7, or t = 1.3 $bohr^2$ for the larger KcsA system, were found. On the

whole, it seems that a good reduced representation of a residue backbone should consist of at least two CG points, which is not the case at t = 1.3 or 1.4 bohr^2. For MEP-based CG representations, the optimal values of t were actually equal to 1.35 and 1.3 bohr^2 for Amber and Gromos43A1, respectively. For HP7, resulting 48- and 49-point CG models were built for Amber and Gromos43A1, respectively. They both were evaluated in terms of their ability to reproduce all-atom MEP grid values and corresponding molecular dipole moments.

Let us finally mention that, from our calculations, it seems that the location of CG steric centers like those defined as ED peaks, differ from the location of CG electrostatic centers. This might be a point to consider in the further development of a CG FF.

The weakness of the procedure first applied to study the HP7 case resided in the necessity to locate the CGs, each residue at a time. An automated procedure was thus implemented, and tested on the selected larger scale system, KcsA. The generation of CGs for each residue was achieved through a superimposition algorithm of CG template motifs on 3D PDB structure, for each residue. The resulting CG descriptions, consisting of 1494 and 1176 CGs and their tabulated charges in the frameworks of Amber and Gromos43A1, respectively, allowed to reproduce MEP trends observed in the all-atom MEP functions. Such agreements were not observed for the ED-based CG descriptions. Though not sufficient to demonstrate the full transferability property of our models, the results are thus encouraging, and open an interesting extension to the present work, for example, by comparing MEP calculated using the Poisson-Boltzmann formalism, or by calculating pKa values [sch01]. One can also imagine two more direct ways for transferability testing of Coulomb potentials. The first one consists in applying our procedure to a larger set of protein structures. The other would ask for a detailed comparison between AA-AA MEP profiles calculated at the all-atom and CG levels, and this, for all possible AA-AA pair.

An extension to our work would thus reside in the evaluation of CG models made of smaller numbers of CGs. Additional applications of our merging/clustering algorithm, that are expected to be extremely long for structures like KcsA, would thus be required; it was however shown that for a small system like HP7, the number of CGs stayed rather constant above t = 1.4 bohr^2. The question is thus raised whether it is still possible to drastically reduce the number of CGs built from topological features of smoothed MEP functions.

Finally, to directly link MEP and experimental ED distribution functions, one can use databases of transferable multipolar ED parameters for calculating atom charges, as presented in [zar07], and then calculate the MEP.

VIII. References

[ama97] Amat, L., Carbó-Dorca, R. Quantum Similarity Measures under Atomic Shell Approximation: First Order Density Fitting Using Elementary Jacobi Rotations, J. Comput. Chem. 18 (1997) 2023-2039.

[bad95] Bader, R.W. Atoms in Molecules - A Quantum Theory. Clarendon Press: Oxford (1995).

[bas07] Basdevant, N., Ha-Duong, T., Borgis, D. A Coarse-Grained Protein-Protein Potential Derived from an All-Atom Force Field, J. Phys. Chem. B 111 (2007) 9390-9399.

[bec03] Becue, A., Meurice, N., Leherte, L., Vercauteren, D.P. Description of Protein-DNA Complexes in Terms of Electron Density Topological Features, Acta Crystallogr. D 59 (2003) 2150-2162.

[bec04a] Becue, A., Meurice, N., Leherte, L., Vercauteren, D.P. Evaluation of the Protein Solvent-Accessible Surface Using Reduced Representations in Terms of Critical Points of the Electron Density, J. Comput. Chem. 25 (2004) 1117-1126.

[bec04b] Becue, A. Development of an Original Genetic Algorithm Method Dedicated to Complementarity Studies Between Protein-Protein and Protein-Nucleic Acid Macromolecular Partners, PhD Thesis, FUNDP: Namur (2004).

[bec07] Becue, A., Meurice, N., Leherte, L., Vercauteren, D.P. Protein-Protein Docking Using Three-Dimensional Reduced Representations and Based on a Genetic Algorithm, In: Models, Mysteries, and Magic of Molecules, Eds. Boeyens, J.C.A, Ogilvie, J.F., Springer (2007).

[boi07] Boiteux, C., Kraszewski, S., Ramseyer, Ch., Girardet, Cl. Ion Conductance vs? Pore Gating and Selectivity in KcsA Channel: Modeling Achievements and Perspectives, J. Mol. Model. 13 (2007) 699-713.

[bon06] Bond, P.J., Sansom, M.S.P. Insertion and Assembly of membrane Proteins via Simulation, J. Amer. Chem. Soc. 128 (2006) 2697-2704.

[bon07] Bond, P.J., Holyoake, J., Ivetac, A., Khalid, S., Sansom, M.S.P. Coarse-Grained Molecular Dynamics Simulations of Membrane Proteins and Peptides, J. Struct. Biol. 157 (2007) 593-605.

[bor] Borodin, O., Smith, G.D. Force Field Fitting Toolkit, The University of Utah: Salt lake City, UT; http://www.eng.utah.edu/~gdsmith/fff.html.

[bot98] Botella, V., Pacios, L.F. Analytic Atomic Electron Densities in Molecular Self-Similarity Measures and Electrostatic Potentials, J. Mol. Struct. (Theochem) 426 (1998) 75-85.

[bul03] Bultinck, P., Carbó-Dorca, R., Van Alsenoy, Ch. Quality of Approximate Electron Densities and Internal Consistency of Molecular Alignment Algorithms in Molecular Quantum Similarity, J. Chem. Inf. Comp. Sci. 43 (2003) 1208-1217.

[car08] Carbone, P., Varnazeh, H.A.K., Chen, X., Müller-Plathe, F. Transferability of Coarse-Grained Force Fields: The Polymer Case, J. Chem. Phys. 128 (2008) 064904/1-064904/11.

[chn08] Chng, Ch.-P., Yang, L.-W. Coarse-Grained Models Reveal Functional Dynamics – II. Molecular Dynamics Simulation at the Coarse-Grained Level – Theories and Biological Applications, Bioinform. Biol. Insights 2 (2008) 171-185.

[cle08] Clementi, C. Coarse-Grained Models of Protein Folding: Toy Models or Predictive Tools ? Curr. Opin. Struct. Biol. 18 (2008) 10-15.

[dob08] Dobbins, S.E., Lesk, V.I., Sternberg, M.J.E. Insights into Protein Flexibility: The Relationship between Normal Modes and Conformational Change upon Protein-Protein Docking, Proc. Nat. Acad. Sci. USA 105 (2008) 10390-10395.

[dor02] Doruker, P., Jernigan, R.L., Bahar, I. Dynamics of Large Proteins through Hierarchical Levels of Coarse-Grained Structures, J. Comput. Chem. 23 (2002) 119-127.

[dow02] Downs, R.T., Gibbs, G.V., Boisen Jr., M.B., Rosso, K.M. A Comparison of Procrystal and ab initio Model Representations of the Electron-Density Distributions of Minerals, Phys. Chem. Miner. 29 (2002) 369-385.

[dua03] Duan, Y., Wu, C., Chowdhury, S., Lee, M.C., Xiong, G.M., Zhang, W., Yang, R., Cieplak, P., Luo, R., Lee, T., Caldwell, J., Wang, J.M., Kollman, P. A Point-Charge Force Field for Molecular Mechanics Simulations of Proteins Based on Condensed-Phase Quantum Mechanical Calculations, J. Comput. Chem. 24 (2003) 1999-2012.

[eya08] Eyal, E., Bahar, I. Toward a Molecular Understanding of the Anisotropic Response of Proteins to External Forces: Insights from Elastic Network Models, Biophys. J. 94 (2008) 3424-3435.

[emp08] Emperador, A., Carrillo, O., Rueda, M., Orozco, M. Exploring the Suitability of Coarse-Grained Techniques for the Representation of Protein Dynamics, Biophys. J. 95 (2008) 2127-2138.

[fug04] Fujitsuka, Y., Takada, S., Luthey-Schulten, Z., Wolynes, P.G. Optimizing Physical Energy Functions for Protein Folding, Proteins 54 (2004) 88-103.

[fuk01] Fukunaga, H., Aoyagi, T., Takimoto, J.-I., Doi, M. Derivation of Coarse-Grained Potential for Polyethylene, Comput. Phys. Comm. 142 (2001) 224-226.

[gad96] Gadre, S.R., Bhadane, P.K., Pundlik, S.S., Pingale, S.S. Molecular Recognition via Electrostatic Potential Topography, Theor. Comput. Chem. 3 (1996) 219-255.

[gas06] Gascon, J.A., Leung, S.S.F., Batista, E.R., Batista, V.S. A Self-Consistent Space-Domain Decomposition Method for QM/MM Computations of Protein Electrostatic Potentials, J. Chem. Theory Comput. 2 (2006) 175-186.

[gil96] Gilbert, D.G. Phylodendron, for Drawing Phylogenetic Trees, Version 0.8d; Indiana University, 1996. Software at http://iubio.bio.indiana.edu/soft/molbio/java/apps/trees/, Web form at http://iubio.bio.indiana.edu/treeapp/treeprint-form.html

[gir98] Gironés, X., Amat, L., Carbó-Dorca, R. A Comparative Study of Isodensity Surfaces Using ab initio and ASA Density Functions, J. Mol. Graph. Model. 16 (1998) 190-196.

[gir01] Gironés, X., Carbó-Dorca, R., Mezey, P.G. Application of Promolecular ASA Densities to Graphical Representation of Density Functions of Macromolecular Systems, J. Mol. Graph. Model. 19 (2001) 343-348.

[goh06] Gohlke, H., Thorpe, M.F. A Natural Coarse Graining for Simulating Large Biomolecular Motion, Biophys. J. 91 (2006) 2115-2120.

[goo95] Good, A.C., Ewing, T.J.A., Gschwend, D.A., Kuntz, I.D. New Molecular Shape Descriptors: Application in Database Screening, J. Comput. Aided Mol. Des. 9 (1995) 1-12.

[gra95] Grant, J.A., Pickup, B.T. A Gaussian Description of Molecular Shape, J. Phys. Chem. 99 (1995), 3503-3510.

[gro94] Grootenhuis, P.D.J., Roe, D.C., Kollman, P.A., Kuntz, I.D. Finding Potential DNA-Binding Compounds by Using Molecular Shape, J. Comput. Aided Mol. Des. 8 (1994) 731-750.

[gue97] Guex, N., Peitsch, M.C. SWISS-MODEL and the Swiss-PdbViewer. An Environment for Comparative Protein Modeling, Electrophoresis 18 (1997) 2714-2723.

[guo99] Guo, D.-Y., Blessing, R.H., Langs, D.A., Smith, G.D. On “Globbicity” of Low-Resolution Protein Structures, Acta Crystallogr. D 55 (1999) 230-237.

[hal90] Hall, S.R., Stewart, J.M. XTAL 3.0 User's Manual, Universities of Western Australia and Maryland (1990).

[hei90] Heisterberg, D.J., Technical report, Ohio Supercomputer Center. Translation from FORTRAN to C and Input/Output by Labanowski, J., Ohio Supercomputer Center, 1990; http://www.ccl.net/cca/software/SOURCES/quaternion-mol-fit/.

[hin08] Hinsen, K. Structural Flexibility in Proteins: Impact of the Crystal Environment, Bioinform. 24 (2008) 521-528.

[hor09] Hori, N., Chikenji, G., Berry, R.S., Takada, S. Folding Energy Landscape and Network Dynamics of Small Globular Proteins, Proc. Natl. Acad. Sci. USA 106 (2009) 73-78.

[izv05] Izvekov, S., Voth, G.A. A Multiscale Coarse-Graining Method for Biomolecular Systems, J. Phys. Chem. B 109 (2005) 2469-2473.

[jia02] Jiang, Y., Lee, A., Chen, J., Cadene, M., Chait, B.T., MacKinnon, R. The Open Pore Conformation of Potassium Channels, Nature 417 (2002) 523-526.

[kon06] Kondrashov, D.A., Cui, Q., Phillips Jr., G.N. Optimization and Evaluation of a Coarse-Grained Model of Protein Motion Using X-Ray Crystal Data, Biophys. J. 91 (2006) 2760-2767.

[kos91] Kostrowicki, J., Piela, L., Cherayil, B.J., Scheraga, H.A. Performance of the Diffusion Equation Method in Searches for Optimum Structures of Clusters of Lennard-Jones Atoms, J. Phys. Chem. 95 (1991) 4113-4119.

[leb99] Leboeuf, M., Köster, A.M., Jug, K., Salahub, D.R. Topological Analysis of the Molecular Electrostatic Potential, J. Chem. Phys. 111 (1999) 4893-4905.

[leh94] Leherte L., Allen, F.H. Shape Information from a Critical Point Analysis of Calculated Electron Density Maps: Application to DNA-Drug Systems, J. Comput. Aided Mol. Des. 8 (1994) 257-272.

[leh95] Leherte, L., Latour, Th., Vercauteren, D.P. Topological Analysis of Electron Density Maps of Chiral Cyclodextrin-Guest Complexes: A Steric Interaction Evaluation, Supr. Mol. Sci. 2 (1995) 209-217.

[leh97] Leherte, L., Vercauteren, D.P. Critical Point Analysis of Calculated Electron Density Maps at Medium Resolution: Application to Shape Analysis of Zeolite-Like Systems, J. Molec. Model. 3 (1997) 156-171.

[leh01] Leherte, L. Applications of Multiresolution Analyses to Electron Density Maps of Small Molecules: Critical Point Representations for Molecular Superposition, J. Math. Chem. 29 (2001) 47-83.

[leh04] Leherte, L. Hierarchical Analysis of Promolecular Full Electron-Density Distributions: Description of Protein Structure Fragments, Acta Crystallogr. D 60 (2004) 1254-1265.

[leh07] Leherte, L., Guillot, B., Vercauteren, D., Pichon-Pesme, V., Jelsch, Ch., Lagoutte, A., Lecomte, Cl. Topological Analysis of Proteins as Derived from Medium and High Resolution Electron Density: Applications to Electrostatic Properties, In Quantum

Theory of Atoms in Molecules - From Solid State to DNA and Drug Design, Eds. Matta, C.F., Boyd, R.J., Wiley-VCH: Weinheim (2007), pp. 285-316.

[leh08a] Leherte, L., Vercauteren, D.P. Collective Motions of Rigid Fragments in Protein Structures from Smoothed Electron Density Distributions, J. Comput. Chem. 29 (2008) 1472-1489.

[leh08b] Leherte, L., Vercauteren, D.P. Collective Motions in Protein Structures: Applications of Elastic Network Models Built from Electron Density Distributions, Comput. Phys. Comm. 179 (2008) 171-180.

[leu00] Leung, Y., Zhang, J.-S., Xu, Z.-B. Clustering by Scale-Space Filtering, IEEE T. Pattern Anal. 22 (2000) 1396-1410.

[lio93] Liotard, D., Rerat, M. Equivariant Morse Theory of the N-Body Problem: Application to Potential Surfaces in Chemistry, Theor. Chim. Acta 86 (1993) 297-313.

[liw09] Liwo, A., Czaplewski, C., Oldziej, S., Rojas, A.V., Kazmierkiewicz, R., Makowski, M., Murarka, R.K., Sheraga, H.A. Simulation of Protein Structure & Dynamics with the Coarse-Grained UNRES Force Field, in reference [voth09], pp. 107-122.

[liu07] Liu, P., Izvekov, S., Voth, G.A. Multiscale Coarse-Graining of Monosaccharides, J. Phys. Chem. B 111 (2007) 11566-11575.

[lym08] Lyman, E., Pfaendtner, J., Voth, G.A. Systematic Multiscale Parametrization of Heterogeneous Elastic network Models of Proteins, Biophys. J. 95 (2008) 4183-4192.

[mar07] Marrink, S.J., Risselada, H.J., Yefimov, S., Tieleman, D.P., de Vries, A.H. The MARTINI Forcefield: Coarse Grained Model for Biomolecular Simulations, J. Phys. Chem. B 111 (2007) 7812-7824.

[mat07] Mata, I., Molins, E., Espinosa, E. Zero-Flux Surfaces of the Electrostatic Potential: The Border of Influence Zones of Nucleophilic and Electrostatic Sites in Crystalline Environment, J. Phys. Chem. A 111 (2007) 9859-9870.

[mit00] Mitchell, A.S., Spackman, M.A. Molecular Surfaces from the Promolecule: A Comparison with Hartree-Fock ab initio Electron Density Surfaces, J. Comput. Chem. 21 (2000) 933-942.

[mon08] Monticelli, L., Kandasamy, S.K., Periole, X., Larson, R.G., Tieleman, D.P., Marrink, S.J. The MARTINI Coarse Grained Forcefield: Extension to Proteins, J. Chem. Theory Comput. 4 (2008) 819-834.

[mor08] Moritsugu, K., Smith, J.C. REACH Coarse-Grained Biomolecular Simulation: Transferability between Different Protein Structural Classes, Biophys. J. 95 (2008) 1639-1648.

[noi08a] Noid, W.G., Chu, J.-W., Ayton, G.S., Krishna, V., Izvekov, S., Voth, G.A., Das, A., Andersen, H.C. The Multiscale Coarse-Graining Method. I. A Rigorous Bridge between Atomistic and Coarse-Grained Models, J. Chem. Phys. 128 (2008) 244114/1-244114/11.

[noi08b] Noid, W.G., Liu, P., Wang, Y., Chu, J.-W., Ayton, G.S., Izvekov, S., Andersen, H.C., Voth, G.A. The Multiscale Coarse-Graining Method. II. Numerical Implementation for Coarse-Grained Molecular Models, J. Chem. Phys. 128 (2008) 244115/1-244115/20.

[nos07] Noskov, S.Y., Roux, B. Importance of Hydration and Dynamics on the Selectivity of the KcsA and NaK Channels, J. Gen. Physiol. 129 (2007) 135-143.

[odx] OpenDX, The Open Source Software Project Based on IBM's Visualization Data Explorer; Visualization and Imagery Solutions, Inc.; http://www.opendx.org/.

[pac92] Pacios, L.F. Simple Analytical Representation of Atomic Electron Charge Densities, Electrostatic Potentials, and Local Exchange Potentials, J. Phys. Chem. 96 (1992) 7294-7301.

[par05] Paramonov, L., Yaliraki, S.N. The Directional Contact Distance of Two Ellipsoids: Coarse-Grained Potentials for Anisotropic Interactions, J. Chem. Phys. 123 (2005) 194111/1-194111/11.

[pat90] Pathak, R.K., Gadre, S.R. Maximal and Minimal Characteristics of Molecular Electrostatic Potentials, J. Chem. Phys. 93 (1990) 1770-1773.

[ped04] Pedretti, A., Villa, L., Vistoli, G. VEGA - An Open Platform To Develop Chemo-Bio-Informatics Applications using Plug-in Architecture and Script Programming, J. Computer. Aided Mol. Des. 18 (2004) 167-173; http://www.ddl.unimi.it/.

[pic08] Piccinini, E., Ceccarelli, M., Affinito, F., Brunetti, R., Jacoboni, C. Biased Molecular Simulations for Free-Energy Mapping: A Comparison on the KcsA Channel as a Test Case, J. Chem. Theory Comput. 4 (2008) 173-183.

[pol82] Politzer, P., Landry, S.J., Wärnheim, T. Proposed Procedure for Using Electrostatic Potentials to Predict and Interpret Nucleophilic Processes, J. Phys. Chem. 86 (1982) 4767-4771.

[pop04] Popelier, P.L.A., Devereux, M., Rafat, M. The Quantum Topological Electrostatic Potential as a Probe for Functional Group Transferability, Acta Crystallogr. A 60 (2004) 427-433.

[roy08] Roy, D., Balanarayan, P., Gadre, S.R. An Appraisal of Poincaré-Hopf Relation and Application to Topography of Molecular Electrostatic Potentials, J. Chem. Phys. 129 (2008) 174103/1-174103/6.

[sch01] Schutz, Cl.N., Warshel, A. What Are the Dielectric "Constants" of Proteins and How To Validate Electrostatic Models? Proteins 44 (2001) 400-417.

[she08] Sherwood, P., Brooks, B.R., Sansom, M.S.P. Multiscale Methods for Macromolecular Simulations, Curr. Opin. Struct. Biol. 18 (2008) 630-640.

[sho93] Shoichet, B., Kuntz, I.D. Matching Chemistry and Shape in Molecular Docking, Prot. Eng. 6 (1993) 723-732.

[sim08] Simms, A.M., Toofanny, R.D., Kehl, C., Benson, N.C., Daggett, V. Dynameomics: Design of a Computational Lab Workflow and Scientific Data Repository for Protein Simulations, Prot. Eng. Des. Sel. 21 (2008) 369-377; http://www.dynameomics.org/.

[sjo90] Sjoberg, P., Politzer, P. Use of the Electrostatic Potential at the Molecular Surface to Interpret and Predict Nucleophilic Processes, J. Phys. Chem. 94 (1990) 3959-3961.

[tre08] Treptow, W., Marrink, S.-J., Tarek, M. Gating Motions in Voltage-Gated Potassium Channels Revealed by Coarse-Grained Molecular Dynamics Simulations, J. Phys. Chem. B 112 (2008) 3277-3282.

[tsi98a] Tsirelson, V.G., Avilov, A.S., Abramov, Y.A., Belokoneva, E.L., Kitaneh, R., Feil, D. X-Ray and Electron Diffraction Study of MgO, Acta Crystallogr. B 54 (1998) 8-17.

[tsi98b] Tsirelson, V.G., Abramov, Y., Zavodnik, V., Stash, A., Belokoneva, E., Stahn, J., Pietsch, U., Feil, D. Critical Points in a Crystal and Procrystal, Struct. Chem. 9 (1998) 249-254.

[vot09] Coarse-Graining of Condensed Phase and Biomolecular Systems, Voth, G.A. (Ed.), CRC Press, Boca Raton, FL, USA (2009).

[war07] Warshel, A., Kato, M., Pisliakov, A.V. Polarizable Force Fields: History, Test Cases, and Prospects, J. Chem. Theory Comput. 3 (2007) 2034-2045.

[yan06] Yang, L., Tan, C.-H., Hsieh, M.-J., Wang, J., Duan, Y., Cieplak, P., Caldwell, J., Kollman, P.A., Luo, R. New-Generation Amber United-Atom Force Field, J. Phys. Chem. B 110 (2006) 13166-13176.

[yan08] Yang, L.-W., Chng, Ch.-P. Coarse-Grained Models Reveal Functional Dynamics – I. Elastic Network Models – Theories, Comparisons and Perspectives, Bioinf. Biol. Insights 2 (2008) 25-45.

[zar07] Zarychta, B., Pichon-Pesme, V., Guillot, B., Lecomte Cl., Jelsch, Ch. On the Application of an Experimental Multipolar Pseudo-Atom Library for Accurate Refinement of Small-Molecule and Protein Crystal Structures, Acta Crystallogr. A 63 (2007) 108-125.

[zha08] Zhang, Z., Lu, L., Noid, W.G., Krishna, V., Pfaendtner, J., Voth, G.A. A Systematic Methodology for Defining Coarse-Grained Sites in Large Biomolecules, Biophys. J. 95 (2008) 5073-5083.

[zho08] Zhou, L., Siegelbaum, S.A. Effects of Surface Water on Protein Dynamics Studied by a Novel Coarse-Grained Normal Mode Approach, Biophys. J. 94 (2008) 3461-3474.

IX. Appendices

Appendix I. Atom charges as defined in the force field Amber [dua03].

Table 3. Atomic Partial Charges (in e.u.) of Standard Amino Acids.

	Gly	Ala	Ser	Cys	Val	Thr	Pro	Ile	Leu	Met	Asp	Asn	Glu	Gln	His	Lys	Arg	Trp	Phe	Tyr
N	−0.374	−0.405	−0.541	−0.396	−0.450	−0.245	−0.088	−0.451	−0.355	−0.395	−0.558	−0.430	−0.423	−0.387	−0.528	−0.436	−0.301	−0.428	−0.371	−0.488
H	0.254	0.294	0.345	0.295	0.440	0.255		0.329	0.262	0.281	0.320	0.255	0.307	0.301	0.282	0.251	0.234	0.242	0.234	0.264
C	0.581	0.570	0.483	0.643	0.447	0.560	0.334	0.569	0.573	0.600	0.443	0.617	0.470	0.419	0.662	0.725	0.730	0.584	0.548	0.622
O	−0.509	−0.555	−0.581	−0.585	−0.405	−0.552	−0.435	−0.620	−0.558	−0.566	−0.501	−0.524	−0.593	−0.565	−0.529	−0.563	−0.578	−0.495	−0.507	−0.527
C_α	−0.129	−0.028	0.118	−0.074	−0.052	−0.271	−0.035	−0.102	−0.101	−0.088	0.007	0.045	0.032	0.037	0.031	−0.039	−0.131	−0.020	−0.030	0.010
[a]H_α	0.089	0.121	0.142	0.141	−0.026	0.164	0.060	0.174	0.137	0.123	0.082	0.060	0.065	0.152	0.085	0.129	0.053	0.107	0.102	0.096
C_β		−0.230	0.147	−0.221	0.395	0.238	−0.003	0.062	−0.144	0.019	−0.048	−0.094	0.075	−0.032	−0.152	−0.108	0.037	−0.098	−0.099	−0.052
[b]H_β		0.077	0.040	0.147	−0.116	0.045	0.019	0.062	0.053	0.049	−0.015	0.043	−0.004	0.031	0.055	0.045	0.028	0.065	0.061	0.019
[c]C_γ, O_γ, S_γ			−0.640	−0.285	−0.090	−0.602	0.013	0.022	0.192	−0.208	0.745	0.584	−0.034	−0.020	0.278	0.033	0.012	−0.100	0.021	0.113
[d]H_γ			0.446	0.189	−0.009	0.405	0.020	0.012	0.001	0.124			−0.004	0.031		0.010	0.003			
$C_{\gamma 2}$						−0.176		−0.130												
$H_{\gamma 2(1,2,3)}$						0.060		0.030												
[e]C_δ, O_δ, N_δ							−0.012	−0.101	−0.123	−0.212	−0.730	−0.527	0.765	0.668	−0.423	−0.048	0.126	−0.174	−0.083	−0.183
[f]H_δ							0.044	0.024	0.022							0.071	0.068	0.171	0.098	0.133
[g]$C_{\delta 2}$, $N_{\delta 2}$												−0.782			−0.298			0.090		
[h]$H_{\delta 2}$												0.355			0.160					
[i]C_ε, O_ε, N_ε										−0.285			−0.824	−0.628	0.026	−0.070	0.465	−0.298	−0.157	−0.182
[j]H_ε										0.128					0.127	0.120	0.326	0.322	0.124	0.137
[k]$C_{\varepsilon 2}$, $N_{\varepsilon 2}$														−0.883	−0.098			0.142		
[l]$H_{\varepsilon 2}$														0.408	0.267					
$C_{\varepsilon 3}$																		−0.154		
$H_{\varepsilon 3}$																		0.123		
[m]C_ζ																−0.250	0.566	−0.211	−0.100	0.206
[n]H_ζ																0.295		0.126	0.115	
[o]$C_{\zeta 3}$, O_H, $N_{H(1,2)}$																	−0.686	−0.164		−0.421
[p]$H_{\zeta 3}$, H_H																	0.391	0.119		0.330
C_{H2}																		−0.133		
H_{H2}																		0.119		

[a]$H_{\alpha(2,3)}$ for Gly.
[b]$H_{\beta(1,2,3)}$ for Ala and H_β for Thr, Ile, and Val, $H_{\beta(2,3)}$ for all others.
[c]C_γ for Glu, Asp, Lys, Pro, Met, Asn, and Gln; $C_{\gamma(1,2)}$ for Val; O_γ for Ser; $O_{\gamma 1}$ for Thr; S_γ for Cys.
[d]$H_{\gamma 1}$ for Thr, $H_{\gamma(2,3)}$ for Gln, Arg, $H_{\gamma 1(2,3)}$ for Ile, $H_{\gamma(1,2)(1,2,3)}$ for Val.
[e]$C_{\delta 1}$ for Ile; Trp; $C_{\delta(1,2)}$ for Leu, Phe, Tyr; S_δ for Met; $O_{\delta 1}$ for Asn; $O_{\delta(1,2)}$ for Asp; C_δ for Pro, Glu, Gln, Lys, Arg; $N_{\delta 1}$ for His.
[f]$H_{\delta 1(1,2,3)}$ for Ile, $H_{\delta(2,3)}$ for Arg, Lys, Pro; $H_{\delta 1}$ for Trp; $H_{\delta(1,2)}$ for Phe, Tyr; $H_{\delta(1,2)(1,2,3)}$ for Leu.
[g]$C_{\delta 2}$ for His, Trp; $N_{\delta 2}$ for Asn.
[h]$H_{\delta 2(1,2)}$ for Asn.
[i]C_ε for Met, Lys; $C_{\varepsilon 1}$ for His; $C_{\varepsilon(1,2)}$ for Tyr, Phe; $O_{\varepsilon 1}$ for Gln; $O_{\varepsilon(1,2)}$ for Glu; N_ε for Arg; $N_{\varepsilon 1}$ for Trp.
[j]$H_{\varepsilon(2,3)}$ for Lys; $H_{\varepsilon(1,2,3)}$ for Met; H_ε for Arg; $H_{\varepsilon 1}$ for His, Trp; $H_{\varepsilon(1,2)}$ for Phe, Tyr.
[k]$C_{\varepsilon 2}$ for Trp, Phe, Tyr; $N_{\varepsilon 2}$ for Gln, His.
[l]$H_{\varepsilon 2(1,2)}$ for Gln.
[m]$C_{\zeta 2}$ for Trp; N_ζ for Lys.
[n]$H_{\zeta 2}$ for Trp; $H_{\zeta(1,2,3)}$ for Lys.
[o]$C_{\zeta 3}$ for Trp; O_H for Tyr; $N_{H(1,2)}$ for Arg.
[p]$H_{\zeta 3}$ for Trp; H_H for Tyr; $H_{H(1,2)(1,2)}$ for Arg.

AMBER Force Field

2005

Appendix II. Atom charges as defined in the force field Gromos43A1 and implemented in the program SwissPdbViewer [gue97].

	ALA	ARG	ASN	ASP	CYSH	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
N	-0.28	-0.28	-0.28	-0.28	-0.28	-0.28	-0.28	-0.28	-0.28	-0.28	-0.28	-0.28	-0.28	-0.28	0	-0.28	-0.28	-0.28	-0.28	-0.28
HN	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0	0.28	0.28	0.28	0.28	0.28
C	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38
O	-0.38	-0.38	-0.38	-0.38	-0.38	-0.38	-0.38	-0.38	-0.38	-0.38	-0.38	-0.38	-0.38	-0.38	-0.38	-0.38	-0.38	-0.38	-0.38	-0.38
CB	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.15	0.15	0	0	0
CG	0	0	0.38	0.27	0	0	0	0	0.13	0	0	0	0	0	0	0	0	-0.14	0	0
OG	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-0.548	-0.548	0	0	0
SG	0	0	0	0	-0.064	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HG	0	0	0	0	0.064	0	0	0	0	0	0	0	0	0	0	0.398	0.398	0	0	0
CD	0	0.09	0	0	0	0.38	0.27	0	0	0	0	0	0	-0.1	0	0	0	-0.1	-0.1	0
OD1	0	0	-0.38	-0.635	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
OD2	0	0	0	-0.635	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ND	0	0	0	0	0	0	0	0	-0.58	0	0	0	0	0	0	0	0	0	0	0
HD	0	0	0	0	0	0	0	0	0	0	0	0	0	0.1	0	0	0	0.1	0.1	0
CD2	0	0	0	0	0	0	0	0	0	0	0	0	0	-0.1	0	0	0	0	-0.1	0
ND2	0	0	-0.83	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HD2	0	0	0.415	0	0	0	0	0	0	0	0	0	0	0.1	0	0	0	0	0.1	0
CE	0	0	0	0	0	0	0	0	0.26	0	0	0.127	0	-0.1	0	0	0	0	-0.1	0
OE1	0	0	0	0	0	-0.38	-0.635	0	0	0	0	0	0	0	0	0	0	0	0	0
OE2	0	0	0	0	0	0	-0.635	0	0	0	0	0	0	0	0	0	0	0	0	0
NE	0	-0.11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-0.05	0	0
HE	0	0.24	0	0	0	0	0	0	0	0	0	0	0	0.1	0	0	0	0.19	0.1	0
CE2	0	0	0	0	0	0	0	0	0	0	0	0	0	-0.1	0	0	0	0	-0.1	0
NE2	0	0	0	0	0	-0.83	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HE2	0	0	0	0	0	0.415	0	0	0.19	0	0	0	0	0.1	0	0	0	0	0.1	0
CE3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-0.1	0	0
HE3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.1	0	0
CZ	0	0.34	0	0	0	0	0	0	0	0	0	0	0	-0.1	0	0	0	0	0.15	0
CZ2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-0.1	0	0
NZ	0	0	0	0	0	0	0	0	0	0	0	0.129	0	0	0	0	0	0	0	0
HZ	0	0	0	0	0	0	0	0	0	0	0	0.248	0	0.1	0	0	0	0	0	0
HZ2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.1	0	0
CZ3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-0.1	0	0
OH	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-0.548	0
NH1	0	-0.26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NH2	0	-0.26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HZ3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.1	0	0
HH	0	0.24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.398	0
CH2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-0.1	0	0
HH2	0	0.24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.1	0	0