Theor Chem Acc (2000) 105:101–109
DOI 10.1007/s002140000183

Theoretical
Chemistry Accounts

Regular article

A generalized Langevin dynamics approach to model solvent dynamics effects on proteins via a solvent-accessible surface. The carboxypeptidase A inhibitor protein as a model

Baldomero Oliva[1], **Xavier Daura**[2], **Enrique Querol**[1], **Francesc X. Avilés**[1], **O. Tapia**[3]

[1] Institut de Biologia Fonamental and Departament de Bioquímica. Universitat Autònoma de Barcelona, 08193 Bellaterra (Barcelona), Spain
[2] Laboratorium für Physikalische Chemie, ETH Zentrum, 8072 Zurich, Switzerland
[3] Department of Physical Chemistry, Uppsala University, Box 532, 75121 Uppsala, Sweden

Received: 2 February 2000 / Revised: 12 March 2000 / Accepted: 26 May 2000 / Published online: 2 November 2000

Abstract. A generalized Langevin dynamics (GLD) scheme is derived for (bio)macromolecules having internal structure, arbitrary shapes and a size larger than solvent molecules (i.e. proteins). The concept of solvent-accessible surface area (SASA) is used to incorporate solvent effects via external forces thereby avoiding its explicit molecular representation. A simulation algorithm is implemented in the GROMOS molecular dynamics (MD) program including random forces and memory effects, while solvation effects enter via derivatives of the surface area. The potato carboxypeptidase inhibitor (PCI), a small protein, is used to numerically test the approach. This molecule has N- and C-terminal tails whose structure and fluctuations are solvent dependent. A 1-ns MD trajectory was analyzed in depth. X-ray and NMR structures are used in conjunction with MD simulations with and without explicit solvent to gauge the quality of the results. All the analyses showed that the GLD simulation approached the results obtained for the MD simulation with explicit simple-point-charge-model water molecules. The SASAs of the polar atoms show a natural exposure towards the solvent direction. A FLS solvent simulation was completed in order to sense memory effects. The approach and results presented here could be of great value for developing alternatives to the use of explicit solvent molecules in the MD simulation of proteins, expanding its use and the time-scale explored.

Key words: Computer simulation of proteins – Generalized Langevin dynamics – Solvation effects – Potato carboxypeptidase inhibitor

Correspondence to: F. X. Avilés
e-mail: fx.aviles@blues.uab.es

1 Introduction

The formalisms of stochastic dynamics and Langevin equations (LE) are well known [1, 2, 3, 4, 5, 6, 7]. For instance, they have been used to derive computing schemes to calculate molecular dynamics (MD) trajectories of biomolecules by numerically simulating thermal baths [8, 9], thereby implicitly including solvent effects [10], and Brownian dynamics simulations of protein folding in torsional angle space [11, 12]. The natural extension of the LE to describe the motion of a generalized Brownian particle, which is not necessarily heavier than the particles of the solvent or surrounding medium, was proposed by Mori [1] and Kubo [2]. Such formalisms, when numerically implemented and applied to molecules having internal atomic structure and showing any molecular shape, may provide useful models for biomolecules immersed in various surrounding media and may generate information of interest to biochemists, molecular biologists and biophysicists.

Realistic simulations of hydration effects require a large a number of solvent molecules. For this reason, the simulations of proteins were carried out with a special parameter set where water was not explicitly included. We expect that such a procedure can be improved if solvent forces are explicitly included, while the solvent is still explicitly absent during the computation of the trajectory. In this article, we start from a generalized LE that can be derived from standard theory [4], where a new explicit solvation force term is obtained which reflects the arbitrary shape of the subsystem of interest that is usually concealed in the standard approach. The memory kernel, $\mathbf{K}(t - t')$, is evaluated using an algorithm introduced by Rey et al. [13]. The autocorrelation matrix of the stochastic forces is related to this matrix kernel; thus, dynamical solvent properties are introduced via the random force, $R(t)$, which is then constructed by introducing atom-type as well as solvent-accessible-surface-area (SASA) dependence for the atomic friction constant. The solvation forces [14, 15] representing hydrophobicity effects are

represented in a manner analogous to the approach used by Cramer and Truhlar [16]. This type of term was previously used in a different context by Wesen and Eisenberg [15] and by Eisenberg and McLachlan [17]. The algorithms are implemented in GROMOS [18] and tested with a MD calculation of the carboxypeptidase A (CPA) protein inhibitor from potatoes (PCI). This protein has structural features that are strongly solvent dependent. Furthermore, PCI has been studied with experimental [19, 20] and computer-assisted simulation techniques [21, 22, 23, 24], so the model results can be carefully gauged.

In GROMOS, the surrounding polarization effects are implicitly incorporated in a noninertial solvent framework [25]. This is a collisionless model for the solvent and is referred to as an in vacuo model in what follows. We expect the generalized Langevin dynamics (GLD) approach including solvophobic effects may contribute to improve the quality of solvation models in the protein's MD by including forces opposing implosion in vacuo. In the present case, a thoroughly studied model system will serve to gauge the quality of the atomic fluctuation patterns by MD simulations carried out with and without explicit water representation. To what extent the inclusion of memory effects (frequency-dependent-friction coefficients) together with the inclusion of solvophobic-like forces are sufficient to yield MD trajectories that would allow the essential features obtained with full solvent representation to be recovered is one basic point to be investigated here. The results reported for the protein used as a model, PCI, are encouraging.

2 Methods

2.1 Theoretical outline

One of the important limitations of the theory of Brownian motion, noticed early by Einstein [26], is the divergence of the mean velocity for indefinitely small values of the time t. In 1930, Uhlenbeck and Ornstein [27] proposed a model to solve this problem. The key issue was to realize that equations such as Langevin's for a particle or for rotational Brownian motion are stochastic differential equations. Doob, Wiener and others developed the theoretical basis leading to fundamentally correct solutions (see Ref. [28], where a number of important articles are reprinted, including Chandrasekhar's [29] 1943 seminal article). However, it was not until 1965 that Mori [1] and Kubo [2] gave a general result with the generalized LE.

The atomic/molecular description can conveniently be incorporated in the formalism of statistical mechanics of irreversible processes [3]. One has to remember that the solute, from a thermodynamic viewpoint, is an open system: it is exchanging energy and momentum with the surrounding media. For practical simulations, Eq. (1) is the starting point. The approximate time evolution equations for the dynamical variables of the system of interest that are used in molecular dynamics simulations have the form

$$\mathbf{P}_m = \partial V(\mathbf{r}_m)/\partial \mathbf{r}_m - \partial \langle V_{ms}\rangle_s/\partial \mathbf{r}_m + \mathbf{R}_m(t) + \int_d t'\mathbf{K}(t - -t')\cdot \mathbf{P}_m(t') \tag{1}$$

P being the momentum, **R** the stochastic force, V the potential energy, **r** the atom position and **K** the kernel memory, whilst subscripts "m" and "s" are for solute and solvent atoms, respectively. The first term describes the intramolecular force field of the given biomacromolecule and the second derives from the solvation free energy. The integral of the last term goes from $t' = 0$ to $t' = t$. This latter term should contain solvophobic forces via the modifications suffered by the solvent probability distribution function, which is assumed to include solute perturbations. The stochastic forces and memory terms are the subject of particular modeling described later.

2.2 Algorithms and definitions implemented

For proteins of arbitrary shape, the solvent-averaged interaction term can be used to introduce particular solvation forces. This solvation force is taken here from the derivative of the accessible surface for the k-th atom of the m system [15, 30]:

$$-\partial\langle V_{ks}\rangle_s/\partial \mathbf{r}_k = \mathbf{F}_k \mathrm{solv} = -\alpha_k \partial S/\partial \mathbf{r}_k \tag{2}$$

where the atomic solvation parameters used are those of Cramer and Truhlar [16] and the derivatives of the accessible surface are extracted analytically by using the Gauss–Bonet theorem [31]. From now on, the subscript m is no longer used as all the variables are referred to this subsystem, so V_{ks} stands for the interaction potential between the k-th atom in the solute interacting with the solvent and is averaged over the solvent degrees of freedom. A word of caution is in place here. In the original work, the parameters account for cavitation and dispersion effects. In our case, they are used to differentiate, at the level of forces, the action of hydrophobic and hydrophilic interactions. No reference to solvation energies is made at this stage. We expect to test the structural behavior of a test system that presents severe defects when simulated in a vacuum.

The friction effects in the atoms' motions have been analyzed by Canales and Padró [32] from simulations of soft spheres using molecular and Langevin dynamics. The study of electrolyte solutions through the Langevin dynamics [33] or GLD [13] has already been shown to be an important tool to reduce the computation time, obtaining reliable results, while for proteins, the stochastic dynamics [34, 35], not so commonly used, also represented an important improvement.

The approximation used to simplify Eq. (1) is to take the **K** matrix as a diagonal one: $\mathbf{K}_{ij} = \delta_{ij}\,\mathbf{K}_i(t')$, with δ_{ij} being the Kronecker delta. Following Rey et al. [13], the memory term, in the leap-frog algorithm can hence be written as

$$\int_{t_o}^{\infty} \mathbf{K}_i(t - t')\mathbf{p}_i(t')dt' = (1/2)\Delta t\mathbf{K}_i(0)\mathbf{p}_i(t) + \Delta t\left\{\sum\left\{\mathbf{K}_i((1/2 + j)\Delta t)\right\}\mathbf{p}_i(t - (1/2 + j)\Delta t)\right\} \tag{3}$$

where the summation is until **K** has vanished and $\Sigma\Delta t$ is the effective memory time [13]. In order to define the random force the following postulates are made:

1. Only one random force can be applied for each step on one atom if this is accessible to the solvent.

2. The strength of the random force is modeled with a Gaussian distribution that fulfills the second theorem of fluctuation–dissipation, the mean and the standard deviation for an atom "i" on the x-axis being defined as

$$\langle \mathbf{R}_{xi}\rangle = 0 \tag{4}$$

This property follows from the neglect of correlations between both subsystems. The kinetic energy of the N-atom system is given by $(1/2)\mathbf{P}^T(t)\cdot\mathbf{M}^{-1}\cdot\mathbf{P}(t)$, where **M** is now the diagonal matrix containing the masses of the atoms in block form, so the three atom degrees of freedom have the same mass value. On average, the kinetic energy $(1/2)\mathbf{P}^T(0)\cdot\mathbf{M}^{-1}\cdot\mathbf{P}(0) = 3N\,k_B T/2$ corresponds to thermal equilibrium. For the i-th atom

$$\langle \mathbf{R}_{xi}(0)\mathbf{R}_{xi}(0)\rangle = m_i\,\gamma_{xi}\,k_B\,T_o \tag{5}$$

T_o is the temperature at which the system should be kept. The friction constant is taken to be anisotropic (γ_{xi}) and it has been provided to fulfill the following conditions (these conditions have been arbitrarily defined in order to use the random force to keep the temperature of the system by small perturbations):

a. Dependence on the accessible surface area of the Datom (the largest area implies the largest friction).

b. Dependence on the atom type and its interaction with the solvent.
c. Dependence on the difference of temperature between the system and the bath.

3. The direction of the random force on an atom "i" depends upon its velocity and the temperature of the system:
When $T - T_o < 0$

if $m_i^{-1}|p_i|^2 - 3kT_o < 0$, then $\mathbf{R}_i$ increases the momentum,
if $m_i^{-1}|p_i|^2 - 3kT_o > 0$, then $\mathbf{R}_i = 0$.

When the bath is colder than the molecule, $T - T_o > 0$

if $m_i^{-1}|p_i|^2 - 3kT_o < 0$, then $\mathbf{R}_i$ decreases the momentum,
if $m_i^{-1}|p_i|^2 - 3kT_o > 0$, then $\mathbf{R}_i = 0$.

4. To complete the definition of the random force we define γ in agreement with the second and third postulates. Given a F_{max} to be the maximum strength at atom "i" produced by the electrostatic force of a ghost water-molecule-dipole located at the minimum of the attractive van der Waals distance between the center of atom "i" and the water oxygen, γ_x is defined as

$$\gamma_{xi} = (1/\tau)(|T - T_o|)^{1/2}(S_{xi}F_{max}/\pi\rho_i) \quad (6)$$

S being the accessible surface and S_x the projection on the x-axis, and τ is defined as the relaxation time [8], which is taken here to be 0.1 ps. ρ_i is the van der Waals radius of the i-th atom increased by the radius of a water molecule. The projection of the surface area on the Cartesian axis is obtained with the numerical algorithm of GEPOL [36].

Therefore, the instantaneous solvent force is randomly applied to those atoms accessible to the solvent until the temperature of the system becomes closest to T_o or a random force has been applied for all the atoms with no no solvent-accessible surface. Moreover, if the temperature of the system is already close to T_o the strengths of the random forces are small and are even zero when $T = T_o$ (where no random force has to be applied). The correlation function of the force $\mathbf{R}$ is calculated at each step of the simulation and the memory function $\mathbf{K}$ is calculated from:

$$\mathbf{K}_{xi} = \langle \mathbf{R}_{xi}(0)\mathbf{R}_{xi}(t)\rangle / 3k_B T, \quad (7)$$

with analogous expressions for the y- and z-axes. For the initial steps of a given simulation the memory is set to zero until meaningful correlations of the force $\mathbf{R}$ can be obtained. At this stage, one point may produce some misunderstanding. The stochastic force (magnitude and direction) does not entirely come from changes on the surface accessible area of the protein atoms. The algorithms used to construct it contain dynamical information on solvent molecules as well as force strength (F_{max}). The colored memory will act locally to damp large velocities that a given atom may acquire; however, as the atom is included in a large body with mobile domains, these latter may move and so the viscous forces will also tend to stop such motions. Regions belonging to mobile parts of a biomolecule may then develop large fluctuations without collapsing onto the core of the biomolecule. Such effects cannot be simulated with simple solvent-accessible-surface effects only.

GROMOS with D4 parameters for the potential-energy function was used for the simulation [21, 22]. The cutoff for the nonbonding potential energy was 8 Å, without use of a switching function, and for the long-range interactions a cutoff of 13 Å and a 2-fs time step were used in the integration, and the bond lengths were constrained by SHAKE [37]. The temperature was set at 3000 K and maintained by the random forces (see earlier). Special care was taken to generate a low-(negative-) energy conformation to initiate the GLD calculations. A 1800 step of steepest descent energy optimization was carried out with the structure; bond lengths were not constrained in the optimization. The GLD calculations were carried out on a Silicon Graphics Origin 2000 of CESCA. The program GROMOS [18] was modified for this study.

The D4-parameter model corresponds to an electroneutral protein. The initially charged groups are neutralized by using the concept of charge polarization [25]. In this manner, the solvent collision effects are absent, while implicit solvent polarization effects are taken into account as potentially charged groups are neutralized. This model is referred to as the in vacuo approach. The results obtained by us with in vacuo on PCI have shown severe limitations in representing the most mobile part of the structure, namely, the N and C terminal domains [21]. This is a case where one may strongly suspect the need for solvent friction effects. In fact, several MD simulations of proteins with the solvent represented with discrete simple-point-charge (SPC) [38] water molecules have shown that the collective motions of protein regions [39, 40, 41], above all those in contact with solvent, have a high friction with respect to the movement in a simulation without explicit solvent [39, 42]. The PCI system is then an appropriate candidate to test the qualities of the GLD approach. An auxiliary frictionless (FLS) simulation, where the last term of Eq. (3) is omitted, was run in order to sense the effect of the friction. This also provides a test for the modeled solvophobic forces on the system on the basis of the derivative of the free energy.

2.3 The protein system

The X-ray crystallographic coordinates [19] of the PCI-IIa isoform [43] (the major one for such a protein) in a complex with CPA are used to seed the GLD simulation of the PCI. The molecular model is identical to the one extensively studied by Oliva and coworkers [21, 22, 23, 24]. From the model, the N terminal residue Glu1 is removed together with the C terminal residue Gly39. For some isoforms the former is lacking, while the latter is cut out from PCI in its complex with CPA. The structure of PCI wild type presents three domains: the core, the N tail and the C tail. The core is defined from residues Cys8 to Cys34, the N tail from residues Gln2 to Ile7 and the C tail from Gly35 to Val38. The NMR conformation of isolated PCI [20] is totally compatible with the extended form found in the PCI–CPA complex and it will be used for the sake of comparison between experimental (NMR plus X-ray data) and modeled (by means of simulations) systems. The regions described earlier will be analyzed by the root mean square (RMS) fluctuations and the RMS deviation (RMSD) fluctuations [23] and compared to the experimental structures (X-ray and NMR [20]) and to the MD models previously obtained for the solvated system [23, 24] and for the nonsolvated system [21, 22].

3 Results and discussion

3.1 Estimators for the equilibrium time and memory function approach

The calculation of the memory function as in Eq. (7) is very time consuming, mainly owing to the correlation function on a long time scale. Therefore, assuming a rapid decay of the memory function [4, 13], this can be expressed as an exponential function ($\mathbf{K}_{t_o}(t) = \mathbf{K}_{t_o} e^{-(t-t_o)/\tau}$) with relaxation time τ [13]:

$$\tau = \int_{t_o}^{\infty} \langle \mathbf{R}_i(t_o) \cdot \mathbf{R}_i(t)\rangle \, dt / \langle \mathbf{R}_i(t_o) \cdot \mathbf{R}_i(t_o)\rangle \ , \quad (8)$$

where the integral is taken here for $t_o = 0$, but, in practice, as the trajectory proceeds different t_o are used as initial points. For a stationary system, $\langle \mathbf{R}_i(t_o) \cdot \mathbf{R}_i(t)\rangle$ is independent of the time origin. In order to know how long it takes for the GLD simulation to give stationary values of the memory function (hereafter defined as equilibrium time), two estimators were used: one for the relaxation time (Θ) and other for K_{t_o} (ζ), defined as

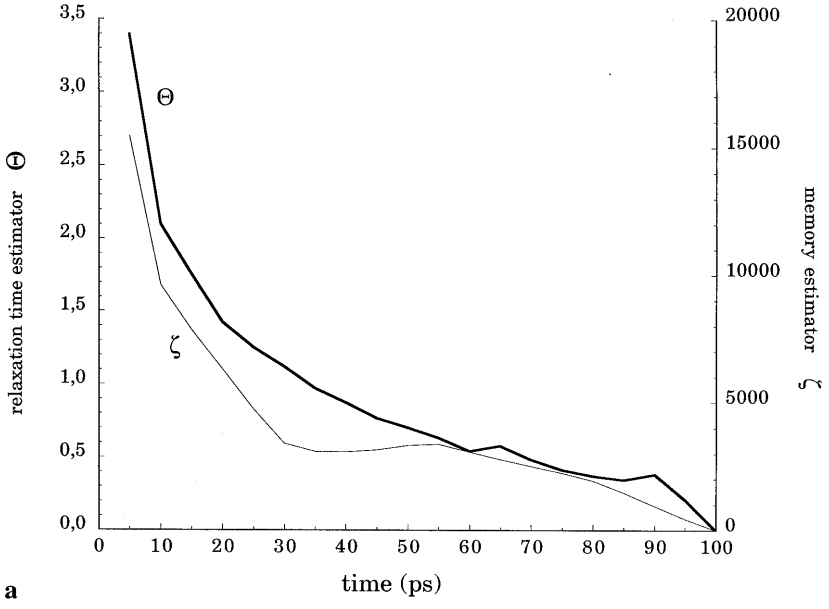
$$\Theta(t_o) = \sqrt{(1/N)\sum_{i=1,N}\left(\tau_i(t_o)_{-\tau i(\infty)}\right)^2}$$

$$\zeta(t_o) = \sqrt{(1/N)\sum_{i=1,N}(\mathbf{K}_{toi} - \mathbf{K}_{\infty i})^2} \ ,$$

where t_o varies parametrically and the values at infinite time are taken from the last point in the given calculation (t 100 ps). The sum is over all atoms having nonzero solvent accessibility surface. The representation of both estimators has to show asymptotic behavior, from which the equilibrium time (∞) can be extracted for the model (i.e., $|\delta\Theta|$ and $|\delta\zeta|$ being smaller than a given $\varepsilon > 0$, or by visual inspection of the graphic). With this approach the integral coefficients of Eq. (7) can be taken from the values of the memory function at the equilibrium time and the correlation function of the random force is not numerically estimated again, thereby saving much computational time.

For the 1-ns trajectory of PCI reported here (see later), these two parameters behave as shown in Fig. 1a. The representation of the two estimators Θ and ζ shows that the stationary position is achieved around 80 ps of GLD simulation. Both estimators present curves asymptotically approaching zero. The perfect equilibrium is achieved when both estimators are equal to zero. Nevertheless, this would imply that only an infinite time of simulation should be taken as the correct equilibrium time; therefore, the limit for the equilibrium for Θ is taken at $\varepsilon = 10^3$ and for ζ the corresponding threshold is taken as $\varepsilon = 0.2$. The values of $K_{t_o}(t)$ are taken from the results of the GLD simulation obtained at the 100-ps step. Between 100 ps and 1 ns the memory functions obtained for each atom were used for the simulation and the correlation of the random forces was not calculated again. Note that at the beginning of a trajectory, the memory function is set to zero, so the solvent dragging effects appear with a delay that is short when compared with the 100-ps trajectory length.

The average dynamical aspects of the method are considered satisfactory. Kinetic energy (temperature coupling) is maintained here by the effect of random forces and not by using Berendsen's algorithm. This leads to fluctuations that are larger than those obtained with a Berendsen bath. The largest fluctuation is around ± 15 kJ/mol (approximately 5 K) and in the equilibrium



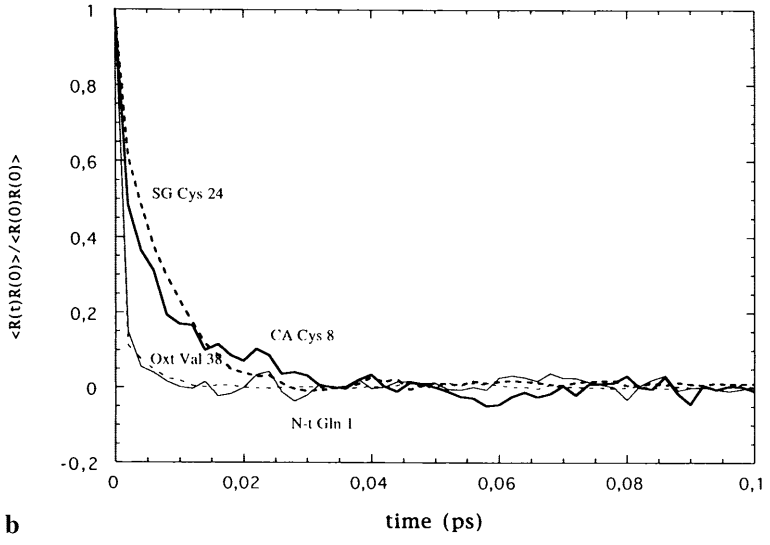


Fig. 1a, b. Analysis of the memory function. **a** Estimators of the equilibrium for the memory function. The *thin line* represents the K(0) estimator and the thick line the relaxation time estimator. **b** Atom stochastic force correlation functions. *Oxt* is the oxygen atom at the C terminal residue Val38

section of the trajectory maintains an average temperature of 287.6 K, about 3 K difference with respect to the targeted $T_o = 300$ K. These results suggest that the system is reasonably well equilibrated with respect to the thermal bath (although not necessarily equilibrated with respect to the accessible configurational space [44]).

The individual atoms accessible to solvent interactions have a different history. At time $t = 0$ (when the trajectory calculation is initiated) they have a $\mathbf{K}(t - t')$ equal to zero. Thereafter, the data are saved and $\mathbf{K}(t - t')$ becomes different from zero. The time dependence of the stochastic force correlation, used to calculate **K** is plotted for selected atoms in Fig. 1b: one from the C tail, another from the N tail and two engaged in residues belonging to the protein core (two cysteines). The exponential decay is quite apparent. The atoms belonging to highly mobile regions (tails) decay faster than those found in zones showing collective fluctuations (core). Figure 1a shows an average property of the protein, while the curves of Fig. 1b represent the solvent effects on short time scales for single atoms.

3.2 Structure, energy and fluctuation behavior

The global structural results of the simulation are qualitatively summarized in Fig. 2, where a series of snapshots are overlayed. The first 100 ps is discarded. The inclusion of surface-mediated effects yields a picture that is very similar to the simulations with the explicit water SPC representation (psW simulation). Both the N tail and the C tail fluctuate around regions populated by the real system as can be inferred from the X-ray and NMR structures.

The standard way to sense the stationary position in a MD trajectory is to study the time series for various energy entries. This is also an issue here, in particular if we take into account that the model is also a representation of a thermal bath.

3.2.1 Potential energy

The time series of the potential energy are shown in Fig. 3. A close analysis shows that the protein becomes equilibrated after 500 ps, with an energy of about −715 kJ/mol. The main energetic components are the electrostatic (−1184 ± 23 kJ/mol) and the van der Waals (−712 ± 41 kJ/mol) energies. The effect of the random forces produces several tensions which are reflected in the energy of the angles (671 ± 12 kJ/mol for bending, 276 ± 13 kJ/mol for dihedrals and 242 ± 7 kJ/mol for improper dihedrals). The total fluctuation of the system is about 37 kJ/mol between 100 ps and 1 ns, the largest fluctuations being found for the van der Waals interaction (about 6% of its energy) and the bending energy (about 4%).

All in all, the system, under the solute–solvent couplings, displays stationary behavior. Given that a new thermal bath is implied in the present approach, the kinetic energy is to be checked now. The algorithms taking care of the solvent also had to answer for thermal equilibration. The stationary position of these entries can be appreciated from Fig. 3. This result is important since no Berendsen bath is used here. The first stringent test is then fulfilled satisfactorily.

3.2.2 RMS fluctuations

The GLD simulation gives rise to a stationary average structure in the 100-ps–1-ns time window of the trajectory. Accordingly, atomic fluctuations along the three main axes of the fluctuation were calculated for this time average. The GLD B factors derived from the matrix of the atomic fluctuations [23] were calculated for the C α atoms of PCI (Fig. 4). For the C tail the B factor is around 60 Å^2, which seems to correlate with the CPA docking and inhibitory function of PCI [21, 23] that requires the necessary mobility of the C tail while still keeping the conformational orientation. In vacuo simulations (absence of solvent collisions and viscosity), N and C tails end up folding onto the core to a great extent. On the other hand, the N-tail conformation became stabilized after 100 ps and presented the largest B factor of the system (about 400 Å^2). The differences with respect to the X-ray structure correlate with the crystal contacts and CPA–PCI contacts analyzed by protein engineering, as reported elsewhere [41].

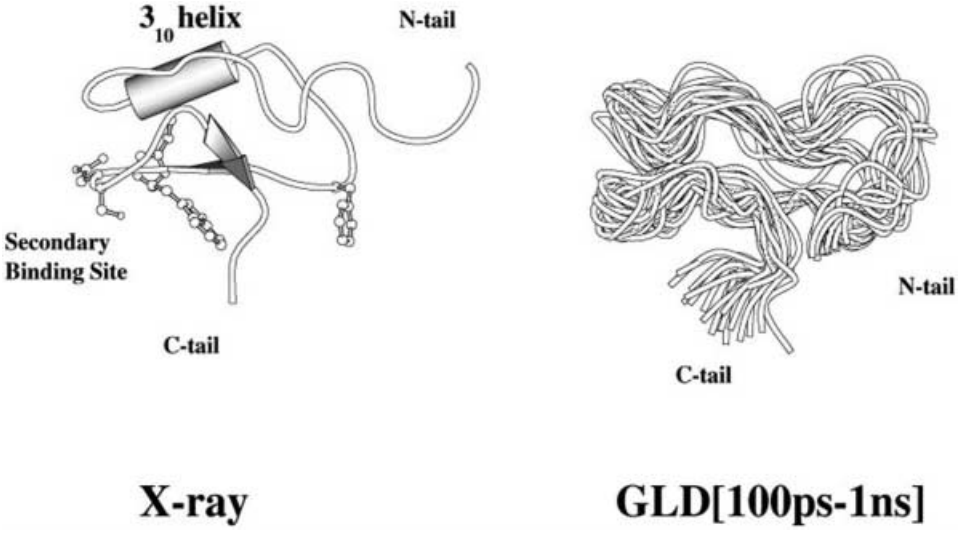


Fig. 2. Ribbon image-plate of the 3D models of potato carboxypeptidase inhibitor (*PCI*). *Left*: PCI X-ray structure showing the three main regions: N tail, core and C tail (see text). *Right*: several conformations of PCI along the generalized Langevin dynamics (*GLD*) simulation taken in snapshots of 50 ps between 100 ps and 1 ns. The images were obtained with PREPI, kindly provided by Dr. Suhail (http://bonsai.lif.icnet.uk)

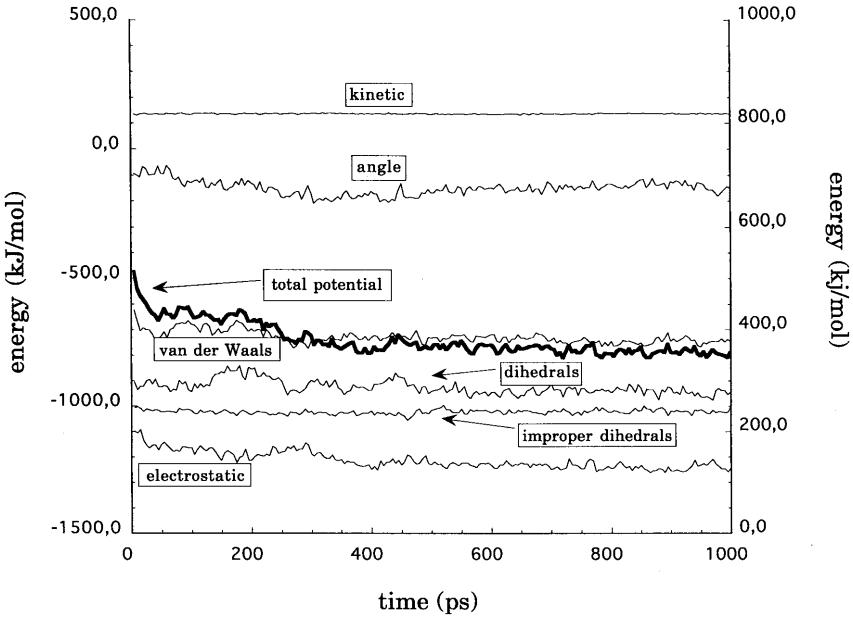


Fig. 3. Time series of the protein potential-energy components of the GLD simulation. The energetic components of the bond interaction (bond angles and bond dihedrals) and the kinetic energy are presented on the *right*. The total potential energy of the protein and the nonbonded interaction (electrostatic and van der Waals) are shown on the *left*

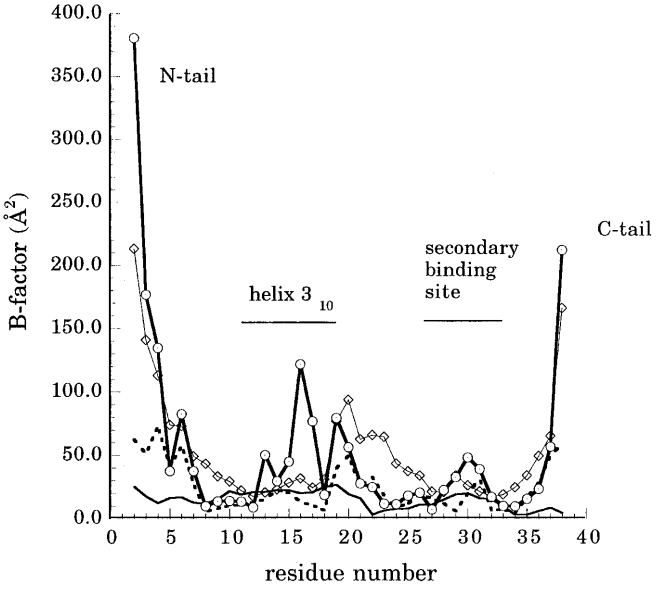


Fig. 4. Comparison of Cα atomic fluctuations of PCI. Crystallographic B factors (*thick line*) of the Cα atoms of PCI are shown with the simulated B factors of the in vacuo model (*dashed line*), the explicit water single-point-charge (*psW*) model (*squares*) and the GLD simulation (*circles*)

3.2.3 RMSD with respect to the X-ray and NMR structures

The averaged RMSDs of the backbone and whole set of atoms of PCI between 100 ps and 1 ns are shown in Table 1. The averaged deviations with respect to the experimental structures of PCI (NMR and X-ray structures) are also shown. The auxiliary simulation with FLS is added to sense the effects of the memory. The experimental coordinates for X-ray and NMR were averaged and are denoted as ⟨Exp⟩. The RMSD is about 3.2 Å when main-chain atoms for the whole protein are used in the comparisons, and it increases to about 4.4 Å for all atoms. In view of the large fluctuations shown by the N and C terminal atoms, such a result is not surprising. In fact, for the core atoms, the RMSD of the main chain deviates by 1.9 Å, while for all atoms yields 3.3 Å. It is apparent that the side chains have somewhat different orientations with respect to either the NMR or the X-ray structures. We note that the GLD simulation compares better with the NMR data than with the crystal structure. As the GLD calculations are designed to represent the protein in solution, this is a positive result.

3.2.4 Comparison with respect to MD models (in vacuo, psW and FLS)

Three auxiliary MD simulations of PCI were performed:

1. The in-solvent (psW) model containing explicit water molecules [24] with a corrected solute–solvent parameter GROMOS force field (C4).
2. A MD model of PCI simulated without explicit solvent [21, 23] using a D4 GROMOS force field (in vacuo).
3. A FLS model derived from the GLD algorithm by eliminating the memory contribution.

Several physical properties analyzed in these MD simulations were compared. First of all, the RMSD (Table 1) with respect to the psW is much smaller for the whole protein when compared to the in vacuo and FLS structures. This suggest that our attempt to simulate the explicit solvent with GLD is satisfactory, although it is certainly not perfect. The comparisons with the average obtained for simulations in vacuo show larger differences, except for the main chain and all-atom core. This situation is encountered again when the FLS run is compared to the average GLD.

The computing time (R8000 processor of Silicon Graphics Indigo Power^2) for the different simulations of

Table 1. Comparison of the root-mean-square deviation(RMSD) for different regions and models of carboxypeptidase inhibitor (*PCI*). The regions considered are the whole PCI, the core of PCI and the C and N terminal tails of PCI, and the RMSD is calculated either for all the atoms or for only the main-chain atoms (Ca, N and C). *X-ray* and *NMR* are the experimentally known structures of PCI, while $\langle Exp \rangle$ is the set constituted by both. *GLD*, *in vacuo* and *psW* are the conformations obtained from the optimization of the average conformation obtained for generalized Langevin dynamics (*GLD*), in vacuo and explicit water single-point-charge (*psW*) simulations, respectively. denotes the set of snapshots taken each 10 ps from the GLD simulation between 100 ps and 1 ns and $\langle MD \rangle$ denotes the set constituted by psW and in vacuo conformations. The simulation with frictionless solvent (*FLS*) is also included

S	Whole PCI RMSD (Å)		PCI core RMSD (Å)		PCI C tail RMSD (Å)		PCI N tail RMSD (Å)	
	Main chain	All atoms	Main chain	All atoms	Main chain	All atoms	Main chain	All atoms
GLD-X-ray	3.20	4.62	2.00	3.40	0.98	1.54	2.59	4.44
GLD-NMR	3.22	4.11	1.85	3.09	0.86	1.49	1.65	2.64
GLD-psW	2.37	3.43	2.19	3.42	0.79	1.27	1.10	2.74
GLD-in vacuo	2.97	4.05	1.67	2.44	2.33	3.76	3.21	4.94
GLD-$\langle$Exp$\rangle$	3.21 ± 0.01	4.36 ± 0.25	1.93 ± 0.08	3.25 ± 0.16	0.92 ± 0.06	1.52 ± 0.03	2.12 + 0.47	3.54 + 0.90
GLD-$\langle$MD$\rangle$	2.67 ± 0.30	3.74 ± 0.31	1.93 ± 0.26	2.93 ± 0.49	1.56 ± 0.77	2.51 ± 1.25	2.16 ± 1.05	3.84 ± 1.10
GLD-$\langle$Exp$\rangle$	2.66 ± 0.37	3.88 ± 0.44	1.48 ± 0.49	2.63 ± 0.43	1.13 ± 0.6	2.87 ± 1.43	2.19 ± 0.25	3.59 ± 0.12
$\langle$GLD$\rangle$-X-ray	3.30 ± 0.32		2.01 ± 0.17		0.69 ± 0.20		2.28 ± 0.10	
$\langle$GLD$\rangle$-NMR	2.86 ± 0.39		2.05 ± 0.45		0.57 ± 0.22		1.18 ± 0.15	
$\langle$GLD$\rangle$-GLD	1.67 ± 0.52		1.42 ± 0.45		0.60 ± 0.28		0.70 ± 0.27	

Table 2. Computation time comparison. Simulation times required to calculate simulations of PCI: FLS; psW (with explicit solvent); in vacuo; GLD_0 (GLD simulation during the first 100 ps in which the memory kernel parameters are calculated); and GLD_f (GLD simulation where the kernel memory is taken from the first 100 ps)

	Number of steps	Time step (fs)	Number of atoms	Computation time (s)
In vacuo	500	2	349	291
PsW	500	2	5376	8724
FLS	500	2	349	519
GLD_0	500	2	349	53567
GLD_f	500	2	349	3171

PCI that are considered for comparison are given in Table 2. The computational time measured to calculate the GLD simulation and the memory function (GLD_0 in Table 2) is far larger than for the simulation with explicit solvent (psW). Nevertheless, after the equilibrium has been achieved and by using the representation of the memory kernel obtained, the GLD simulation is faster (GLD_f in Table 2). The number of atoms used in the simulations is the same for all, except for simulation psW (that increases the number of atoms because of explicit water molecules).

Simplified representations of the 3D-modeled structures of PCI are shown in Fig. 5. These were obtained by optimization of the averaged structures extracted from the equilibrium time window of MD simulations with discrete solvent (psW) with the GROMOS D4 force field (in vacuo) and the GLD simulation (GLD). Remarkable conformational differences are observed between the in vacuo and the X-ray structures, while the conformations of the psW and GLD models are in better agreement with both experimental structures (X-ray and NMR). Taken together with the snapshots displayed in Fig. 2, the inclusion of memory and stochastic forces effects to represent the solvent can be considered as satisfactory.

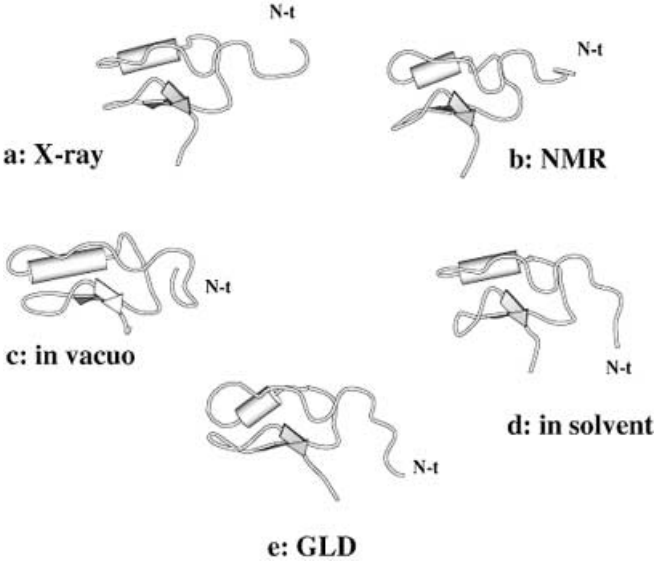


Fig. 5a–e. Simplified representation of the 3D molecular models of PCI. **a** PCI X-ray structure; **b** PCI NMR structure; **c** average conformation of in vacuo simulation; **d** average conformation of the psW simulation; and **e** the average conformation of the GLD simulation

3.2.5 Energy

As noted earlier, the protein's kinetic energy is well maintained in the MD and GLD simulations. Some energy results are compared in Table 3. The total potential energy of the protein showed a large fluctuation for the psW model due to the flux of energy towards the solvent [24]. This effect can also be found in the GLD simulation, the fluctuation being 5% with respect to an averaged energy of about −715 kJ/mol. Therefore, the friction and random forces identified in the generalized LE yield a virtual effect that simulates the energy flux. The total potential energy calculated with the GROMOS force field [18] for the optimized averaged conformation

Table 3. Protein potential-energy components of simulated and experimental models of PCI. Electrostatic and van der Waals energies of the optimized conformations of the experimental structures and the models obtained from simulation. The X-ray and NMR structures are taken for the experimental set of structures, and the averaged conformations of the in vacuo, psW, FLS and GLD for the models are obtained from their respective simulations

	X-ray	NMR	In vacuo	psW	GLD	FLS
Total potential (kJ/mol)	−1893.0	−1915.0	−2262.9	−2026.3	−2136.2	−2175.9
Electrostatic (kJ/mol)	−1259.8	−1245.6	−1461.0	−1300.0	−1395.0	−1404.0
van der Waals (kJ/mol)	−1274.6	−1323.6	−1408.0	−1287.0	−1319.0	−1372.0

Table 4. Comparison of solvent-accessible surface areas (*SASA*). Total solvent-accessible surface areas of PCI and its decomposition into the area due to nonpolar and polar atoms for the optimized experimental structures (X-ray and NMR) and the optimized conformation of the models obtained from simulation (in vacuo, psW, FLS and GLD)

	X-ray	NMR	In vacuo	psW	GLD	FLS
Total SASA (nm^2)	27.87	26.36	23.97	28.17	27.27	25.57
Nonpolar SASA (nm^2)	17.68	17.66	17.83	17.05	17.02	17.29
Polar SASA (nm^2)	10.19	8.7	6.14	11.12	10.24	8.28

in the equilibrium time window (100 ps–1 ns) of the GLD model was −2136 kJ/mol, while for both MD models, psW and in vacuo, the energies were −2026 and −2262 kJ/mol, respectively. This difference confirms that the conformational space scouted by the GLD simulation is within the ranges of the in vacuo and psW simulations. Moreover, the ranges obtained are mainly due to the electrostatic and van der Waals energies (Table 3). From the energetic side, it can be said that the GLD results mimic those obtained with the explicit water simulation (psW).

3.2.6 Solvent-accessible surface area

Table 4 gives the SASA for the averaged and optimized conformations in the equilibrium of the GLD, in vacuo, psW and FLS simulations, these being compared with the SASA calculated for the NMR and X-ray structures. As shown in Table 4, there are small differences in the SASA between the two sets of experimental coordinates. The nonpolar contribution dominates over the polar atoms and the X-ray structure appears to have a little bit more SASA, although for nonpolar atoms both (X-ray and NMR) have the same values. The differences appear for the polar ones. Interestingly, the NMR structure appears to "hide" more polar atoms than the X-ray structure if this latter were to survive in solution dissociated from CPA. The results of simulations in vacuo, where the structure tends to implode, show a decrease in the SASA of the polar atoms below the NMR value. The nonpolar atoms always appear to have the same SASA, irrespective of the data generation. The effect of water is clearly shown by an increase in the SASA of the polar atoms, while the SASA of the nonpolar atoms is conserved. The GLD SASA for the polar atoms is smaller compared to the psW simulation, while it shows a good correlation with the X-ray data. The effect of the memory function can be sensed from the FLS results. The SASA of the polar atoms moves towards the in vacuo simulation. One can clearly see that it makes a difference to have a memory function in the simulation or not.

4 Conclusions

A GLD method has been developed and applied to the study of a protein system. A specific algorithm was developed in order to define the stochastic random force and to keep the system at constant temperature. The main objective was to achieve a methodology that is able to reproduce the physical properties obtained by a MD simulation with explicit solvent by using external forces whose form is derived from the general theory. Besides the fact that computing time may be saved, particularly at long times and for large systems, the success of such a strategy may help the development of further refined computational programs for studying surrounding-medium effects of varying complexity on large biomolecules. From the exhaustive analyses presented here, the feasibility of such an endeavor may be concluded. The results are far from being perfect, particularly for the flexible parts of proteins; however, if we consider the possibility of simulating, for instance, a membrane-surrounding medium or any other specific solvent at different temperatures and pressures, procedures such as the GLD may produce reliable results.

The method presented here, compared to simulations with explicit solvent, will speed the simulation of large protein-solvated systems. This is because the increase in the number of protein atoms involves a large increase in the number of surrounding water molecules (i.e., a globular protein molecule of about 300 residues, such as CPA, has to be embedded in about 50,000 water molecules); therefore, the number of interactions to be calculated increases exponentially in the presence of water, while for the GLD simulation the lack of explicit water molecules reduces exponentially the number of calculations for nonbonded atom–atom interactions. On the other hand, the algorithm can be modified in order to perturb the system by changing some of the parameters used to calculate the random force. This perturbation can be used to obtain the simplest way to explore a simulated unfolding under "hydrophobic" conditions or in high-viscosity media. Besides, the method uses the random force to perturb the system and maintain the temperature, while the system itself is maintained by

frictional forces that do not constrain the conformational space. This implies that the configurational space can be explored fast and securely with neither disturbing additional elements (such as high temperature or additional dimensionality) nor experimental constraints (such as distance constraints from NMR or an electron density map from X-ray studies).

Acknowledgements. The authors wish to acknowledge L. Wesson and D. Eisenberg for providing their subroutines to calculate the solvation energy. B.O. wishes to thank X. Mora from the Mathematics Department and D. Jou from the Thermodynamics Department of the Universitat Autònoma de Barcelona for their help with the development of the algorithms, and M. Orozco and F.J. Luque from the Biochemistry Department of the Universitat de Barcelona and E. Guardia from the Physics Department of the Universitat Politècnica de Catalunya for their helpful discussions on the algorithms. This work was supported by CICYT (Ministerio de Educación y Cultura, Spain, grant BIO98–0362 and BIO97–511), by CERBA (Generalitat de Catalunya, Spain) and by CESCA/C4 (Fundació Catalana de la Recerca). O.T. thanks NFR for financial support.

References

1. Mori H (1965) Prog Theor Phys 33: 423–455
2. Kubo R (1966) Rep Prog Phys 29: 255–284
3. Zwanzig RW (1960) J Chem Phys 33: 1338–1341
4. Berne BJ (1977) In: Berne BJ (ed) Statistical mechanics. Part B: Time-dependent processes. Plenum, New York, pp 233–257
5. Evans M, Evans GJ, Coffey WT, Grigolini P (1982) Molecular dynamics. Wiley, New York
6. Moss F, McClintock PVE (1989) In: Moss F, McClintock PVE (eds.) Noise in nonlinear dynamical systems. Cambridge University Press, Cambridge, pp 1–72, (Chapter 1: Theory of continuous Fokker-Planck systems)
7. Sancho JM, San Miguel M (1989) In: Moss F, McClintock PVE (eds) Noise in nonlinear dynamical systems. Cambridge University Press, Cambridge, pp 72–109
8. Berendsen HJC, Postma JPM, van Gunsteren WF, DiNola A, Haak JR (1984) J Chem Phys 81: 3684–3690
9. Van Gunsteren WF (1993) In: Van Gunsteren WF, Weiner PK, Wilkinson AJ (eds) Computer simulation of biomolecular systems. ESCOM, Leiden, pp 3–36
10. Tapia O (1992) J Math Chem 10: 139–181
11. He S, Scheraga HA (1998) J Chem Phys 108: 271–286
12. He S, Scheraga HA (1998) J Chem Phys 108: 287–300
13. Rey R, Guardia E, Padrós JA (1992) J Chem Phys 97: 8276–8284
14. Israelachvili J (1991) Intermolecular forces and surface forces. Academic, London
15. Wesson L, Eisenberg D (1992) Protein Sci 1: 227–235
16. Cramer CJ, Truhlar DG (1991) J Am Chem Soc 113: 8305–8311
17. Eisenberg D, McLachlan AD (1986) Nature 319: 199–203
18. van Gunsteren WF, Berendsen HJC (1987) Groningen Molecular Simulation (GROMOS) Library Manual. BIOMOS, Groningen
19. Rees DC, Lipscomb WN (1982) J Mol Biol 160: 475–498
20. Clore GM, Gronenborn AM, Nilges M, Ryan CA (1987) Biochemistry 26: 8012–8023
21. Oliva B, Wästlund M, Nilsson O, Cardenas R, Querol E, Avilés FX, Tapia O (1991) Biochem Biophys Res Commun 176: 616–621
22. Oliva B, Nilsson O, Wästlund M, Cardenas R, Querol E, Avilés FX, Tapia O (1991) Biochem Biophy. Res Commun 176: 627–632
23. Oliva B, Daura X, Querol E, Avilés FX, Tapia O (1995) Eur Biophys J 24: 89–103
24. Daura X, Oliva B, Querol E, Avilés FX, Tapia O (1996) Proteins 25: 89–103
25. Åqvist J, van Gunsteren WF, Leijonmarck M, Tapia O (1985) J Mol Biol 83: 461–477
26. Einstein A (1956) Investigations on the theory of Brownian movement. Dover, New York
27. Uhlenbeck GE, Ornstein LS (1930) Phys Rev 36: 823–841
28. Wax N (1954) Selected papers on noise and stochastic processes. Dover, New York
29. Chandrasekhar S (1943) Rev Mod Phys 15: 1–89
30. Cossi M, Mennucci B, Cammi R (1996) J Comput Chem 17: 57–73
31. von Feyberg B, Braun W (1993) J Comput Chem 14: 510–521
32. Canales M, Padró JA (1988) Mol Simul 1: 403–414
33. Trullàs J, Giró A, Padró JA (1990) J Chem Phys 93: 5177
34. van Gunsteren WF, Berendsen HJC (1988) Mol Simul 1: 173–185
35. Yun-Yu S, Lu W, van Gunsteren WF (1988) Mol Simul 1: 369–383
36. Pascual-Ahuir JL, Silla E, Tomasi J, Bonacorsi R (1987) J Comput Chem 8: 778–787
37. Ryckaert J-P, Ciccotti G, Berendsen HJC (1977) J Comput Phys 23: 327–341
38. Berndsen HJC, Postma JPM, van Gunsteren WF, Hermans J (1981) In: Pullman B (ed) Intermolecular forces. Reidel, Dordrecht, pp 331–342
39. Hayward S, Kitao A, Hirata F, Go N (1993) J Mol Biol 234: 1207–1217
40. Jackson MB (1993) J Chem Phys 99: 7253–7259
41. Oliva B, Marino C, Daura X, Molina MA, Avilés FX, Querol E (1995) J Mol Model 1: 54–67
42. Pierleoni C, Ryckaert JP (1991) Phys Rev Lett 66: 2992–2995
43. Hass GM, Ryan CA (1982) Methods Enzymol 80: 778–791
44. Troyer JM, Cohen FE (1995) Proteins 23: 97–110