Coarse-Grained Model of Proteins Incorporating Atomistic Detail of the Active Site

Marilisa Neri,[1] Claudio Anselmi,[1] Michele Cascella,[2] Amos Maritan,[3] and Paolo Carloni[1,*]

[1]*SISSA/ISAS and INFM-DEMOCRITOS Modeling Center, Via Beirut 4, I-34014 Trieste, Italy*
[2]*Ecole Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland*
[3]*Dipartimento di Fisica and INFN, Università degli Studi di Padova, Via Marzolo 8, I-35131 Padova, Italy*

(Received 20 April 2005; published 16 November 2005)

We present a novel approach to explore the conformational space of globular proteins near their native state. It combines the advantages of coarse-grained models with those of all-atoms simulations, required to treat molecular recognition processes. The comparison between calculated structural properties with those obtained with all-atoms molecular dynamics simulations establishes the accuracy of the model. Our method has the potential to be extended to molecular recognition processes in systems whose characteristic size and time scale prevent an analysis based on all-atoms molecular dynamics.

DOI: 10.1103/PhysRevLett.95.218102 PACS numbers: 87.14.Ee

All-atoms molecular dynamics (MD) simulations are a very powerful tool to predict structural, dynamical, and thermodynamical properties of biological molecules [1]. Unfortunately, the current computational power constrains this analysis to time scales of ~100 ns, too short to follow several important biological processes, such as ligand-protein recognition, protein-protein interactions, signaling, etc., which evolve on a much longer time scale. In addition, the number of degrees of freedom of biological systems is very large, and an appropriate exploration of the phase space is possible only if a small number of approximate reaction coordinates can be identified [2,3].

To bridge the gap between time scales of feasible simulations and those of biologically relevant motions, several simplified methods [4] have been proposed. As most of the degrees of freedom of a system are generally associated with the solvent, hybrid models combining an explicit molecular mechanical (MM) treatment for the solute (i.e., the biomolecule) and a continuum model for the environment have been put forward. These methods have been successfully used for evaluating solvation free energy of DNA, RNA, and proteins complexes in aqueous solutions [5] and more recently have been extended to the study of peptides in biological membranes [6]. Alternatively, coarse-grained (CG) models have been proposed in which a small number of interaction sites is used to represent the systems. These approaches have been able to describe bulk water [7], phospholipid bilayers [8], and to follow complex processes such as the fusion between two liposomes [9] and the interaction of a nanotube with the cell membrane [10].

A key contribution in CG simulations of globular proteins has been provided by the observation that the overdamped dynamics of a protein in its solvent can be described as occurring in a highly dimensional quadratic effective potential [11–14]. Consequently, CG models have been introduced, in which the potential energy is expressed in terms of harmonic springs between spatially close effective centroids representing aminoacids, located on the C_α and/or C_β atoms. Several properties calculated with these approaches agree well with experimental and/or MD data [15–20], in spite of the very modest computer resources generally required. They have been successfully applied to the study of the dependence of the folding time of proteins on the number of amino acids [21]. Unfortunately, since the chemical details of the protein sequence are neglected, these models cannot be used to investigate processes that involve molecular recognition, which are crucial for protein functionality and pharmacological applications. Such processes, however, are usually highly localized and involve small portions of proteins.

Here we present a method that overcomes some of the drawbacks of CG models of globular proteins that is a hybrid approach (MM/CG), in which the small biologically relevant region of the protein is treated at the level of detail allowed by classical MD, while the rest of the protein is treated at the CG level, by only considering C_α centroids. An interface region is located between the two MM and CG regions, bridging the large discontinuity between full-atom and CG descriptions (Fig. 1).

We tested our method on two proteins of great pharmacological relevance, belonging to the aspartic protease class: the HIV type 1 virus aspartic protease (HIV-1 PR) [22,23] and the human β-secretase (BACE) [24,25]. The first is a major target for anti-AIDS therapy [26], while BACE plays a role in the progression of Alzheimer's disease [27]. The biological function of these proteins is the catalyzed hydrolysis of peptide chains at specific locations. The dynamical peculiarities of these two proteins make them a sort of ideal benchmark for our model. In fact,

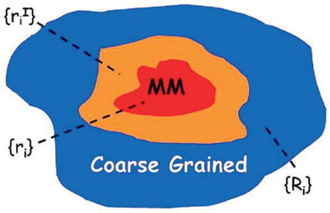

FIG. 1 (color). Schematic representation of the regions described by the MM/CG model.

first, large-scale motions of these proteins have been well characterized both at an atomistic [23,25] and a CG level [19,20]; second, these calculations have shown that their large-scale motions are crucially coupled to their enzymatic activity [23,25]; finally, in spite of the identical cleavage site (a dyad composed of two aspartic residues), these proteins exhibit a large structural diversity: HIV-1 PR is a homodimer containing mostly β strands, while BACE is a monomer with both α and β secondary structure elements.

Our MM/CG simulations were able to reproduce both the mesoscopic [i.e., the residue root mean square fluctuations (RMSF) and the principal normal modes] and the local microscopic details (i.e., distances between key atoms in the active sites and H-bond patterns) of the two proteins, suggesting that our method can be conveniently applied to those systems for which either the size or the necessity of long-time sampling prevents the application of standard MD techniques.

In our method, the MM region (set of atoms $\{\mathbf{r}_i\}$) is treated using a standard molecular force field, the CG region (set of atoms $\{\mathbf{R}_i\}$) using a simplified Go potential [12] and the interface region (I, set of atoms $\{\mathbf{r}_i^{\mathrm{I}}\}$), located between the first two (see Fig. 1), is also treated with the MM force field. Solvent-protein interactions are only treated in terms of viscosity and environment random forces in the framework of stochastic dynamics (SD) [28]. Within our approach, the total potential energy of the system reads:

$$V = E_{\mathrm{MM}} + E_{\mathrm{CG}} + E_{\mathrm{I}} + E_{\mathrm{MM/I}} + E_{\mathrm{CG/I}} + E_{\mathrm{SD}}, \qquad (1)$$

where the first three terms represent, respectively, the interactions within the MM, CG, and I regions, whereas the fourth and fifth represent the cross-terms potentials. The last term, E_{SD}, mimics the stochastic and frictional forces acting on the system, due to the solvent, proportional to the particle velocity and mass [28,29]. E_{MM} is represented by the GROMOS96 43a1 force field, with only polar hydrogens explicitly considered [30]:

$$E_{\mathrm{MM}} = E_{\mathrm{bond}} + E_{\mathrm{vdW}} + \sum_{i>j} \frac{q_i q_j}{\varepsilon|\mathbf{r}_i - \mathbf{r}_j|}, \qquad (2)$$

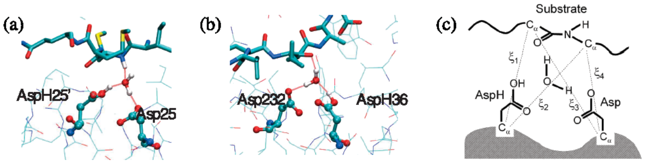

FIG. 2 (color). HIV-1 PR (a) and BACE (b) active sites. The balls and sticks representation depicts the aspartic dyad with the catalytic water, whereas the licorice representation depicts the substrate (MM region). Red lines represent the H-bond network. The line representation depicts a portion of I region. (c) Schematic representation of the catalytic site for aspartic proteases; ξ_i indicates the distances between the C_α atoms of the aspartic dyad and those of the closest two residues of the substrate.

where E_{bond} and E_{vdW} represent, respectively, the total bonded and van der Waals interactions inside the MM region, whereas the last term is the Coulomb interactions between the ith and the jth atoms of charges q_i and q_j, respectively. Electrostatic interactions are unscreened ($\varepsilon = 1$). E_{CG} takes the following form:

$$E_{\mathrm{CG}} = \frac{1}{4}\sum_i K_b(|\mathbf{R}_i - \mathbf{R}_{i+1}|^2 - b_{ii+1}^2)^2 + \sum_{i>j} V_0\{1 - \exp[-B_{ij}(|\mathbf{R}_i - \mathbf{R}_j| - b_{ij})]\}^2. \qquad (3)$$

The first term in Eq. (3) takes into account bonded interactions between consecutive CG (C_α) centroids, identified by the position vectors $\mathbf{R}_i$ and $\mathbf{R}_{i+1}$, and K_b is the relative bond force constant [31]. b_{ij} is the equilibrium distance, corresponding to the native distance between CG atoms. The second term in Eq. (3) describes the nonbonded interactions between CG atoms. V_0 is the interaction well depth [32]. B_{ij} is the modulating exponent of the Morse potential. In region I, all atoms are explicitly considered, as in the MM part, and both E_{I} and $E_{\mathrm{MM/I}}$ energy terms have the same form of Eq. (2).

At the interface between I and CG regions, we have to ensure the protein backbone integrity. Thus, we impose bonds between consecutive C_α belonging to I and CG regions. Similarly to the first term in Eq. (3), we take

$$E_{\mathrm{CG/I}}^{\mathrm{bonded}} = \frac{1}{4}\sum_{i,j} K_b(|\mathbf{r}_{i,C_\alpha}^{I} - \mathbf{R}_j|^2 - b_{ij}^2)^2. \qquad (4)$$

To this, a term describing the nonbonded interactions is added. It reads:

$$E_{\mathrm{CG/I}}^{\mathrm{nonbonded}} = \frac{1}{2}\sum_{i\in[C_\alpha,C_\beta],j} V_0\{1 - \exp[-B_{ij}(|\mathbf{r}_i^{\mathrm{I}} - \mathbf{R}_j| - b_{ij})]\}^2, \qquad (5)$$

where the interface ith atom is either a C_α or a C_β atom and the factor 1/2 stands for the interaction energy equally distributed between the two types of atoms. All the coefficients are chosen similarly to those in Eq. (3).

HIV-1 PR and BACE MM regions include all residues that play a crucial role in the enzymatic catalysis, namely, the aspartic dyad (Asp25 and AspH25' for HIV-1 PR and AspH36 and Asp232 for BACE), the catalytic water molecule, and the substrate in the active site (Fig. 2) [23,25]. Three parameters [32,33] were calibrated so as to reproduce the RMSF of HIV-1 PR, calculated with MD and a CG model [19] [Fig. 3(a)]. In particular, a very critical task is the choice of the interface thickness [33]. The interface region has to guarantee the geometrical positions and orientation of MM residues, the local electrostatics, and to transmit the modes of vibration of the remainder of the protein, which is mimicked by the CG region. We have then performed MM/CG simulations on BACE and the comparison between our results and those available for BACE and HIV-1 PR (except, of course, the RMSF of

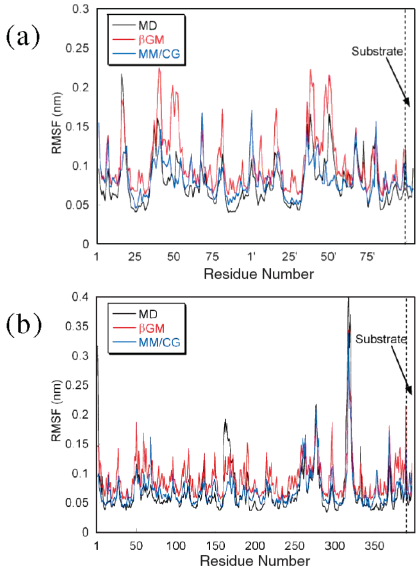

FIG. 3 (color). RMSF of HIV-1 PR (a) and BACE (b), calculated by means of MD simulations, βGM model calculations, and MM/CG simulations.

the latter) constitutes an appropriate test for the general validity of our proposed MM/CG force field. As a reference CG approach, we chose to adopt the β-Gaussian model (βGM) [19], which has been shown to reproduce vibrational modes of both HIV-1 PR and BACE [19,20]. In addition, we tested whether the MM/CG simulations were able to reproduce local structural features, which can be studied by MD, but not by standard CG models.

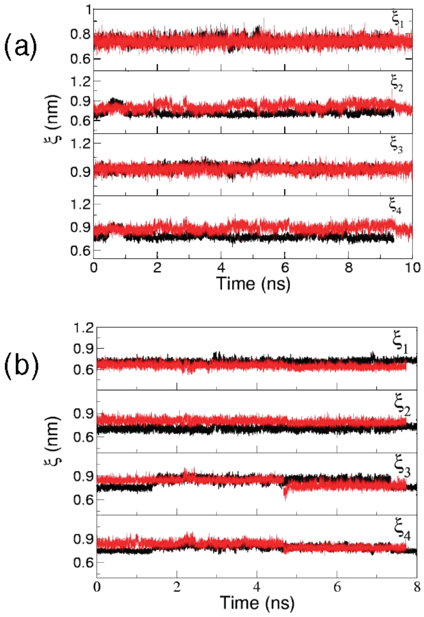

FIG. 4 (color). Time evolution of the distance between the substrate and the active site for HIV-1 PR (a) and BACE (b). ξ_i are defined as in Fig. 2(c). Black and red lines refer to MD and MM/CG simulations, respectively.

Figure 3(b) depicted the RMSF of BACE, computed by classical MD, βGM, and our MM/CG simulations [34]. As shown in the figure, MM/CG data follow the trend of MD simulations with a straightforward correlation for both proteins. In general, our data agree better than the βGM ones with the MD results. To quantify the accuracy of the vibrational modes of the two proteins obtained by MM/CG simulations and compare them with βGM calculations, we have represented the meaningful MM/CG (βGM) eigenvectors, $\mathbf{v}_i$, in terms of the largest N MD eigenvectors, $\mathbf{v}_j^{\mathrm{MD}}$, of the corresponding dynamic cross-correlation matrices [36]. These matrices, calculated for the C$_\alpha$ atoms, provide information on the degree of correlation between pairs of residues at equal time. $\mathbf{v}_i$ reads:

$$\mathbf{v}_i = \sum_{j=1}^{N} c_j^{(i)} \mathbf{v}_j^{\mathrm{MD}}. \tag{6}$$

In particular, we considered the value C_i, defined as the square root of the summation over j of the first N $c_j^{(i)}$. In this way, it is possible to quantify the ability of the MD eigenvectors to represent the largest MM/CG or βGM ones. The calculated high values of C_i (Table I) indicate that the subspace of the most relevant eigenvectors computed with the MM/CG model and MD almost coincide. For instance, in the case of HIV-1 PR, $C_1 = 0.95$ means that 95% of the first MM/CG eigenvector overlaps with the overall motion of the MD simulation, if the MD trajectory is projected onto the relative first 20 normal modes (i.e., $N = 20$). The results are similar to those of βGM, especially for the most important eigenvectors. βGM calculations seem to perform better than MM/CG for HIV-1 PR, but not in the case of BACE.

As a test prediction accuracy of the microscopic dynamical features, i.e., the chemical interactions in MM regions, we focused on the motion of the substrate in the active site of the two enzymes, which has been shown to play a functional rule. In fact, the catalytic activity of both HIV-1 PR and BACE is directly related to the distance

TABLE I. First 10 C_i, calculated as the square root of the sum of the squared principal MD eigenvectors decomposition coefficient, in the case of $N = 20$.

	HIV-1 PR		BACE	
	MM/CG	βGM	MM/CG	βGM
C_1	0.95	0.97	0.81	0.23
C_2	0.96	0.98	0.86	0.86
C_3	0.92	0.96	0.78	0.87
C_4	0.81	0.92	0.85	0.67
C_5	0.81	0.88	0.87	0.82
C_6	0.85	0.82	0.66	0.84
C_7	0.77	0.88	0.81	0.62
C_8	0.55	0.79	0.75	0.67
C_9	0.55	0.82	0.59	0.64
C_{10}	0.54	0.63	0.65	0.46

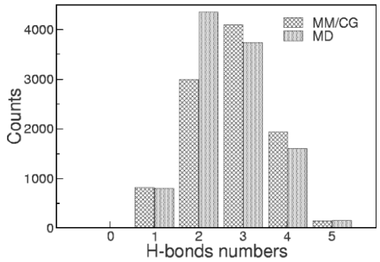

FIG. 5. H-bond number counts between the aspartic dyad along with the catalytic water and the substrate in BACE. Values were derived by MD and MM/CG simulations.

between the catalytic aspartic dyad and the substrate [ξ_{1-4} in Fig. 2(c)]. These distances are directly modulated by the large-scale motions of the protein scaffold [23,25] and fluctuate between two characteristic values in a few ns [23,25]. Figures 4(a) and 4(b) show that MM/CG method is able to reproduce these characteristic distances very well.

Comparing the H bonds in the active site in BACE (Fig. 5) further establishes the accuracy of our method. The correlation between MD and MM/CG data is very high (0.92).

In summary, we have presented a new computational approach for globular proteins, based on a mixed mesoscopic-atomistic description. This approach may provide a reliable way of overcoming the time- or size-limit bottlenecks that constitute one of the major drawbacks of MD simulations. At present, the particular form of our force field is mostly suited for simulating globular proteins with buried active sites; the future challenge will therefore lie in the implementation of efficient schemes for solvent flow in the MM part.

We thank Gianluca Lattanzi, Vincenzo Carnevale, Cristian Micheletti, and Stefano Piana for suggestions and useful discussions. We thank MURST-COFIN for financial support.

*Electronic address: carloni@sissa.it

[1] M. Karplus, Acc. Chem. Res. **35**, 321 (2002).
[2] F. Wang and D.P. Landau, Phys. Rev. Lett. **86**, 2050 (2001).
[3] A. Laio and M. Parrinello, Proc. Natl. Acad. Sci. U.S.A. **99**, 12 562 (2002).
[4] S.O. Nielsen, C.F. Lopez, G. Srinivas, and M.L. Klein, J. Phys. Condens. Matter **16**, R481 (2004).
[5] P.A. Kollman *et al.*, Acc. Chem. Res. **33**, 889 (2000).
[6] W. Im and C.L. Brooks, III, Proc. Natl. Acad. Sci. U.S.A. **102**, 6771 (2005).
[7] A. Malevanets and R. Kapral, J. Chem. Phys. **110**, 8605 (1999).
[8] J.C. Shelley *et al.*, J. Phys. Chem. B **105**, 4464 (2001).
[9] M.J. Stevens, J.H. Hoh, and T.B. Woolf, Phys. Rev. Lett. **91**, 188102 (2003).
[10] S.O. Nielsen *et al.*, Biophys. J. **88**, 3822 (2005).
[11] M.M. Tirion, Phys. Rev. Lett. **77**, 1905 (1996).
[12] T. Noguti and N. Go, Nature (London) **296**, 776 (1982).
[13] S. Swaminathan, T. Ichiye, W. van Gusteren, and M. Karplus, Biochemistry **21**, 5230 (1982).
[14] K. Hinsen, Proteins **33**, 417 (1998).
[15] I. Bahar, A.R. Atilgan, and B. Erman, Folding Des. **2**, 173 (1997).
[16] P. Doruker, A. Atilgan, and I. Bahar, Proteins: Struct., Funct., Genet. **40**, 512 (2000).
[17] A.R. Atilgan *et al.*, Biophys. J. **80**, 505 (2001).
[18] I. Navizet, F. Cailliez, and R. Lavery, Biophys. J. **87**, 1426 (2004).
[19] C. Micheletti, P. Carloni, and A. Maritan, Proteins: Struct., Funct., Genet. **55**, 635 (2004).
[20] M. Neri, M. Cascella, and C. Micheletti, J. Phys. Condens. Matter **17**, S1581 (2005).
[21] M. Cieplak and T.X. Hoang, Biophys. J. **84**, 475 (2003).
[22] M. Miller *et al.*, Science **246**, 1149 (1989).
[23] S. Piana, P. Carloni, and M. Parrinello, J. Mol. Biol. **319**, 567 (2002).
[24] L. Hong *et al.*, Science **290**, 150 (2000).
[25] M. Cascella, C. Micheletti, U. Rothlisberger, and P. Carloni, J. Am. Chem. Soc. **127**, 3734 (2005).
[26] M. Boffito *et al.*, Antivir. Ther. **10**, 375 (2005).
[27] S.J. Pollack and H. Lewis, Current Opinion in Investigational Drugs **6**, 35 (2005).
[28] M. Doi, *Introduction to Polymer Physics* (Oxford Science Publications, Great Britain, 1996), 1st ed..
[29] M. Doi, *Introduction to Polymer Physics* (Oxford University Press, Oxford, Great Britain, 1996), 1st ed..
[30] W.F. van Gunsteren *et al.*, *Biomolecular Simulation: The GROMOS96 Manual and User Guide* (Hochschulverlag AG an der ETH Zurich, Zurich, Switzerland, 1996).
[31] We tentatively chose $K_b = 7.2 \times 10^6$ kJ/mol nm^{-4}, the typical force constant of a bond between sp^3 carbons in the GROMOS96 force field [30].
[32] The interaction well depth, V_0, was conveniently chosen equal to 5.3 kJ mol^{-1}. The nonbonded interactions are computed between CG atoms within a cutoff distance of $r_{\mathrm{cut}} = 1.0$ nm.
[33] In our simulations, a residue belongs to the interface region if at least one of its atoms is less distant than a cutoff, $r_{\mathrm{int}} = 0.60$ nm. This means that the actual interface thickness is much larger, in the order of ~1.0 nm.
[34] All simulations were performed using Gromacs 3.2 program [29]. The whole systems were composed of 429 and 984 particles for HIV-1 PR and BACE, respectively. The leapfrog stochastic dynamics algorithm was used to integrate the equations of motion with a time step $\Delta t = 2$ fs. The SHAKE algorithm [35] was used to keep fixed the distance of bonds containing hydrogen(s). The friction coefficient was computed as $\gamma_i = m_i/\tau$, where $\tau = 0.5$ ps is the time constant for the coupling and m_i is the mass of ith particle. No cutoff distance was used for the electrostatic and vdW interactions. The MM/CG simulations were carried out for the time scale of MD simulations: ~10 ns and 8 ns for HIV-1 PR and BACE, respectively, required few hours on a Intel xeon 3.06 GHz PC.
[35] J.P. Ryckaert, G. Ciccotti, and H.J.C. Berendsen, J. Comput. Phys. **23**, 327 (1977).
[36] B. Hess, Phys. Rev. E **65**, 031910 (2002).