JCTC Journal of Chemical Theory and Computation

United-Atom Discrete Molecular Dynamics of Proteins Using Physics-Based Potentials

Agustí Emperador,[†,‡] Tim Meyer,[†,‡] and Modesto Orozco*[,†,‡,§,||]

Joint IRB-BSC research program in Computational Biology, Institute for Research in Biomedicine (IRB), Josep Samitier 1-5, Barcelona 08028, Spain, Barcelona Supercomputing Centre (BSC), Jordi Girona 29, Barcelona 08034, Spain, Departament de Bioquímica i Biología Molecular, Facultat de Biología, Universitat de Barcelona, Avgda Diagonal 645, Barcelona 08028, Spain, and National Institute of Bioinformatics, Parc Científic de Barcelona, Josep Samitier 1-5, Barcelona 08028, Spain

Received September 12, 2008

Abstract: We present a method for the efficient simulation of the equilibrium dynamics of proteins based on the well established discrete molecular dynamics algorithm, which avoids integration of Newton equations of motion at short time steps, allowing then the derivation of very large trajectories for proteins with a reduced computational cost. In the presented implementation we used an all heavy-atoms description of proteins, with simple potentials describing the conformational region around the experimental structure based on local physical interactions (covalent structure, hydrogen bonds, hydrophobic contacts, solvation, steric hindrance, and bulk dispersion interactions). The method shows a good ability to describe the flexibility of 33 diverse proteins in water as determined by atomistic molecular dynamics simulation and can be useful for massive simulation of proteins in crowded environments or for refinement of protein structure in large complexes.

Introduction

Under native conditions proteins exist not as unique structures but as dynamic ensembles of conformers, some of them which can be quite distant from the most probable structure of the protein. The spontaneous sampling of different conformations helps the proteins to perform their biological action, since most protein actions imply relevant conformational changes. Thus, many studies[1−10] have reported that the biological relevant transition (i.e. that required for biological action) corresponds in many cases to the easiest deformation movements of the relaxed protein, suggesting that evolution has driven proteins to optimize not only their structure but also their intrinsic deformability patterns.[11−16] A better understanding of the dynamical behavior of proteins is then crucial to bridge structure and function in proteins.

Despite recent advances[10] the experimental study of protein flexibility is still very difficult and impossible to carry out at the full proteome-level. Theoretical methods are then the logical alternative, and many examples of the power of these methods have been published in the last years.[17,18] One of the most useful tools to describe protein dynamics is atomistic molecular dynamics (MD), where solvated proteins are represented at full atomistic detail by means of physical potentials[19−25] derived from high level quantum mechanical calculations and experimental data on condensed phase and model systems. The impact of MD in the study of protein flexibility has been impressive.[26,27] However, despite recent improvements in software and hardware, atomistic MD simulations of proteins are still too expensive to allow a full-proteome description of flexibility or to simulate simultaneously the dynamics of hundreds of thousands of proteins in crowded cellular-like environments. Therefore, less rigorous, but more efficient simulation, tools, such as normal

* Corresponding author e-mail: modesto@mmb.pcb.ub.es.

[†] Institute for Research in Biomedicine.

[‡] Barcelona Supercomputing Centre.

[§] Universitat de Barcelona.

[||] National Institute of Bioinformatics, Parc Científic de Barcelona.

10.1021/ct8003832 CCC: $40.75
Published on Web 11/12/2008

mode analysis,[18] Brownian MD,[28] or discrete molecular dynamics[29] (DMD), are necessary.

DMD is an ultrafast but still physically-based technique where interactions are represented with discontinuous potentials with a steplike profile. In the simple case of coarse grained models with strictly structure-based potentials at residue-level resolution, the interactions between protein residues are represented by means of simple infinite square wells.[30] Within this model, the particle moves in ballistic regime (constant velocity) in a flat potential region (within the well), interrupted by collisions at the walls of the potential wells. Upon each collision momentum and energy conservation laws are imposed. The simplicity of the potential functional avoids the need to integrate Newton's equations of motion every femtosecond, since trajectory becomes fully deterministic and is followed from collision to collision. This allows the obtaining of very long trajectories in a simple PC, the performance being increased in diffusive or slow transitional movements. Despite its simplicity, DMD has been successful in describing different aspects of protein dynamics[31−35] and nucleic acids,[36,37] including macromolecular aggregation[38−44] and transitions.[45] The reliability of the DMD discontinuous potentials to describe protein interactions has been shown, for example, in *ab initio* protein folding[46] and protein oligomerization studies.[47−49] Very recently, our group has shown that DMD coupled to a simple Gō-like Hamiltonian[50] (similar to those used in normal mode analysis) is able to reproduce reasonably well the C_α dynamics of proteins,[30] with a very small computational cost, suggesting that DMD can be an alternative to Gō-based Brownian dynamics[28] or normal mode analysis.[18] Unfortunately, as formulated in our previous work the technique presents two intrinsic shortcomings. First, since it is based exclusively on Gō-like potentials it can only reproduce pseudoharmonic deformability around a reference structure, without any possibility of describing large (nonharmonic) conformational transitions. Second, no atomic-detailed information is provided, limiting the range of applicability of the technique in cases where resolution higher than the backbone is needed.

In this paper we present and validate a new atomistic DMD model based on a simple pseudophysical force-field, which represents a hybrid between pure physical potentials[51−57] and empirical Gō-like models linked to strictly harmonic/pseudoharmonic potentials.[50] The method defines the Hamiltonian based on a reference structure, but instead of using all atomic or residue contacts as Gō-restraints, only physically meaningful interactions are considered. The method provides atomic-detailed information on the protein, while guaranteeing that the sampling will not produce protein conformations unrealistically different from the experimentally determined structure. The technique is fast and seems to be a promising tool for the structure refinement, for the representation of moderate conformational transitions, and for the analysis of protein-protein docking, and multiprotein systems.

Methods

Discrete Molecular Dynamics Algorithm. DMD assumes that particles defining a molecule follow fully deterministic trajectories defined by their starting positions and velocities and simple (mono-or multiple) square-well potentials for particle-particle interactions. Particles move with constant velocities which change only after collision with the walls of the wells (see Figure 1 for explanation). As a consequence, trajectory can be analytically determined, integration of equations of motions at fixed time steps (as in MD) is not needed, and the calculation progresses directly from collision

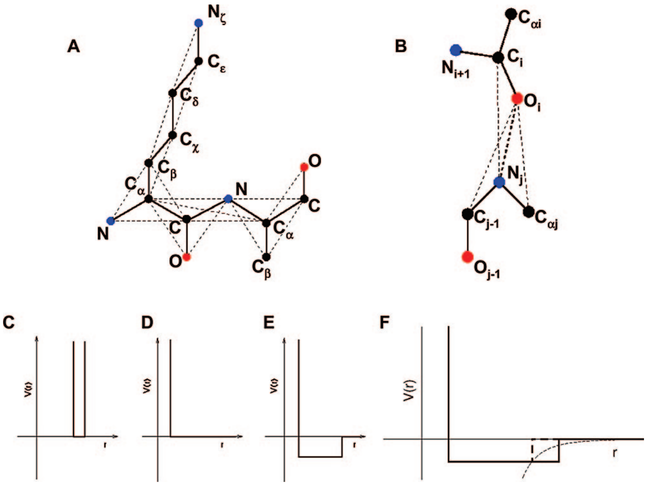

Figure 1. (A) Schematic diagram for the all-atom protein model. A Lys-Ala sequence is shown. Solid thick lines represent covalent bonds, and thin dashed lines show pseudobonds (see text). (B) Schematic diagram for the hydrogen bond. The pseudobonds (dashed lines) keep the directionality of the bond. (C) DMD square well potential for covalently or noncovalently bound particles. In the case of covalent bonding the well depth is infinite. (D) DMD hardcore repulsive potential for unbound particles. (E) Square well for hydrophobic interactions and salt bridges, composed by the hardcore repulsion at short distances and a potential step at the cutoff distance for these interactions. (F) Dispersion potential established between C_αs. The well depth depends on the distance in the native conformation (see text).

Table 1. Different Metrics Indicating the Similarities of DMD Ensembles to Reference Structures (MD-Averaged: MD) or Experimental Conformation (EXP) and with Ensembles Obtained by Atomistic MD Simulations in Water

parameter	DMD vs MD	DMD vs EXP[d]
RMSd[a]	2.6 ± 0.8	3.2 ± 0.7
Tm-score[a]	0.4 ± 0.2	1.5 ± 0.3
R_{gyr}[a,b]	1.1 ± 0.6	0.8 ± 0.6
SAS(tot)[b,c]	12% ± 7%	11% ± 7%
SAS(apolar)[b,c]	10% ± 7%	9% ± 7%
Conserv helix	96% ± 8%	93% ± 12%
Conserv β-sheet	85% ± 12%	85% ± 12%
Conserv turn	96% ± 3%	96% ± 3%
no. of native contacts	80% ± 11%	87% ± 10%

[a] Average RMSd, R_{gyr}, and Tm-scores are in Å. [b] Computed as $dif = ((1/n)\sum_n(\alpha_{ref} - \alpha_{DMD})^2)^{0.5}$, where α is the descriptor, n is the number of proteins in the test set, α_{ref} is the value of the descriptor in the reference conformation (MD-averaged or experimental conformation), and α_{DMD} is the average deviation obtained along the 5-10 ns portion of the DMD trajectory. [c] Solvent accessible surfaces are represented as relative values for every protein. [d] Values in the last column always refer to average relative deviations. All values were obtained considering an all-heavy atom representation of proteins.

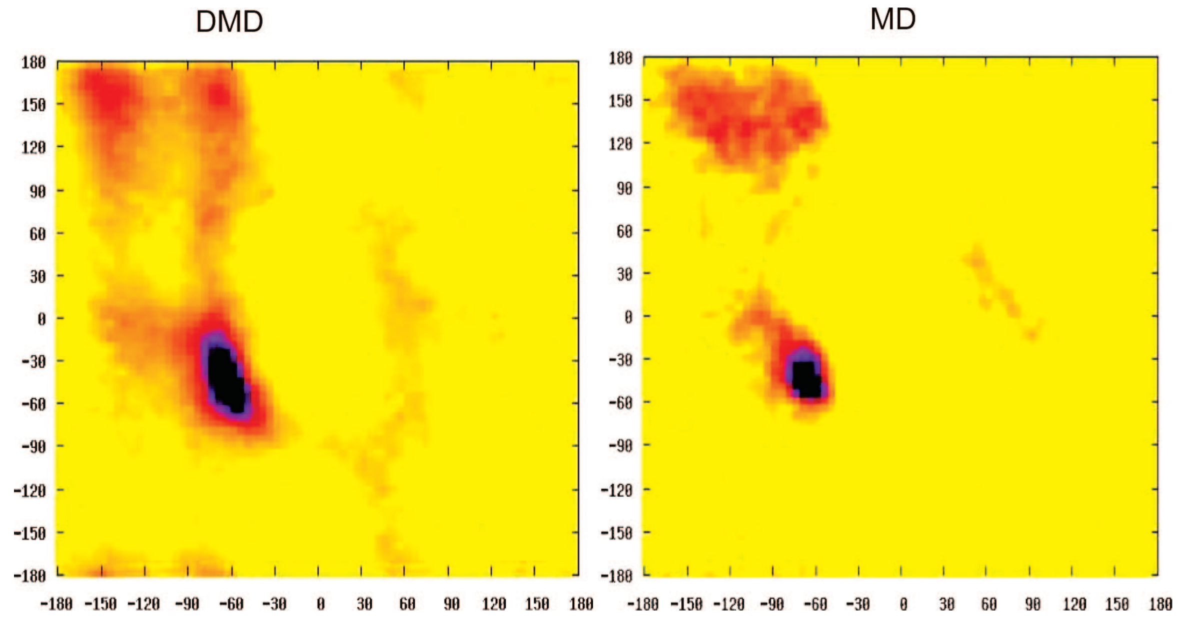

Figure 2. Ramachandran map for the proteins calculated with the current DMD sampling technique and all-atoms MD simulations with AMBER. In both cases 10 ns trajectories were used.

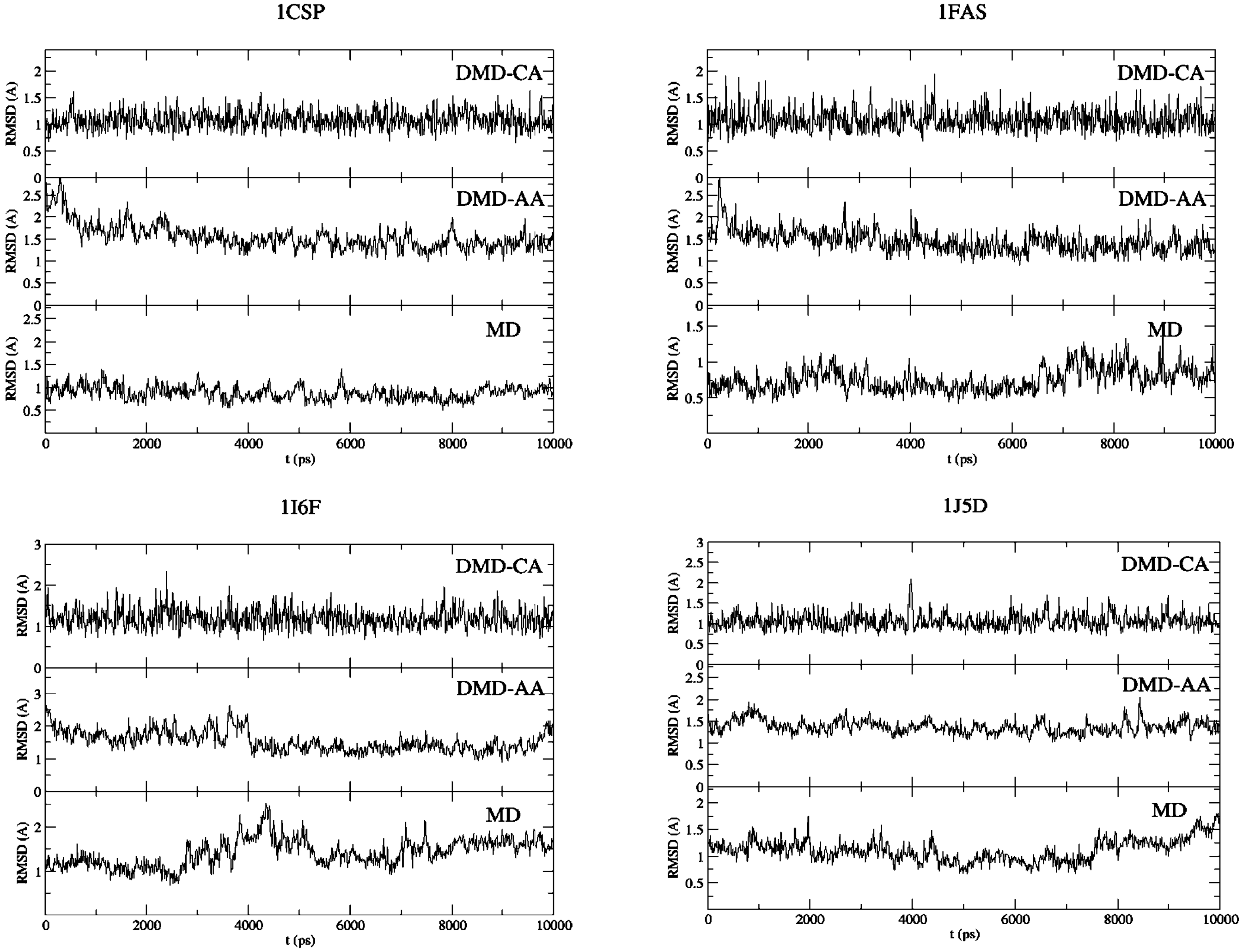

Figure 3. RMSd (in Å^2) with respect to reference structure (MD-averaged one) of selected proteins in DMD using a Gō-like potential and C$_\alpha$ representation of proteins (DMD-CA), present "all atoms" DMD implementation (DMD-AA) and atomistic MD simulation with AMBER-force-field (MD). In all cases only the C$_\alpha$ coordinates are considered to compute the RMSd.

(event) to collision. If an efficient algorithm for predicting collisions is used[58] the method can be extremely efficient allowing simulation of very long time periods. In the present implementation the CPU time needed to determine collisions scales as ~$N^{1.1}$ with the number of particles (N) considered.

The basic DMD formalism assumes that the position of a particle after some period of time (the minimum collision time) is determined analytically using

$$\vec{r}_i(t+t_c)=\vec{r}_i(t)+\vec{v}_i(t)t_c \qquad (1)$$

where $\vec{r}_i$ and $\vec{v}_i$ stand for positions and velocities, and t_c is the minimum amongst the collision times t_{ij} between each pair of particles i and j, given by

$$t_{ij}=\frac{-b_{ij}\pm\sqrt{b_{ij}^2-v_{ij}^2(r_{ij}^2-d^2)}}{v_{ij}^2} \qquad (2)$$

Table 2. Indexes Representative of the Dynamic Behavior Obtained in DMD and MD Samplings[a]

	DMD	MD	MD vs MD
variance (heavy atoms) [b]	25.6 ± 11.4	13.4 ± 6.4	6.1-53.3
reduced variance[b]	6.1 ± 4.5	6.2 ± 4.8	1.8-30.3
entropy (Schlitter, heavy atoms)[b]	56.9 ± 15.2	47.1 ± 2.7	39.9-64.3
complexity[c]	9% ± 2%	4% ± 2%	1 %-8%
Z-score	71.4 ± 24.6	----	62-255
Γ index	0.61 ± 0.13	----	0.8-1.0
correlation B-factors (to exp)	0.43 ± 0.12	0.66 ± 0.10	0.4-0.8
correlation B-factors (to MD)	0.54 ± 0.15	----	0.4-0.9
Δ_L (all)	0.40 ± 0.08	0.32 ± 0.07	0.20-0.57
Δ_L (exposed) - Δ_L (buried)	0.13 ± 0.07	0.18 ± 0.06	0.08-0.30
Δ_L (α-helix) - Δ_L (β-sheet)	−0.19 ± 0.15	−0.12 ± 0.13	−0.33-0.16

[a] In all cases an all-heavy atoms representation of the protein is used, and represented values correspond to averages over 33 proteins. See Methods for description of the different metrics. Values in the column labeled by MD correspond to the meta-trajectory obtained by adding individual CHARMM, OPLS, and AMBER trajectories for individual proteins. The values in the last column (labeled as MD vs MD) correspond to those obtained when instead of the original meta-trajectory the original CHARMM, OPLS, and AMBER trajectories are compared among each other. The last column gives the range of uncertainty expected in atomistic MD simulation as reported by comparing the 4 different trajectories available for each protein. [b] Values used in the average were previously referred to as the number of residues in the protein. [c] Values used in the average were previously referred to as the total number of eigenvectors in the protein.

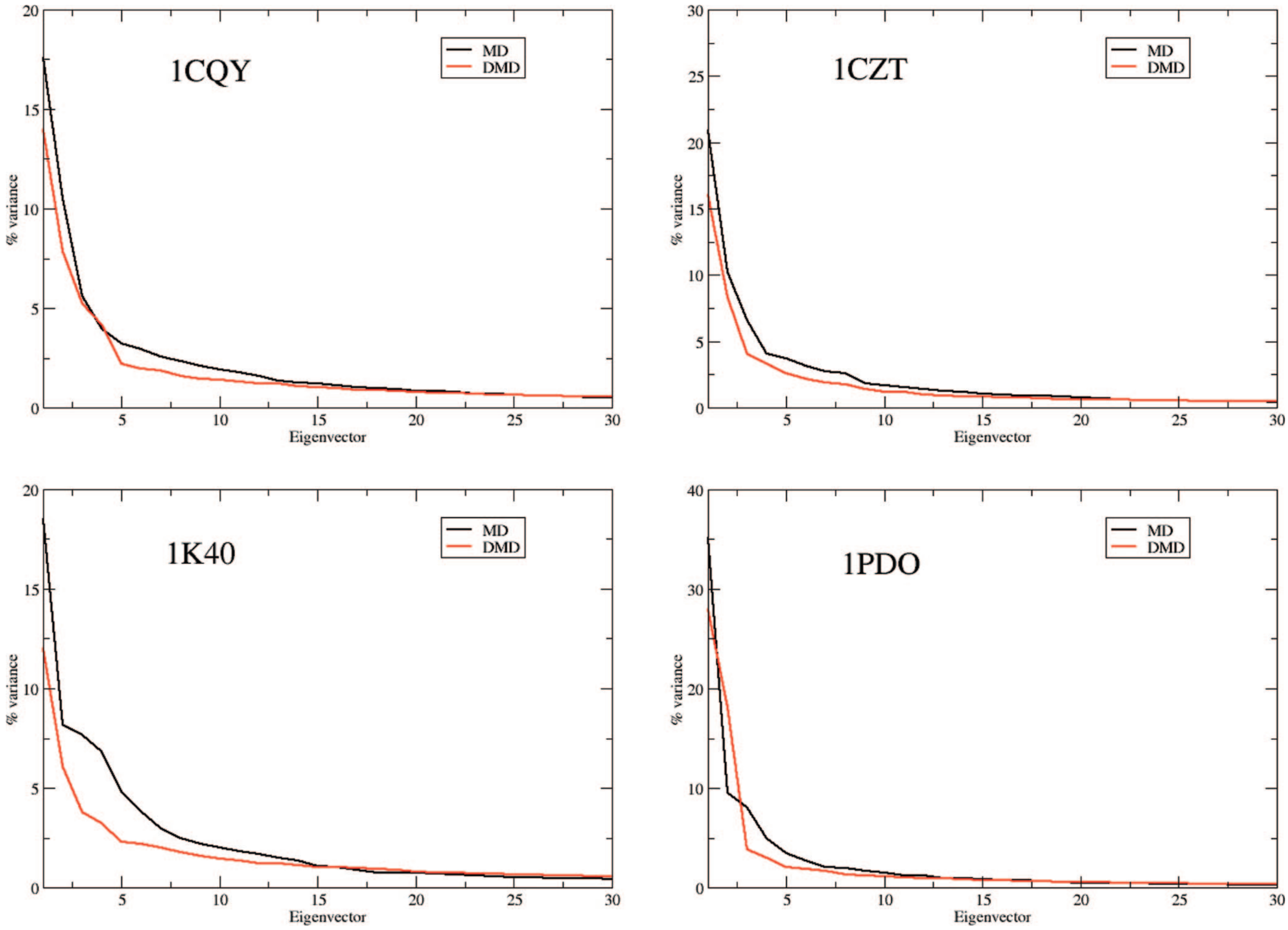

Figure 4. Percentage of the total variance explained by individual eigenvectors of different rank for selected proteins as determined from MD and DMD ensembles.

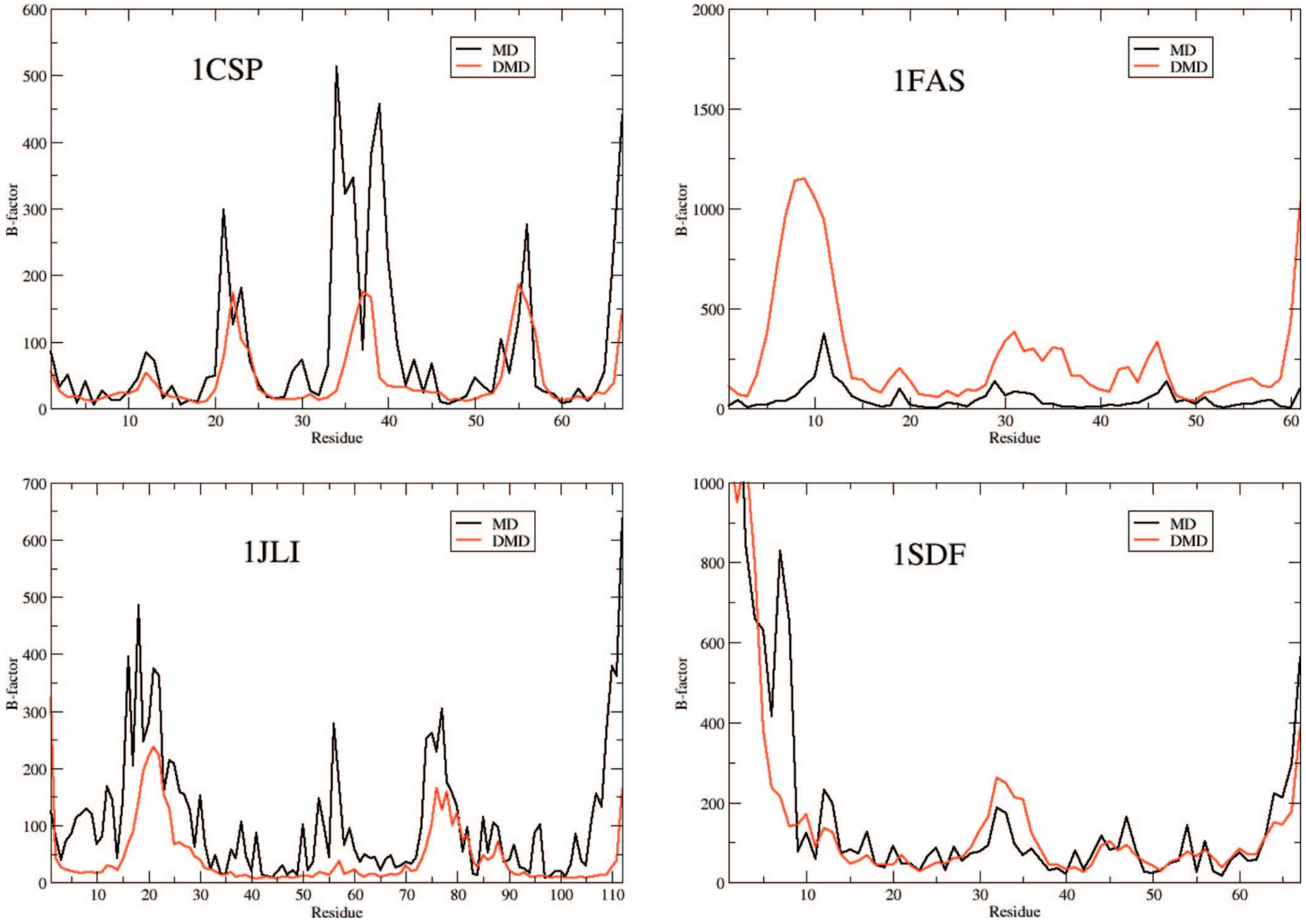

Figure 5. B-factors (in Å^2) for selected proteins as determined from MD and DMD ensembles.

where r_{ij} is the square modulus of $\vec{r}_{ij} = \vec{r}_j - \vec{r}_i$, v_{ij} is the square modulus of $\vec{v}_{ij} = \vec{v}_j - \vec{v}_i$, $b_{ij} = \vec{r}_{ij}\cdot\vec{v}_{ij}$, and d is the distance corresponding to a discontinuity (the wall) in the potential (the signs + and − before the square root are used for particles approaching one another and moving apart, respectively).

When two particles collide, there is a transfer of linear momentum into the direction of the vector $\vec{r}_{ij}$

$$m_i\vec{v}_i = m_i\vec{v}_i{}' + \Delta\vec{p} \tag{3}$$

$$m_j\vec{v}_j + \Delta\vec{p} = m_j\vec{v}_j{}' \tag{4}$$

where the prime indices denote the variables after the event (collision).

In order to calculate the change in velocities upon collision the velocity of each particle is projected in the direction of the vector $\vec{r}_{ij}$ so that the conservation equations become one-dimensional along the interatomic coordinate, which implies that

$$m_iv_i + m_jv_j = m_iv_i{}'\, m_jv_j{}' \tag{5}$$

$$\frac{1}{2}m_iv_i^2 + \frac{1}{2}m_iv_i^2 = \frac{1}{2}m_iv_i'^2 + \frac{1}{2}m_iv_i'^2 + \Delta V \tag{6}$$

where the change in potential energy (ΔV) is the depth of the square well that defines the interatomic potential.

The transferred momentum can be easily determined from

$$\Delta p = \frac{m_im_j}{m_i + m_j}\left\{\sqrt{(v_j - v_i)^2 - 2\frac{m_i + m_j}{m_im_j}\Delta V} - (v_j - v_i)\right\} \tag{7}$$

Note that the two particles can go out of the well as long as

$$\Delta V < \frac{m_1m_2}{2(m_1 + m_2)}(v_j - v_i)^2 \tag{8}$$

Otherwise, the particles do not cross the potential energy discontinuity at the wall of the well and remain inside the well (ΔV=0). In this case eq 7 reduces to

$$\Delta p = \frac{m_im_j}{m_i + m_j}\left\{\sqrt{(v_j - v_i)^2} - (v_j - v_i)\right\} \tag{9}$$

which taking the negative solution of the root leads to the definition of the transferred momentum:

$$\Delta p = \frac{2m_im_j}{m_i + m_j}(v_i - v_j) \tag{10}$$

Since the proteins we want to simulate are considered to be in a thermal bath, but kinetic and potential energy are continuously exchanged during the simulation, we maintain the temperature through a Berendsen coupling.[25]

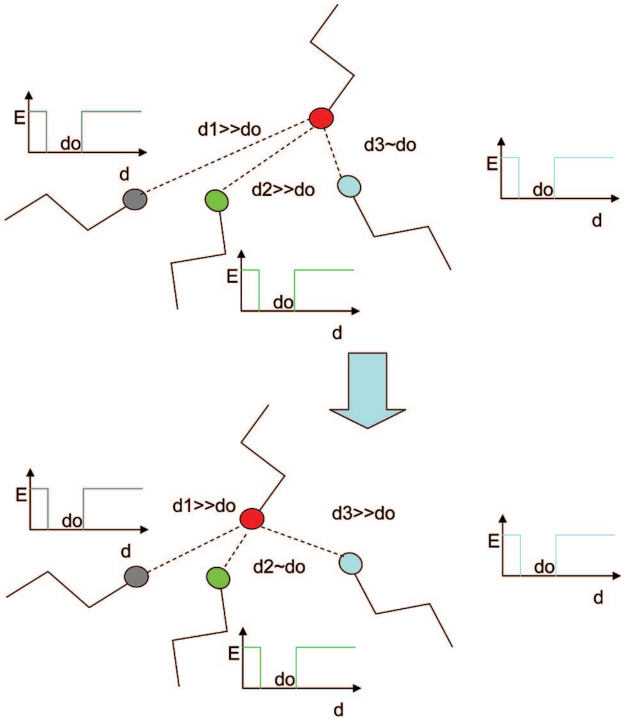

Figure 6. Schematic representation of how DMD can be used with competing interactions for side-chain (or backbone) refinement. In the example a bead in red is at an interaction distance for another bead (in blue), but the system Hamiltonian is defined by three competing finite square wells, all centered at the same equilibrium distance d_o. During the simulation the system is then allowed to exit the original well (in blue) entering into any of the two other alternating wells (in gray and green). This opens the possibility of changing from a native (red-blue) to a non-native contact (red-gray or red-green).

Force-Field Definition for DMD. The force-field was aimed at performing atomistic simulations of proteins around equilibrium structures but enabling moderate nonharmonic conformational transitions (not allowed in a pure Gō-Cα model) and introducing physical considerations (missing in current normal mode or Gaussian elastic model systems). To this purpose, we defined a simple united atom force-field consisting of "bonded" and "nonbonded" terms. As usual "bonded" terms account for stretchings (a-b), bendings (a-b-c), and torsions (a-b-c-d). Stretching, bendings, and torsions involving double or conjugated bonds were represented by means of infinite square wells with a width corresponding to 5% of the length of the bond/pseudobond distance[41] (a-b bond for stretching, a-c pseudobond for bending, and a-d pseudobond for torsions; see Figure 1). Equilibrium values for bonds and pseudobonds were taken from the reference structure (in our case the MD-averaged one). In this version of the force-field no explicit terms were used to represent torsions around single bonds (the only fixed torsions were that of the peptide bonds and sidechain rings), but forbidden regions in the Ramachandran plot were not accessible due to the steric interactions between backbone O and C_β atoms (see Results).

Up to six nonbonded terms were considered: i) a hardcore infinite repulsive term to avoid nuclei overlap, ii) a backbone hydrogen bond potential which is designed to enforce the maintenance of the secondary structure of the reference conformation, iii) a salt bridge term to preserve native ionic contacts, iv) a potential accounting for residue-residue hydrophobic interactions, v) a weak C_α-C_α based dispersion term, and finally vi) a pseudosolvation term for exposed residues to avoid the tendency of exposed polar side chains to collapse onto the surface of the protein.

Hardcore repulsions between all nonbonded particles at $d_{min}=R_{hc}^{i} + R_{hc}^{j}$ were established (see Table S1 in the Supporting Information for hardcore radii of each atom type) using an infinite repulsive wall (see Figure 1). Hydrogen bonds were defined with pseudobonds between backbone O_i and N_j that were located at a distance $R_c < 4$ Å in the reference conformation as long as i) the N_jO_i and N_jC_i axes are almost parallel (maximum angle allowed of 45°) and ii) the $C_{j-1}O_i$, $C_{\alpha j}O_i$, and N_jO_i axes are almost coplanar (see schematic diagram in Figure 1). To keep the geometry of the hydrogen bonds, pseudobonds between C_i and N_j, O_i and C_{j-1}, and O_i and $C_{\alpha j}$ were added. These pseudobonds are defined as finite potential wells centered at the interatomic distance d_{ab} in the native conformation, whose width is 5% of d_{ab}. Salt bridges were defined between suitable atom types (see Table S2 in the Supporting Information) that were located at a distance lower than $R_c = 6$ Å in the native conformation. Hydrophobic interactions were defined between the central atoms of hydrophobic side chains (see Table S3 in the Supporting Information) that were located at $R_c \leq 6$ Å. Salt bridges and hydrophobic interactions were defined as wells between d_{min} and R_c. Hydrogen bonds and hydrophobic and saline interactions are defined by deep (15 kcal/mol) finite wells in order to preserve the contact topology of the native conformation. Therefore the method is not well suited for *ab initio* folding prediction. The pseudosolvation term is based on a volume exclusion term for exposed polar residues as a hardcore potential of 4 Å for the atoms specified in Table S4 in the Supporting Information. Finally, dispersion terms at the residue level were introduced by using a C_α-based finite square well. The potential associated to this well is defined by an infinite wall at the hardcore repulsion distance $R_{hc}^{i} + R_{hc}^{j}$ and a step of depth $V(r_0)$ at the interparticle distance r_0 plus 1 Å in the native conformation. The distance-dependent $V(r_0)$ term is computed as

$$V(r_0) = -\varepsilon\left(\frac{d}{r_0}\right)^6 \tag{11}$$

where $d = R_{hc}^{i} + R_{hc}^{j}$, and the factor hardness constant ε was fitted to 2 after training with a selected set of ultrarepresentative proteins in our database (see below).

Overall, the force-field tries to maintain the structure close to the reference conformation but enables large movements if they do not disrupt favorable physical interactions. The force-field is then expected to keep sampling within reasonable limits around the reference structure but providing a more realistic exploration of conformational space.

Molecular Dynamics Simulations. MD trajectories were taken when available from our μMODEL database (http://mmb.pcb.ub.es/MODEL) or computed directly for this work. In all cases protein structures were titrated, neutralized by ions, minimized, hydrated, heated, and equilibrated (for at least 0.5 ns) using an established protocol,[17] followed by at least 10 ns of production. Trajectories in the micromodel database (see below) were collected using three all-atom force fields (AMBER,[51] CHARMM,[52,53] OPLS/AA[54−57]). The μMODEL proteins considered in this work are 1AGI, 1BFG, 1BJ7, 1BSN, 1CHN, 1CQY, 1CSP, 1CZT, 1EMR, 1FAS, 1FVQ, 1I6F, 1IL6, 1J5D, 1JLI, 1K40, 1KTE, 1KXA, 1LIT, 1LKI, 1OOI, 1OPC, 1PDO, 1PHT, 1SDF. Additional proteins which are not present in the μmodel databases and whose MD simulations are presented here for the first time were obtained using only the AMBER[51] force-field: 1ARK, 1BPI, 1CEI, 1SRO, 1UBQ, 2GB1, 3CI2, 4ICB.

The Particle Mesh Ewald approach in conjunction with periodic boundary conditions was used to address long-range nonbonded interactions.[59] Integration of the equations of motion proceeded with a time step of 1 fs; vibrations of bonds involving hydrogen atoms were removed by the SHAKE/RATTLE algorithm.[60,61] Jorgensen's TIP3P model[62,63] was used to represent aqueous solvent. Calculations were performed with AMBER8[64] and NAMD2.6[65,66] computer programs. Results presented in the manuscript are always referred to as the AMBER force-field results (those for which more proteins trajectories are available), but comparison between trajectories obtained with the three different atomistic force-fields are used as reference.

Training and Testing Databases. A reduced training set of proteins (1I6F, 1SDF, 1JLI) simulated with very long (0.1 μs) trajectories was used to fit the adjustable parameters in our model. Testing was then performed with proteins representing all metafolds as described in the micromodel database,[17] and several extra proteins were selected as examples of mobile structures in the protein databank (see above). Proteins in the database are structurally and functionally diverse and represent a quite complete sampling of single domain proteins, which are the most challenging for representation with simplified models.

Metrics for Trajectory Comparison. The samplings obtained from MD and DMD simulations were compared by using a variety of metrics, which are briefly discussed (the reader is addressed to suitable references[18] for more details). Most of the comparison relies on a preprocessing of the trajectories by the essential dynamics procedure,[67] where the covariance matrix built from the ensemble of configuration is treated by principal component analysis (PCA) to derive a set of eigenvectors ($\upsilon_{\rm i}$) determining the nature of the essential deformation movements and a set of eigenvalues ($\lambda_{\rm i}$) defining the amount of variance explained by each deformation movement. Note that eigenvalues can be transformed into stiffness constants using the harmonic expression

$$k_l = \frac{k_b T}{\lambda_l} \tag{12}$$

where T is the absolute temperature, and k_b is the Boltzman's constant.

Size of the accessible configuration space was analyzed by inspection of the corresponding variance and the entropy computed by diagonalization of the mass-weighted covariance matrix using Schlitter's method[68]

$$S = \frac{k_B}{2}\ln\left(1 + \frac{e^2}{\alpha^2}\right) \tag{13}$$

where $\alpha_i = \hbar\omega_i/k_bT$, with ω being the eigenvalues (in frequency units) obtained by diagonalization of the mass-weighted covariance matrix, and the sum extends to all the nontrivial vibrations of the system.

The reduced variance (Var$_{\rm red}$) measures the size of the deformation space when only the most prevalent deformation modes are considered (here only the first 3 eigenvectors were considered)

$$Var_{red} = \sum_{i=1}^{3} \lambda_i \tag{14}$$

where i stands for the rank order of the eigenvalues.

The complexity of the deformability space is measured as the minimum number of eigenvectors needed to explain 90% of variance. Note that this parameter does not necessarily correlate with the size of the conformational space, since samplings with large variance might be explained by a small number of very soft modes, indicating a wide but simple deformation space.

The overlap of deformation space indicates the similarity between the essential spaces of two trajectories and was computed by analyzing the overlap of the respective important spaces,[69−71] defined as those necessary to complete 90% of the total variance in the trajectories generated with MD

$$\gamma_{XY} = \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{m}(\nu_i^X \cdot \nu_j^Y)^2 \tag{15}$$

where X and Y index the two methods to be compared, i and j index the eigenvectors (ranked on the basis of their contribution to structural variance), and m is the number of eigenvectors in the "important space". For a finite time trajectory structural variation generates noise in the similarity computation that needs to be corrected by including the self-similarity terms, deriving then a relative similarity index[69−71]

$$\Gamma = \frac{2\sum_{i=1}^{m}\sum_{j=1}^{m}(\nu_i^X \cdot \nu_j^Y)^2}{\sum_{i=1}^{m}\sum_{j=1}^{m}(\nu_i^{Y\prime} \cdot \nu_j^{Y\prime})^2 + \sum_{i=1}^{m}\sum_{j=1}^{m}(\nu_i^{X\prime} \cdot \nu_j^{X\prime})^2} \tag{16}$$

where the self-similarity products $(\nu_i^{Y'} \cdot \nu_j^{Y'})$ were obtained by comparing first and second halves of trajectories.

Note that similarity indexes are dependent on the number of eigenvalues considered converging to 1 for the entire eigenvector/eigenvalue (complete basis set). Statistical significance of a given similarity index should be then quantified by Z-score

$$Z_{score} = \frac{(\gamma_{XY}(observed)) - (\gamma_{XY}(random))}{std(\gamma_{XY}(random))} \tag{17}$$

where random models were obtained by diagonalization of a pseudocovariance matrix obtained from a simple coarse grained DMD simulation, where only chemical bond and hard sphere potentials were considered. As described elsewhere,[30] this pseudorandom model is more restrictive (yielding to lower but more realistic Z-scores) than a pure background model obtained by assuming gas phase behavior in the proteins. The standard deviation (*std* in the equation) was obtained by considering 500 different pseudotrajectories.

Secondary structure was determined for DMD_{AA} and MD trajectories using standard H-bond annotation criteria. Additionally, Ramachandran's maps were created to investigate the population of forbidden regions along DMD simulations. In all cases calculations were performed using 5000 equally-spaced snapshots collected from the last 5 ns portion of the trajectories.

Local residue mobility was analyzed by means of the isotropic B-factors

$$\text{B-factor} = \frac{8}{3}\pi^2\langle\Delta r^2\rangle \tag{18}$$

where $\langle\Delta r^2\rangle$ stands for the oscillations of residues around equilibrium positions. B-factors distributions were compared by means of the Spearman correlation coefficient to obtain a dimensionless measure of the distribution of residue fluctuations.

Regional mobility was analyzed by means of Lindemann's index[17,72] determined as

$$\Delta_L = \frac{\left(\sum_i \langle\Delta r_i^2\rangle/N\right)^{1/2}}{a'} \tag{19}$$

where a' is the most probable nonbonded near-neighbor distance, N is the number of atoms, and $\langle\Delta r^2\rangle$ is the mean-square displacement of an atom from its equilibrium position. Lindemann's index provides a well accepted description of the macroscopic behavior of a molecular system or a part of it: if Δ_L is lower than 1.5, it is considered a solid; if higher, it is considered a molten solid; and for ever higher values, it is considered a liquid.[73]

Results and Discussion

We generated 10 ns trajectories for the proteins mentioned above and compared them with the trajectories obtained from state-of-the-art all-atom explicit solvent molecular dynamics simulations. It is worth pointing out that within 10 ns DMD allows sampling a bigger conformational space than within 10 ns of MD, because the implicit treatment of the solvent and hydrogen atoms increases the conformational freedom of the protein and reduces the complexity of the energy landscape.

Despite the lack of a complex network of residue-residue harmonic (or infinite) restraints, as happened in Gō-based models, the structures sampled in current DMD simulations remain quite close to the reference conformation (see Table 1). Thermal noise explains around 1-1.5 Å of the deviation between DMD and the MD-averaged conformation, the rest being due mostly to flexible loops as noted in the TM-scores shown in Table 1.

We were concerned that the use of physical interactions and the removal of many residue-residue restriction potentials from our current force-field might expand and perhaps distort the protein in those regions with less physical contacts. Fortunately, such an effect is negligible as seen in measures like the solvent accessible surface and radii of gyration, which are not larger than those expected from a normal MD simulation.[11] Another potential concern was the structural stability of secondary elements, since no specific torsional terms are applied to Φ and Ψ bonds. However, the analysis demonstrates that the secondary structure is well preserved (from both MD and experimental values; see Table 1), with just a moderate decrease expected from thermal vibration. Furthermore, Ramachandran's maps obtained from DMD simulations are close to those that can be derived from the same proteins from atomistic MD simulations (see Figure 2), indicating that besides its simplicity DMD is not sampling artefactual local structures. Finally, the analysis shows that DMD simulations maintain very well the pattern of native residue-residue contacts (Table 1), even in cases where there are not obvious physical interactions that can be traced between the contact partners.

At this point we should remark that we have used averaged MD structures (obtained by averaging and partial minimization of atomistic MD ensembles) as reference structures to define the DMD force-field, and, accordingly, structural analysis should always be performed with respect to these structures. However, for the sake of completeness, we extend the analysis to consider experimental structures (X-ray or NMR) as reference. Results in Table 1 demonstrate DMD ensembles are always close to experimental structure, remarking the suitability of the technique to represent real structural properties of proteins.

The force-field definition used in this DMD implementation pursues not only to maintain the samplings close to the reference structure but also to reduce the full harmonicity intrinsic to Gō-like methods. Inspection of RMSd fluctuation profiles (see randomly selected examples in Figure 3) demonstrates that the pseudophysical potential is able to maintain the samplings close to the reference conformations but at the same time allow local transitions, temporal oscillations in the trajectories, and in summary a more realistic deformation pattern than the pure harmonic behavior observed in sampling generated by NMA-predicted essential movements, Brownian MD based on C_α-C_α harmonic restraints, or our Gō-like C_α implementation of DMD (see Figure 3).

The total size of the deformation space sampled by DMD and MD simulations is quite similar, as noted in the total variances and molecular entropies (both given as values per residue) shown in Table 2. Furthermore, the distribution of variance along modes in DMD simulations is quite realistic as shown in the variance vs eigenvector profiles shown in Figure 4, the complexity and the reduced variance metrics displayed in Table 2, giving a clear improvement with respect to Gō-like C_α methods.[24] In summary, present heavy atom DMD simulations reproduce well the extension of the deformation pattern determined by atomistic MD simulation and at the same time balance properly the importance of the different deformation modes. It seems then that the Hamiltonian definition used here is able to reduce some of the artefacts arising from the use of residue-residue harmonic (or Gō-like) potentials but keeping at the same time structures close to those used as reference.

We also compared DMD and MD essential deformation spaces to determine to what extent both methodologies detect the same type of essential deformations. Results in Table 2 demonstrate that there is a good overlap between MD and DMD trajectories, since Γ values between DMD samplings and AMBER MD trajectories are not far from those obtained when the individual MD trajectories (CHARMM, AMBER, and OPLS) are compared. The statistical significance of the computed similarity becomes evident when looking at the associated Z-scores which are in the order of 10^2 (within the range of Z-scores obtained when atomistic MD trajectories with different force-fields are compared), thus ruling out the possibility of a fortuitous similarity. In other words, the DMD algorithm reported here is able to capture not only the global pattern of flexibility of proteins but also the intrinsic nature of the deformation modes.

In order to determine whether or not DMD is able to reproduce also well the residue flexibility we computed residue B-factors from the DMD ensembles comparing the values with those obtained by MD simulations and when available X-ray data. Results summarized in Table 2 (for randomly selected examples see Figure 5) demonstrate that MD and DMD values correlate well with Spearman's coefficients within the range of those obtained when the MD trajectories with different force-fields are compared (Table 2). Moreover, DMD computed B-factors correlate also well with available X-ray values (see Table 2); even such experimental data were never considered to refine the method. Finally, the ability of the DMD to distribute properly flexibility among different protein regions is also clear by inspecting Lindemann's indexes (see Table 2), which demonstrate that the balance between solid (interior of protein)/ liquid (exterior of the protein) which is found in atomistic MD simulation is well reproduced by our DMD calculations.

Conclusions

All the preceding analysis, performed on a very large set of representative proteins and using state of the art methods as reference, demonstrates that the pseudophysical DMD method can reasonably reproduce the flexibility of proteins as determined by atomistic MD simulations in explicit solvent, avoiding the need to integrate the equations of motions every femtosecond. Without important modifications the method can be used to refine portions of the protein just adding competing wells that will allow a given residue to change partners to optimize the overall energy (see Figure 6), and just adding a long range attractive potential the technique can be used to study protein/ligand diffusion and protein-protein interactions in the context of flexible macromolecules. The performance of the method in these scenarios will be the subject of future investigation.

Acknowledgment. We thank Dr. M. Rueda for help in the analysis of data, the MODEL team for help in managing the MD simulations, and Prof. F. J. Luque for many helpful discussions. This work was supported by the Spanish Ministry of Education and Science (BIO2006-01602), the Instituto de Salud Carlos III (COMBIOMED-project), the Fundación Marcelino Botín, and the National Institute of Bioinformatics (Structural Bioinformatics Node).

Supporting Information Available: Hardcore radii of each atom type, atom types defining salt bridges and hydrophobic interactions, atom types used to simulate solvation effects in our model, and a figure comparing the sampling obtained with DMD to that obtained with MD for the protein 1FVQ. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Ma, J.; Karplus, M. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 8502–8507.

(2) Daniel, R. M.; Dumm, R. V.; Finney, J. L.; Smith, C. J. *Annu. Rev. Biophys. Biomol. Struct.* **2003**, *32*, 69–92.

(3) Eisenmesser, E. Z.; Bosco, D. A.; Akke, M.; Kern, D. *Science* **2002**, *295*, 1520–1523.

(4) Luo, J.; Bruice, T. C. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 13152–13156.

(5) Hinsen, K.; Thomas, A.; Field, M. J. *Proteins* **1999**, *34*, 369–382.

(6) Waldron, T. T.; Murphy, K. P. *Biochemistry* **2003**, *42*, 5058–5064.

(7) Yang, L.-W.; Bahar, I. *Structure* **2005**, *13*, 893–904.

(8) Sacquin-Mora, S.; Lavery, R. *Biophys. J.* **2006**, *90*, 2706–2717.

(9) Remy, I.; Wilson, I. A.; Michnick, S. W. *Science* **1999**, *283*, 990–993.

(10) Henzler-Wildman, K. A.; Ming, L.; Vu, T.; Kerns, S. J.; Karplus, M.; Kern, D. *Nature* **2007**, *450*, 913–916.

(11) Teague, S. J. *Nat. Rev. Drug Discovery* **2008**, *2*, 527–541.

(12) Marvin, J. S.; Hellinga, H. W. *Nature Struct. Biol.* **2001**, *8*, 795–798.

(13) Falke, J. J. *Science* **2002**, *295*, 1480–1481.

(14) Kenakin, T. *Trends Pharmacol. Sci.* **1995**, *16*, 188–192.

(15) Ma, B.; Shatsky, M.; Wolfson, H. J.; Nussinov, R. *Protein Sci.* **2002**, *11*, 184–197.

(16) Shoichet, B. K.; Baase, W. A.; Kuroki, R.; Matthews, B. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 452–456.

(17) Rueda, M.; Ferrer-Costa, C.; Meyer, T.; Perez, A.; Camps, J.; Hospital, A.; Gelpi, J. L.; Orozco, M. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 796–801.

(18) Rueda, M.; Chacon, P.; Orozco, M. *Structure* **2007**, *15*, 565–575, 2007.

(19) Karplus, M.; McCammon, J. A. *Sci. Am.* **1986**, *254*, 42–51.

(20) McCammon, J. A.; Gelin, B. R.; Karplus, M. *Nature* **1977**, *267*, 585–590.

(21) Allen, M. P.; Tildesley, D. J. Computer Simulation of Liquids; Clarendon Press: Oxford, 1989.

(22) Brooks, C. L., III; Karplus, M.; Pettitt, B. M. *Proteins: A Theoretical Perspective of Dynamics, Structure and Thermodynamics*; Cambridge University Press: Cambridge, 1987.

(23) Warshel, A. *Nature* **1976**, *260*, 679–683.

(24) Van Gunsteren, W. F.; Karplus, M. *Biochemistry* **1982**, *21*, 2259–74.

(25) Berendsen, H. J. C.; Postma, J. P. M.; Van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684–3690.

(26) Karplus, M.; McCammon, J. A. *Nat. Struct. Biol.* **2002**, *9*, 646–652.

(27) Karplus, M.; Kuriyan, J. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6679–6685.

(28) McCammon, J. A.; Harvey, S. C. *Dynamics of Proteins and Nucleic Acids*; Cambridge University Press: Cambridge, 1987.

(29) Alder, B. J.; Wainwright, T. E. Studies in Molecular Dynamics. I. General Method. *J. Chem. Phys* **1959**, *31*, 459–466.

(30) Emperador, A.; Carrillo, O.; Rueda, M.; Orozco, M. *Biophys. J.* **2008**, *95*, 2127–2138.

(31) Ding, F.; Buldyrev, S. V.; Dokholyan, N. V. *Biophys. J* **2005**, *88*, 147–155, 2005.

(32) (a) Zhou, Y. Q.; Karplus, M. *Nature* **1999**, *401*, 400–403. (b) Dokholyan, N. V.; Buldyrev, S. V.; Stanley, H. E.; Shakhnovich, E. I. *Folding Des.* **1998**, *3*, 577–587.

(33) Linhananta, A.; Zhou, Y. *J. Chem. Phys.* **2002**, *117*, 8983–8995.

(34) Luo, Z.; Ding, J.; Zhou, Y. *Biophys. J.* **2007**, *93*, 2152–2161.

(35) Zhou, Y.; Linhananta, A. *Proteins: Struct., Funct., Genet.* **2002**, *47*, 154–162.

(36) Ding, F.; Sharma, S.; Chalasani, P.; Demidov, V. V.; Broude, N. E.; Dokholyan, N. V. *RNA* **2008**, *14*, 1164–73.

(37) Sharma, S. F.; Ding, F.; Dokholyan, N. V. *Biophys. J.* **2007**, *92*, 1457–1470.

(38) Ding, F.; Dokholyan, N. V.; Buldyrev, S. V.; Stanly, E. H.; Shakhnovich, E. I. *J. Mol. Biol.* **2002**, *324*, 851–857.

(39) Chen, Y.; Dokholyan, N. V. *J. Mol. Biol.* **2005**, *354*, 473–482.

(40) Ding, F.; LaRocque, J. J.; Dokholyan, N. V. *J. Biol. Chem.* **2005**, *280*, 40235–40240.

(41) Khare, S. D.; Ding, F.; Gwanmesia, K. M.; Dokholyan, N. V. *PLoS Comput. Biol.* **2005**, *1*, 230–235.

(42) Marchut, A. J.; Hall, C. K. *Biophys. J.* **2006**, *90*, 4574–4584.

(43) Nguyen, H. D.; Hall, C. K. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 16180–16185.

(44) Peng, S.; Ding, F.; Urbanc, B.; Buldyrev, S. V.; Cruz, L.; Stanley, H. E.; Dokholyan, N. V. *Phys. Rev. E* **2004**, *69*, 041908.

(45) Ding, F.; Borreguero, J. M.; Buldyrey, S. V.; Stanley, H. E.; Dokholyan, N. V. *Proteins: Struct., Funct., Genet.* **2003**, *53*, 220–228.

(46) Ding, F.; Tsao, D.; Nie, H.; Dokholyan, N. V. *Structure* **2008**, *16*, 1010–1018.

(47) Urbanc, B.; Cruz, L.; Buldyrev, S. V.; Bitan, G.; Teplow, D. B.; Stanley, H. E. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 17345–17350.

(48) Yun, S.; Urbanc, B.; Cruz, L.; Bitan, G.; Teplow, D. B.; Stanley, H. E. *Biophys. J.* **2007**, *92*, 4064–4077.

(49) Urbanc, B.; Cruz, L.; Buldyrev, S.V.; Bitan, G.; Teplow, D. B.; Stanley, H. E. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6015–6020.

(50) Taketomi, H.; Ueda, Y.; Gô, N. *Int. J. Pept. Protein Res.* **1975**, *7*, 45–459.

(51) Cornell, W.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc* **1995**, *117*, 5179–5197.

(52) MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.

(53) MacKerell, A. D., Jr.; Karplus, M. *J. Am. Chem. Soc.* **1995**, *117*, 11946–11975.

(54) Damm, W.; Frontera, A.; Tirado-Rives, J.; Jorgensen, W. L. *J. Comput. Chem.* **1997**, *18*, 1955–1970.

(55) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225–11236.

(56) Kaminski, G.; Duffy, E. M.; Matsui, T.; Jorgensen, W. L. *J. Phys. Chem.* **1994**, *98*, 13077–13082.

(57) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. *J. Phys. Chem. B* **2001**, *105*, 6474–6487.

(58) Smith, W. S.; Hall, C. K.; Freeman, B. D. *J. Comput. Phys.* **1997**, *134*, 16–30.

(59) Darden, T. L.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089–10092.

(60) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1997**, *23*, 327–341.

(61) Andersen, H. C. *J. Comput. Phys.* **1983**, *52*, 24–34.

(62) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.

(63) Mahoney, M. W.; Jorgensen, W. L. *J. Chem. Phys.* **2000**, *112*, 8910–8922.

(64) Case, D. A., Pearlman, D. A.; Caldwell, J. W.; Cheatham, T. E.; Ross, W. S.; Simmerling, C. L.; Darden, T. L.; Merz, K. M.; Stanton, R. V.; Cheng, A. L.; Vincent, J. J.; Crowley, M.; Tsui, V.; Radmer, R. J.; Duan, Y.; Pitera, J.; Massova, I.; Seibel, G. L.; Singh, U. C.; Weiner, P. K.; Kollman, P. A. University of California, San Francisco, 2004.

(65) Kale, L.; Skeel, R.; Bhandarkar, M.; Brunner, R.; Gursoy, A.; Krawetz, N.; Phillips, J.; Shinozaki, A.; Varadarajan, K.; Schulten, K. *J. Comput. Phys.* **1999**, *151*, 283–312.

(66) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K. *J. Comput. Chem.* **2005**, *26*, 1781–1802.

(67) Amadei, A.; Linssen, A. B.; Berendsen, H. J. *Proteins* **1993**, *17*, 412–425.

(68) Schlitter, J. *Chem. Phys. Lett.* **1993**, *215*, 617–621.

(69) Hess, B. *Phys. Rev. E* **2000**, *62*, 8438–8448.

(70) Noy, A.; Meyer, T.; Rueda, M.; Ferrer, C.; Valencia, A.; Perez, A.; Orozco, M.; de la Cruz, X.; Luque, F. J. *J. Biomol. Struct. Dyn.* **2006**, *23*, 357–484.

(71) Orozco, M.; Perez, A.; Noy, A.; Luque, F. J. *Chem. Soc. Rev.* **2003**, *32*, 350–364.

(72) Zhou, Y.; Vitkup, D.; Karplus, M. *J. Mol. Biol.* **1999**, *285*, 1371–1375.

(73) Zhou, Y.; Karplus, M. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 14429–14432.

CT8003832