Fast Atomistic Molecular Dynamics Simulations from Essential Dynamics Samplings

Oliver Carrillo,[†] Charles A. Laughton,*[,‡] and Modesto Orozco*[,†,§,∥]

[†]Joint IRB-BSC Program on Computational Biology, Barcelona Supercomputing Center and Institute of Research in Biomedicine, Barcelona, Spain

[‡]School of Pharmacy and Centre for Biomolecular Sciences, University of Nottingham, Nottingham, United Kingdom

[§]National Institute of Bioinformatics, Parc Científic de Barcelona, Barcelona, Spain

[∥]Departament de Bioquímica, Facultat de Biología, University of Barcelona, Barcelona, Spain

ABSTRACT: We present a new method for fast molecular dynamics simulations in cases where the new trajectories can be considered a perturbation or a combination of previously stored ones. The method is designed for the postgenomic scenario, where databases such as MoDEL will store curated equilibrium trajectories of all biomolecules (proteins, nucleic acids, etc.) of human interest. We demonstrate that the approach outlined here can, with accuracy and great computational efficiency, reproduce and extend original trajectories, describe dynamical effects due to perturbations (e.g., protein−ligand and protein−protein interactions and protein mutations) and predict the dynamics of large polymeric systems built up from previously studied fragments. The method can work simultaneously with low- and high-resolution pictures of the macromolecule, allowing the level of detail to be matched to that required for obtaining the information of biological interest.

■ INTRODUCTION

Biological macromolecules (proteins and nucleic acids) are flexible entities whose functionality is nearly always related to their ability to adjust their shape.[1] The experimental determination of molecular flexibility (particularly with atomic resolution) is difficult; this has encouraged the development and use of computational modeling approaches, among which molecular dynamics (MD) is probably the most rigorous and accurate one.[2] MD allows the determination of the conformational ensemble of a system by numerical integration of the Newton equations of motion of the constituting atoms over extended periods of time. The required forces are computed from physical potentials (the force field) represented by simple equations, which are fitted to reproduce quantum mechanical calculations and condensed phase experimental values.[3] Current MD simulation protocols are able to predict the dynamics of biological macromolecules in full atomistic detail in a nearly physiological environment over multi-nanosecond time scales (approaching a microsecond for small systems[4] or even longer if specific-purpose hardware is used[5]). However, despite its power and universality, practical use of MD is still limited by its computational cost and by the technical expertise required in both: the setup of the system and the analysis of the collected data. Thus, the current *state of the art* in MD makes it a feasible approach to, for example, study the behavior of a medium sized protein bound to an interesting ligand over a multinanosecond time scale, but not to study 10^4−10^5 potential complexes, as typically required in drug design, or hundreds of protein variants, as may well be needed in pharmacogenomic studies. Similarly, it is possible to study a 12-mer DNA in the microsecond range,[6] but it is not possible to study a section of chromatin (kilobases size).

Initiatives, such as MoDEL[7] or Dynameomics,[8] currently provide the community with databases of atomistic trajectories for more than a thousand representative proteins in near-physiological conditions, and we can expect that as computer power increases, these (or similar) databases will eventually contain trajectories for the major variants of all nonredundant proteins of known structure. In a similar spirit, the international Ascona B-DNA Consortium[9,10] has collected accurate trajectories for all possible unique 4-mer fragments in duplex DNA, going in the direction to describe an atlas of DNA structure and flexibility. The question is then how this massive amount of information can be used to predict the dynamics of related proteins, protein complexes, or large DNA polymers of different sequences without the need to perform thousands of additional simulations.

In this contribution, we present a new method (named Essential Dynamics-Molecular Dynamics; ED/MD) for deriving trajectories of macromolecules using pre-existing dynamic data on related systems. Simulations are done using a hybrid Hamiltonian, partly projected in the essential dynamics space and partly in normal Cartesian space, allowing a very fast calculation of the dynamics of macromolecules and complexes. In contrast to normal MD, the user has full control of the desired level of resolution of the trajectory moving from fully atomistic to a coarse grained level which can in fact coexist during the calculation without a loss of accuracy in the atomistic part. We provide examples of how the method can provide useful results in a variety of systems and demonstrate the excellent computational efficiency of the method.

■ METHODS

The Basic ED/MD Formalism. As noted above, the basic idea of our method is that the trajectory of a system can be inferred from previously computed trajectories of closely related systems. For example, the dynamics of a particular

Received: October 17, 2011
Published: January 27, 2012

protein–ligand complex can be inferred from the dynamics of the complex of the same protein with another ligand. The starting point is always an atomistic MD simulation from which an ensemble of conformations of the macromolecule $\langle r \rangle$ is obtained. This in turn defines a covariance matrix, whose diagonalization yields a set of $(3N - 6)$ eigenvectors $(\{\hat{e}_i\})$, describing the nature of the essential deformation movements, and the associated set of eigenvalues $(\{\lambda_i\})$ determining the magnitude of the deformation sampled in each deformation movement. As demonstrated elsewhere,[11,12] a (Cartesian coordinate) trajectory can be projected into the entire set of eigenvectors with no loss of accuracy. Furthermore, an approximated trajectory can be obtained by a back transformation that only considers projections along the most important eigenvectors (those explaining the largest amount of variance in the trajectory either for the entire protein or fragments of it[12]). Within the harmonic limit, the movements of the macromolecule along the essential movements can be represented by an effective Hamiltonian:

$$H^{(M)} = \sum_{i=1}^{3N} \frac{p_i^2}{2m_i} + \frac{1}{2} \sum_{i=1}^{M} \frac{k_B T}{\lambda_i} (\Delta \vec{r} \cdot \hat{e}_i)^2 \tag{1}$$

where i stands for a particle, p_i is the ith component of the momentum vector, m_i the particle mass, $\Delta \vec{r} = \vec{r} - \vec{r}_0$ is the displacement from the equilibrium position of atoms $(\vec{r}_0)$, k_B is Boltzmann's constant, and T is the absolute temperature. It is worth noting that this Hamiltonian can be evaluated to a given level of accuracy by simply restricting the sum to the M most significant deformation modes (instead of considering all $3N - 6$ modes). The restriction of the eigenvector space $(M < 3N - 6)$ allows us to concentrate the computational effort on relevant modes and to use large integration time steps. Note also that it is trivial to restrict the Hamiltonian to a given set of atoms or to reproduce a part of the molecule at the coarse-grained level (for example at the C_α level), keeping all of the atomistic detail in another region. In both cases, the part of interest is reproduced with the same detail as in the original unrestricted-fully atomistic trajectory.

For a system where interactions other than those from the original MD simulation exist, an effective Hamiltonian can be defined by adding a perturbation term (V), typically expressed in Cartesian terms:

$$H = H^{(M)} + V \tag{2}$$

The corresponding perturbational $3N$ force $\vec{F} = -\vec{\nabla} V$ acting on the protein $(\vec{\nabla} = (\partial/\partial r_1, ..., \partial/\partial r_{3N}))$ is reduced to avoid divergences due to higher mode contributions, leading to a reduced force:

$$\vec{F}^* = \vec{F} - \sum_{k=M+1}^{3N} \phi_k \hat{e}_k \tag{3}$$

where $\{\varphi_1, ..., \varphi_M, \varphi_{M+1}, ..., \varphi_{3N}\}$ are the components of $\vec{F}$ in the base $\hat{e}$.

The resulting force acting on the macromolecule under the influence of a perturbational potential is then computed as

$$\vec{F}_i^{(M)} = -\sum_{k=1}^{M} \frac{k_B T e_k^i}{\lambda_k} (\Delta \vec{r} \cdot \hat{e}_i) + \vec{F}_i^* \tag{4}$$

where the sum on the right-hand side is the resulting force derived from the Hamiltonian $H^{(M)}$, where e_k^i is the component i of the eigenvector corresponding to mode k.

Integration of Equations of Motion. The effective forces computed using eq 4 define a Langevin equation[13−15] which needs to be integrated to obtain new trajectories:

$$m_i \ddot{\vec{r}}_i = \vec{F}_i^{(M)} - \gamma \dot{\vec{r}}_i + \vec{\xi}_i^* \tag{5}$$

where γ stands for a friction coefficient and stochastic terms should satisfy

$$\langle \vec{\xi}_i^*(t) \rangle = 0$$
$$\langle \vec{\xi}_i^*(t)\, \vec{\xi}_j^*(t') \rangle = 2 k_B T^* \gamma \delta_{ij} \delta(t - t') \tag{6}$$

and the effective temperature (T^*) is defined as

$$T^* = \frac{3N}{M} T \tag{7}$$

due to the fact that the M-modes system must have the same equilibrium energy as the $3N$-modes one.

Numerical integration of eq 5 is done using Verlet algorithm:[15]

$$\vec{r}_i = \vec{r}_i^{\,0} + \tau(1 - e^{-\Delta t/\tau}) \vec{v}_i^{\,0} + \frac{\Delta t}{\tau}\left(1 - \frac{\tau}{\Delta t}(1 - e^{-\Delta t/\tau})\right) \vec{F}_i^{\,0} + \Delta \vec{r}_i^{\,G} \tag{8}$$

and

$$\vec{v}_i = e^{-\Delta t/\tau} \vec{v}_i^{\,0} + \frac{1}{\gamma}(1 - e^{-\Delta t/\tau}) \vec{F}_i^{\,0} + \Delta \vec{v}_i^{\,G} \tag{9}$$

with stochastic integrals:

$$\Delta \vec{r}_i^{\,G} = \frac{1}{\gamma m_i} \int_t^{t+\Delta t} [1 - e^{-\gamma(t+\Delta t - t')}] \vec{\xi}_i^*(t')\, dt' \tag{10}$$

$$\Delta \vec{v}_i^{\,G} = \frac{1}{m_i} [e^{-\gamma(t+\Delta t - t')}\, \vec{\xi}_i^*(t')]\, dt' \tag{11}$$

In order to avoid excitation of the negligible modes, the stochastic term $\vec{\xi}^*$ is truncated summing up to the Mth mode of the exact noise vector term:

$$\vec{\xi}^* = \vec{\xi} - \sum_{k=M+1}^{3N} \rho_k \hat{e}_k \tag{12}$$

where $\{\rho_1, ..., \rho_{M+1}, ..., \rho_{3N}\}$ are the components of the noise vector $\vec{\xi}$ in the eigenvector base $\hat{e}$.

Trajectories are obtained by activating movements along a set of essential deformations, representing movements with different associated frequencies. The optimum integration time step (Δt) can be determined for each mode based on its characteristic frequency:

$$\Delta t_i < 2\pi \left(\frac{m_s \lambda_i}{k_B T} \right)^{1/2} \tag{13}$$

where Δt_i is the time step associated with the integration of movements along mode i, λ_i is its eigenvalue, T is the nonreduced temperature, and m_s is the mass of the smallest particle (to accelerate calculations and avoid divergences, hydrogen trajectories are not integrated, and accordingly m_s corresponds to the carbon mass).

The use of multiple time steps, which accelerates calculation, can be made evident by rewriting eq 5 explicitly:

$$m_i\ddot{\vec{r}}_i = -\sum_{k=1}^{M}\frac{k_BTe_k^i}{\lambda_k}(\Delta\vec{r}\cdot\hat{e}_k) + \vec{F}_i^* - \gamma\dot{\vec{r}}_i + \vec{\xi}_i^* \tag{14}$$

where the first term in the right-hand part of eq 14 accounts for the internal force of the protein, and the second term is the external force. If we divide both sides by m_i and develop the dot product of the sum, we get

$$\ddot{\vec{r}}_i = -k_BT\sum_{j=1}^{3N}\left(\sum_{k=1}^{M}\frac{e_k^je_k^j}{m_i\lambda_k}\right)\Delta\vec{r}_j + \frac{\vec{F}_i^*}{m_i} - \frac{\gamma}{m_i}\dot{\vec{r}}_i + \frac{\vec{\xi}_i^*}{m_i} \tag{15}$$

which allows us to define a set of frequencies as

$$\omega_{ij}^2 = \sum_{k=1}^{M}\frac{e_k^je_k^j}{m_i\lambda_k} \tag{16}$$

and rewrite eq 15 as

$$\ddot{\vec{r}}_i = -k_BT\sum_{j=1}^{3N}\omega_{ij}^2\Delta\vec{r}_j + \frac{\vec{F}_i^*}{m_i} - \frac{\gamma}{m_i}\dot{\vec{r}}_i + \frac{\vec{\xi}_i^*}{m_i} \tag{17}$$

By defining a threshold frequency ω_0, modes can be divided into fast ($\omega_{jk} \geq \omega_0$) and slow ($\omega_{jk} \leq \omega_0$), so that the sum in eq 17 can be rewritten as

$$\sum_{j=1}^{3N}\omega_{ij}^2\Delta\vec{r}_j = \sum_{\omega\geq\omega_0}\omega_{ij}^2\Delta\vec{r}_j + \sum_{\omega\leq\omega_0}\omega_{ij}^2\Delta\vec{r}_j \tag{18}$$

Integrations along fast motions need to be done frequently, but slow modes can be integrated less often, adding their contribution to the displacement along fast modes as a perturbation ($\vec{\delta}_i(t)$), which is re-evaluated every n steps (i.e., when we consider that slow forces need to be re-evaluated):

$$\vec{r}_i(t) = \vec{r}_i^{(0)}(t) + \vec{\delta}_i(t) \tag{19}$$

where $\vec{r}_i^{(0)}(t)$ is the fast-mode solution, which has to be evaluated at each time step Δt. Substituting eq 19 into eq 17, we get an acceleration equation concerned with coordinate i at time t:

$$\ddot{\vec{r}}_i(t) = \ddot{\vec{r}}_i^{(0)} + \ddot{\vec{\delta}}_i \tag{20}$$

from which we can distinguish the contribution of the fast and slow motion terms:

$$\begin{aligned}\ddot{\vec{r}}_i^{(0)} &= -k_BT\sum_{\text{fast}}\omega_{ij}^2\Delta\vec{r}_j^{\,0} + \frac{\vec{F}_i^*}{m_i} - \frac{\gamma}{m_i}\dot{\vec{r}}_i^{(0)} + \frac{\vec{\xi}_i^*}{m_i}\\ \ddot{\vec{\delta}}_i &= -k_BT\sum_{\text{slow}}\omega_{ij}^2\Delta\vec{r}_j^{\,0} - k_BT\sum_{j=1}^{3N}\omega_{ij}^2\vec{\delta}_j - \gamma\dot{\vec{\delta}}_i\end{aligned} \tag{21}$$

Note that the selection of two frequency intervals to distinguish between fast and slow motions is arbitrary, and the method can be generalized to several different intervals for which integration is done at different time steps.[16,17]

The basic equations outlined above can be easily adapted to study with efficiency and accuracy the dynamics of a variety of systems, such as protein−ligand complexes, protein−protein complexes, protein variants, or very large polymers.

Implementation for the Study of Protein−Ligand Interactions. As presented above, the method is general and can be tuned to a different scenario. For example, in the case of a drug, we can introduce a perturbational potential (see eq 2) accounting for solvation (sol) and nonbonded (van der Waals (vw) and electrostatic (ele)) interactions:

$$H = H_{\text{prot}}^{(M)} + H_{\text{lig}} + V_{\text{ele}}(\vec{d}_{ij}, \varepsilon_{ij}) + V_{\text{vw}}(\vec{d}_{ij}) + V_{\text{solv}}(\vec{d}_{ij}) \tag{22}$$

with

$$H_{\text{lig}} = \sum_{j=1}^{3N_{\text{lig}}}\frac{p_{\text{lig},j}^2}{2m_{\text{lig},j}} + V_{\text{lig}}(\vec{d}_{jk}^{\,\text{int}}) \tag{23}$$

where $\vec{d}_{ij}$ is the vector distance between atom i of the protein and atom j of the drug. N_{lig} is the number of atoms of the drug, $P_{\text{lig},j}$ is the jth component of its momentum vector, and $m_{\text{lig},j}$ is its corresponding mass. V_{lig} is the potential energy of the atoms of the drug due to their internal interactions, and $\vec{d}_{jk}^{\,\text{int}}$ is the intramolecular vector distance between atom j and atom k in the drug.

The related perturbation force is computed as

$$\vec{F}_i = -\vec{\nabla}_iV_{\text{ele}}(\vec{d}_{ij}, \varepsilon_{ij}) - \vec{\nabla}_iV_{\text{vw}}(\vec{d}_{ij}) - \vec{\nabla}_iV_{\text{sol}}(\vec{d}_{ij}) \tag{24}$$

while for drug atoms, forces are computed as

$$\begin{aligned}\vec{F}_j = &-\vec{\nabla}_jV_{\text{ele}}(\vec{d}_{ij}, \varepsilon_{ij}) - \vec{\nabla}_jV_{\text{vw}}(\vec{d}_{ij}) - \vec{\nabla}_jV_{\text{sol}}(\vec{d}_{ij})\\ &- \vec{\nabla}_jV_{\text{lig}}(\vec{d}_{jk})\end{aligned} \tag{25}$$

where the electrostatic forces are computed using the Mehler and Solmajer formalism,[18] van der Waals is represented using a soft potential, and solvation is accounted for by means of the Lazaridis−Karplus functional.[19] Note that perturbational forces modulate protein motion, so the approach captures ligand-induced changes in protein dynamics. As a first approximation in this work, we have used rigid drugs, and we neglect the additional term ($\vec{\nabla}_jV_{\text{lig}}$).

Implementation for Protein−Protein Interactions. The method outlined before can be generalized to the motion of a protein, A, in the field of another one, B, to simulate for example their relative diffusion or direct interaction by using

$$H = H_{\text{A}}^{(M_{\text{A}})} + H_{\text{B}}^{(M_{\text{B}})} + V_{\text{AB}}(\vec{d}_{ij}) \tag{26}$$

This Hamiltonian accounts for the unperturbed Hamiltonian of both proteins ($H_{\text{A}}^{(M_{\text{A}})}$ and ($H_{\text{B}}^{(M_{\text{B}})}$) up to a given degree of accuracy (M_{A} for protein A and M_{B} for protein B) and the interaction potentials V_{AB}, which are defined either as the sum of a soft van der Waals term and electrostatic and solvation energies or as a coarse-grained potential, which can be expressed using either physical or knowledge-based potentials. For the demonstration purposes of this paper, we have used a soft residue-centered van der Waals term, but extension of the method to other energy functionals is straightforward.

Table 1. Different Metrics Showing the Goodness of the ED/MD Method for a Few Representative Proteins Considering Essential Deformation Spaces Defined by Three Thresholds of Variance[a]

PID	# atoms	metrics	variance 85%	variance 90%	variance 95%
1ERG	1078/70	# modes	32/10	53/16	107/28
		speed-up	62/2673	45/1922	40/1744
		simil. index	0.80/0.98	0.82/0.98	0.84/0.99
		rel. cRMSd	0.158/0.049	0.141/0.039	0.119/0.052
1FVQ	1085/72	# modes	66/12	99/20	166/38
		speed-up	80/2525	73/2306	58/1864
		simil. index	0.83/0.94	0.86/0.94	0.88/0.96
		rel. cRMSd	0.294/0.087	0.271/0.117	0.328/0.145
2J8B	1192/77	# modes	63/23	97/25	168/59
		speed-up	40/2237	33/2043	30/1651
		simil. index	0.74/0.98	0.75/0.99	0.78/0.99
		rel. cRMSd	0.136/0.266	0.151/0.251	0.161/0.234
1A19	1433/89	# modes	61/29	97/42	173/66
		speed-up	23/1824	21/735	14/626
		sim. index	0.85/0.99	0.87/0.99	0.90/0.99
		rel. cRMSd	0.287/0.195	0.281/0.178	0.217/0.162
1A2P	1700/108	# modes	64/33	102/50	185/80
		speed-up	35/2278	22/1481	17/834
		sim. index	0.84/0.98	0.85/0.99	0.89/0.99
		rel. cRMSd	0.293/0.047	0.311/0.016	0.299/0.014
1BJ7	2387/150	# modes	114/36	171/55	285/97
		speed-up	26/1747	21/1464	17/1134
		sim. index	0.85/0.99	0.87/0.99	0.89/0.99
		rel. cRMSd	0.33/0.223	0.327/0.205	0.333/0.181
1B39	4689/290	# modes	113/61	180/94	321/163
		speed-up	23/625	18/485	12/276
		sim. index	0.89/0.99	0.90/1.00	0.90/1.00
		rel. cRMSd	0.32/0.180	0.287/0.165	0.293/0.156

[a] Values before the slash correspond to an all-heavy atoms model and those after the slash to a C_α reduced protein model. In all cases, the size of the essential space is indicated (the total space can be easily computed as $3N - 6$; with N = number of atoms of the protein), as well as the relative cross RMSD between ED/MD and MD trajectories[25] and Hess's similarity index for the importance space defined by variance threshold. The speed-up refers to a GROMACS4 reference calculation performed using MoDEL standard set-up.[5]

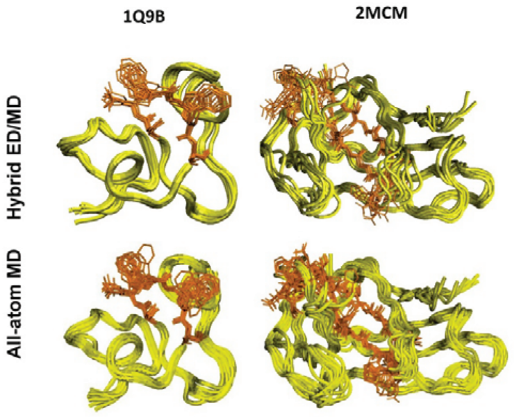

Figure 1. Performance of the ED/MD method. Schematic representation of the ensembles obtained for two randomly selected ligand-binding proteins (2MCM and 1Q9B) using (top) an ED/MD simulation using a hybrid representation of the protein, atomistic for residues near the binding site and coarse-grained (C_α only) for the rest, and (bottom) an all atom MD simulation. More details are given in the caption to Table 2.

Table 2. SimIlarity in Structure and Dynamics at Binding Site Residues for Two Proteins, 2MCM and 1Q9B, As Determined from Atomistic MD Trajectories and ED/MD Calculations Using a Hybrid All (Heavy)Atoms/Coarse Grained Model[a]

	RMSD from time-averaged structure			
protein	MD	ED	cross RMSD (MD/ED)	similarity index (MD/ED)
2MCM	1.7	1.4	0.18	0.68
1Q9B	0.8	0.8	0.07	0.59

[a] RMSD (in Å) refers to the corresponding average structure (before slash, MD; after slash, ED/MD). Residues at binding site were from 32 to 41 and from 91 to 100 for 2MCM and residues 19 to 24 for 1Q9B.

Implementation for the Study of Mutated Proteins. Within our approach, this is done by replacing at each mutant site the "r" atoms from the wild type protein with the "s" new atoms of the mutated amino acid. The MD-average structure of the wild type is taken as a reference; standard residue libraries and relaxation protocols are used to refine the starting model of the mutated protein. The trajectory for the wild type is then obtained by using the effective Hamiltonian:

$$H = H_{<}^{(M)} - H_{\text{prot}-\text{r}} + H_{\text{s}} + H_{\text{prot}-\text{s}} \tag{27}$$

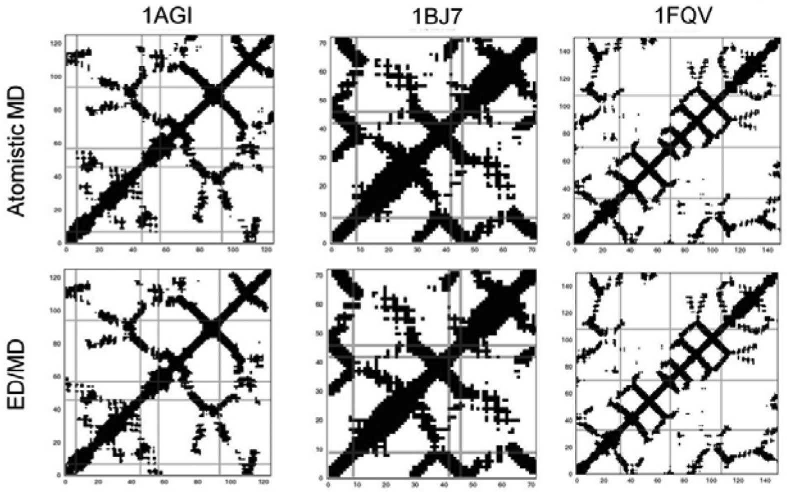

Figure 2. Performance of the ED/MD method for protein mutants. Contact maps for three randomly selected proteins (1AGI, 1BJ7, and 1FQV) subjected to Ala mutations (residues Y7, F46, I57, and R94 for 1AGI; Y33, Y70, and T108 for 1BJ7; and H9, C42, and Y46 for 1FQV) as determined from atomistic MD simulations (top) and ED/MD residues (bottom). Mutated residues can be easily localized by straight lines in gray (contact radii 5 Å).

where $H_<^{(M)}$ is the basic Hamiltonian defined above without the contributions of the removed atoms, $H_{\text{prot}-\text{r}}$ is the energy due to the interactions between removed atoms (r) and the rest of the protein, H_s is the intramolecular Hamiltonian of the added mutation atoms (s), and $H_{\text{prot}-\text{s}}$ reproduces the interactions between added atoms and the protein. All three of these Hamiltonians can be reproduced by means of simple potentials, like those used in the elastic network model:

$$H_{\text{prot}-x} = \frac{1}{2}\sum_{j=1}^{x}\sum_{i=1}^{N-r} K_{ij}^{x}(\vec{D}_{ij} - \vec{D}_{ij}^{0})^{2} \qquad (28)$$

$$H_{x} = \frac{1}{2}\sum_{i,j=1}^{x} L_{ij}^{x}(\vec{D}_{ij} - \vec{D}_{ij}^{0})^{2} \qquad (29)$$

where x stands for r or s depending on if the Hamiltonian is referring to removed or added atoms, $\vec{D}_{ij}$ is the vector between i and j, and $\vec{D}_{ij}^{0}$ is the corresponding equilibrium distance. K_{ij}^{x} is the protein-mutation site stiffness matrix (product of Kirchoff matrix with a spring constant of 10 kcal/mol Å^2 and a cutoff radius of 8 Å[20−22]), and L_{ij}^{x} corresponds to the stiffness matrix for the x atoms.

Polymer Simulation. As a linear polymer, DNA is particularly suited to be studied by this methodology. Through the work of the Ascona B-DNA consortium,[9] we have a description of the essential dynamics of every possible B-DNA tetramer sequence with *state of the art* force field and simulation conditions. We can therefore represent the structure of any DNA duplex of any length as a superposition of its constituent tetramers. For example, the Dickerson−Drew dodecamer sequence d(CGCGAATTCGCG)$_2$ can be modeled as nine overlapping tetramers: CGCG, GCGA, CGAA, etc. To perform MD on the polymer, we take its coordinates, distribute them to give the coordinates of the constituent tetramers, use the equations above to calculate the associated essential-dynamics derived forces, and then merge these to give the forces acting on the whole polymer. Note that, except at the termini, each atom appears in four overlapping tetramers, so the force acting on it is taken as the simple average:

$$\vec{F}_i = \frac{1}{4}\sum_{\text{te}=1}^{4} \vec{F}_i^{\text{te}} \qquad (30)$$

where "te" stands for each of the four tetrads to which any given atom of the DNA can be assigned.

RESULTS

Reference MD simulations of proteins were taken from our MoDEL database,[5] while those for the DNA dodecamer were generated using the standard ABC protocols.[9]

Basic Performance. The ED/MD method produces trajectories for entire proteins, which are very similar by all of the analyzed metrics to the original MD simulations, even when a reduced essential space is selected (see Table 1). Thus, for variance thresholds in the range 85−95%, we find similarity indexes between MD and ED/MD trajectories typically around 0.80−0.85 for all heavy atom representations.

Considering the noise of the original MD simulation (self-similarities around 0.8), this indicates that in terms of essential movements of heavy atoms, MD and ED/MD trajectories are indistinguishable. The same conclusion can be reached by looking at the RMSD between individual snapshots of ED/MD and MD simulations, which are very similar to those obtained by comparing different MD snapshots. The performance of ED/MD simulations is even better when comparisons are limited to C$_\alpha$ representations of the protein (Table 1), since in this case ED/MD and MD trajectories are almost identical.

The performance of the ED/MD method in the intermediate case—when part of the system is represented with atomistic detail and the rest at the C$_\alpha$ level—is illustrated in Figure 1 and Table 2. As expected from the simplicity of the method, the ED/MD approach (even without any code optimization) leads to dramatic speed-ups (up to 10^3 in some cases) with respect to the highly optimized GROMACS[24] program. Altogether, our results demonstrate that our method works efficiently in both atomistic and coarse-grained models and is able to recover with high accuracy the dynamic characteristics of the proteins as determined by reference atomistic MD simulation.

Analysis of Protein Variants and Mutations. We have evaluated the ability of the ED/MD method to reproduce the dynamics of a mutated protein starting from the wild-type (WT) trajectory. We have studied three proteins (1AGI, 1BJ7, and 1FVQ) for which the dynamics of variant proteins containing at least three mutations to alanine in randomly selected residues were determined from atomistic MD as well as for all-(heavy) atom ED/MD calculations (using the ED space determined from WT-MD trajectory; see Methods). Global trajectories obtained from MD and ED/MD are very similar in all cases; e.g., RMSDs between MD and ED/MD simulations are around 0.1−0.3 Å, and contact maps are perfectly preserved, even in the vicinities of the mutated residues (Figure 2).

Interestingly, the flexibilities around mutated residues found in our ED/MD simulations are very similar to those found in the reference atomistic MD calculations (see Table 3). In summary, ED/MD seems able to correctly predict the structure and dynamics of protein variants containing (even multiple) site mutations, saving a dramatic amount of computer time compared to conventional simulations, especially if CG representation is used in regions of reduced interest.

Table 3. Fluctuations (RMSF in Å) for Selected Residues in Three Proteins before (As Determined from Atomistic MD Simulations) and after (As Determined from Both MD and ED/MD Simulations) Mutation to Alanine

	mutated residue	wild type MD	mutated MD	mutated ED/MD
1AGI	TYR7	1.2	0.5	0.5
	PHE46	0.9	0.2	0.5
	ILE57	1.1	0.5	0.4
	ARG94	1.9	0.7	0.7
1BJ7	TYR33	0.7	0.4	0.2
	TYR70	0.6	0.6	0.6
	THR108	0.4	0.3	0.2
1FVQ	HIS9	0.8	0.8	0.5
	CYS42	0.5	0.5	0.4
	TYR46	0.6	0.6	0.5

Study of Protein Interactions. A very common scenario in medicinal chemistry research is that computational methods are desired to analyze the mode and strength of binding to a target protein of thousands or maybe millions of potential drug molecules. The use of MD to analyze one protein−drug interaction is straightforward, but the study of thousands of complexes is simply impossible due to time and resource constraints. The approach outlined here combined with a simple model of protein−drug interaction and a reduced drug friction provides the user with a useful and inexpensive alternative to analyze the structural and dynamic properties of drug−protein complexes. A couple of examples are provided in Figure 3, where we represent snapshots obtained for the multi-microsecond diffusion of two ligands to their binding sites. The method is very efficient, especially when we use a CG representation of the protein away from the binding site (accelerations above 10^3 from GROMACS, thus opening up the possibility of the high-throughput use of this technique.

Generalization of the method to study protein−protein complexes is straightforward (see Methods), which means that our approach can be used to study protein aggregation using flexible rather than rigid models of proteins. This is illustrated in the study of the Barnase−Barstar complex (see Figure 3 bottom), where the ED/MD method allows the simulation of a multi-microsecond-long process in a few hours using a laptop computer.

The ED/MD Method for the Study of Polymers. Simulations on the Dickerson−Drew dodecamer[25] d-$(CGCGAATTCGCG)_2$ performed from prestored dynamics of tetramer segments using the "build-up" ED/MD method are in excellent agreement with the results of a conventional microsecond-long atomistic MD simulation.[6] Both the average

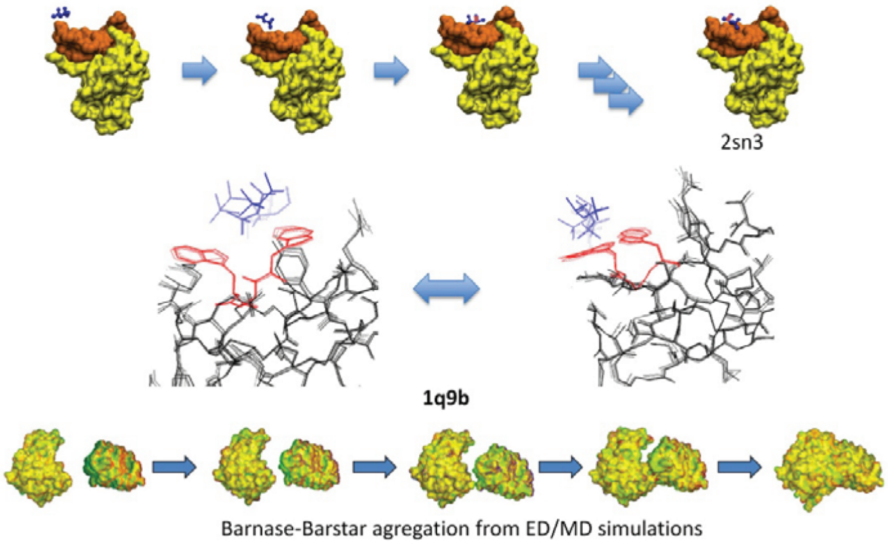

Figure 3. Performance of the ED/MD method for protein−ligand and protein−protein interactions. Top: Detail of the migration of binding of MDP molecule ((4*S*)-2-methyl-2,4-pentanediol) around the binding site of 2SN3 diffusion process is expected to happen in the multi-microsecond scale. Middle: Detail of the flexibility of key residues at the binding site explored during the binding of MDP molecule to 1Q9B; such flexibility is explicitly explored during docking of the drug. Bottom: General view of the diffusion and binding of the Barnase−Barstar complex obtained in an unbiased ED/MD simulation using full dynamics of both proteins with atomistic detail.

values (Figure 4) and fluctuations (Table 4) of helical parameters match the conventional MD values closely.

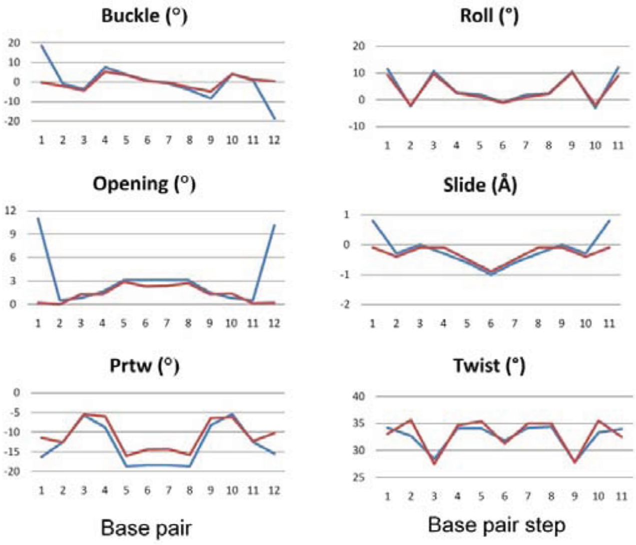

Figure 4. Performance of the ED/MD method for DNA simulations. Averaged helical parameters along sequence (translations in Å, rotations in degrees) for the Dickerson dodecamer d(CGCGAATTCGCG)$_2$ obtained from atomistic MD simulations (blue lines) and from ED/MD simulations (red lines) using the tetramer-based "build-up" approach.

Table 4. Average RMS Fluctuations in Selected Base Pair-Step Helical Parameters for the Simulations of the Dickerson−Drew Dodecamer[a]

parameter	MD	ED/MD
shift	0.75	0.99
slide	0.69	0.76
rise	0.42	0.27
tilt	5.4	3.2
roll	6.7	6.1
twist	6.2	8.0

[a] Values of translational parameters are in Å, and rotational ones are in degrees.

The larger deviations at the termini are as expected, as the ED-based method used data from tetramers embedded in the center of long oligomers and so cannot capture end effects. In terms of computer efficiency, the present simulations (performed at the 90% variance level) are 30−40 times faster than the original atomistic MD simulations. The performance of ED/MD is expected to increase for larger systems; as an example, simulations of GC microcircles of 90 base pairs are also in good agreement with the results obtained in previous conventional MD studies.[7] Circles built with LK = 7 or 9 remain as open circles while the LK = 5 and LK = 11 circles writhe to negatively and positively supercoiled structures, respectively, that resemble closely the previously obtained conformations (Figure 5), reproducing well some fine details of supercoiling such as the ratio of the bending and twist persistence lengths of the DNA.[9] In computational terms, this calculation is around 50 times faster than reference calculations. Considering the nearly linear scaling of ED/MD calculations, we can expect dramatic speed-ups for much larger DNA systems that would be completely impracticable to study with current atomistic MD algorithms.

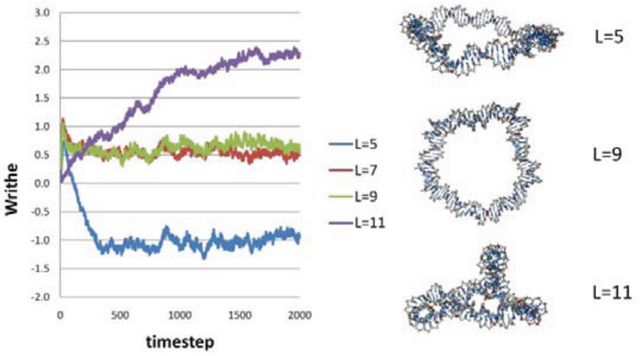

Figure 5. ED/MD simulations of alternating GC 90-mer microcircles with varying degrees of superhelical stress. Moderately under- or overwound circles (linking numbers 7 and 9) remain unwrithed. A strongly underwound circle (L = 5) rapidly supercoils to a negatively writhed structure, while a strongly overwound circle (L = 11) more slowly adopts a positively writhed configuration. Examples of the final structures from the simulations are shown in the right-hand panel.

■ CONCLUSIONS

We present here a new method for simulating the dynamics of macromolecules with a quality comparable to that with atomistic MD, but with a computational cost that can be 1/1000 of the original one. The method relies on the observation that most macromolecules move in a quite harmonic regime around the equilibrium geometry, and a small number of essential deformation modes (obtained from Berendsen's essential dynamics procedure[11]) can represent most of the dynamics sampled during the trajectory (we do however emphasize that in situations where this is not the case, this method is less applicable). Accordingly, we can obtain an accurate and inexpensive picture of the dynamics of a protein by applying Langevin equations on the precomputed essential deformation space. This general idea has a direct application for reinforcing sampling in previously computed trajectories. More interestingly, the method can be easily extended to study systems other than those for which an MD trajectory is already available. For polymeric structures such as DNA, the method can use a "divide-and-conquer" approach where the dynamics of the entire polymer are obtained by combining the previously computed dynamics of small overlapping fragments. In the case of proteins, where the divide-and-conquer approach does not have a straightforward implementation, the dynamics can be computed by perturbing the dynamical characteristics of a similar, previously studied, protein. Similar concepts can be used to simulate the dynamics of protein−drug and protein−protein complexes.

The method can work with different levels of accuracy and any level of resolution. A unique aspect of this approach is that the reduction in resolution of distant zones does not affect the accuracy of the trajectory in the atomistic region at all. The method is extremely efficient from a computational point of view and complements very well the current scenario, where initiatives such as MoDEL,[7] Dynameomics,[8] or the ABC consortium[9] are increasingly providing multi-terabyte databases of the dynamics of macromolecules.

We would emphasize that the approach we have developed here is quite distinct from those based on analyzing MD trajectories in terms of Markov state models.[26−28] That approach also reduces the dimensionality of the system, but by defining a limited number of states and calculating the rates of transitions between them. The method then allows one to rapidly evaluate (for fully converged trajectories) complex kinetics in cases where sampling populates many metastable conformations. Those approaches however provide no method to generate new trajectories for systems which are related to, but not identical to, that used to generate the MSM and, in any case, only include a very small number of "atomically detailed" models of the system of interest.

The code implementing our ED/MD algorithm is just a beta version, which has not been yet optimized to improve efficiency. Despite that, it provides an excellent speed-up with respect to the very heavily optimized GROMACS code[24] by making use of a standard, full, all-heavy atom representation of MoDEL setups:[7] around 50 times for an 85% variance threshold and around 35 times for a stricter 95% variance threshold (see Table 1). Molecules with simpler deformational space, such as nucleic acids (see below) and proteins with well-defined hinges, provide the largest gains. As expected, dramatic gains in efficiency are achieved if only C_α representations are considered, leading in some cases to a speed-up on the order of 10^3 with respect to reference GROMACS calculations (Table 1), suggesting that for selected cases, millisecond to second simulations could be feasible with our methodology. More interestingly, in many practical cases where atomistic detail is only required for a limited part of the protein (for example the binding site), the speed-up of our unoptimized ED/MD code can easily reach a factor of 1000 with respect to the fastest atomistic MD code without decreasing the quality in the representation in the atomistic region (Figure 1 and Table 2).

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: modesto@mmb.pcb.ub.es or Charles.laughton@nottingham.ac.uk.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the Spanish Ministry of Science and Innovation (BIO2009-10964 and Consolider E-Science), Instituto de Salud Carlos III (INB-Genoma España and COMBIOMED RETICS), Fundación Marcelino Botín, the UK Biotechnology and Biological Sciences Research Council (BB/F011407/1), EU-SCALALIFE project, and the HPC-Europa project.

■ REFERENCES

(1) Teilum, K.; Olesen, J. G.; Kragelund, B. B. Functional aspects of protein flexibility. *Cell. Mol. Life Sci.* **2009**, *66*, 2231−2247.

(2) Karplus, M.; Kuriyan, J. Molecular dynamics and protein function. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6679−6685.

(3) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* **1996**, *118*, 11225−11236.

(4) Roy, J.; Laughton, C. A. Long-timescale molecular-dynamics simulations of the major urinary protein provide atomistic interpretations of the unusual thermodynamics of ligand binding. *Biophys. J.* **2010**, *99*, 218−226.

(5) Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y.; Wrtiggers, W. Atomic-level characterization of the structural dynamics of proteins. *Science* **2010**, *330*, 341−346.

(6) Perez, A.; Luque, F. J.; Orozco, M. Dynamics of B-DNA on the microsecond timescale. *J. Am. Chem. Soc.* **2007**, *129*, 14739−14734.

(7) Meyer, T.; D'Abramo, M.; Hospital, A.; Rueda, M.; Ferrer-Costa, C.; Pérez, A.; Carrillo, O.; Camps, J.; Fenollosa, C.; Repchevsky, D.; Gelpí, J. L.; Orozco, M. MoDEL (Molecular Dynamics Extended Library): A database of atomistic molecular dynamics trajectories. *Structure* **2010**, *18*, 1399−1409.

(8) Van der Kamp, M. W.; Schaeffer, R. D.; Jonsson, A. L.; Scouras, A. D.; Simms, A. M.; Toofanny, R. D.; Benson, N. C.; Anderson, P. C.; Merkley, E. D.; Rysavy, S.; Bromley, D.; Beck, D. A. C.; Daggett., V. Dynameomics: A comprehensive database of protein dynamics. *Structure* **2010**, *18*, 423−435.

(9) Lavery, R.; Zakrzewska, K.; Beveridge, D. A systematic molecular dynamics study of nearest-neighbor effects on base pair and base pair step conformations and fluctuations in B-DNA. *Nucl. Acid Res.* **2010**, *38*, 299−313.

(10) Curuksu, J.; Zacharias, M.; Lavery, R.; Zakrzewska, K. Local and global effects of strong DNA bending induced during molecular dynamics simulations. *Nucl. Acid Res.* **2009**, *37*, 3766−3773.

(11) Amadei, A.; Linssen, A. B. M.; Berendsen, H. J. C. Essential dynamics of proteins. *Proteins* **1993**, *17*, 412−415.

(12) Meyer, T.; Ferrer-Costa, C.; Pérez, A.; Rueda, M.; Bidon-Chanal, A.; Luque, F. J.; Laughton, C. A.; Orozco, M. Essential Dynamics: A tool for efficient trajectory compression and management. *J. Chem. Theory Comput.* **2006**, *2*, 251−258.

(13) Gardiner, C. W. *Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences*; Springer: Berlin, 1989.

(14) Van Kampen, N. G. *Stochastic Processes in Physics and Chemistry*; North-Holland: Amsterdam, 1981.

(15) Allen, M. P.; Tildesley, D. J. *Computer Simulations of Liquids*; Clarendon Press: Oxford, U. K., 1989.

(16) Humphreys, D. D.; Friesner, R. A.; Berne, B. J. A multiple-time-step molecular dynamics algorithm for macromolecules. *J. Phys. Chem.* **1994**, *98*, 6885−6892.

(17) Streett, W. B.; Tildesley, D. J.; Saville, G. Multiple time-step methods in molecular dynamics. *Mol. Phys.* **1978**, *35*, 639−648.

(18) Orozco, M.; Luque, F. J. Theoretical methods for the representation of solvent in biomolecular systems. *Chem. Rev.* **2000**, *100*, 4187−4225.

(19) Lazaridis, T.; Karplus, M. Effective energy function for proteins in solution. *Proteins* **1999**, *35*, 133−152.

(20) Kovacs, J. A.; Chacón, P.; Abagyan, R. Predictions of protein flexibility: first-order measures. *Proteins* **2004**, *56*, 661−668.

(21) Emperador, A.; Carrillo, O.; Rueda, M.; Orozco, M. Exploring the suitability of coarse-grained techniques for the representation of protein dynamics. *Biophys. J.* **2008**, *95*, 2127−2138.

(22) Rueda, M.; Chacón, P.; Orozco, M. Thorough validation of protein normal mode analysis: A comparative study with essential dynamics. *Structure* **2007**, *15*, 565−575.

(23) Betancourt, M. R.; Skolnick, J. Universal similarity measure for comparing protein structures. *Biopolymers* **2001**, *59*, 305−309.

(24) Hess, B.; Kutzner, C.; Van der Spoel, D.; Lindahl, E. Gromacs 4: Algorithms for highly efficient, load balanced, and scalable molecular simulation. *J. Chem. Theory Comput.* **2008**, *4*, 435−447.

(25) Drew, H. R.; Wing, R. M.; Takano, T.; Broka, C.; Tanaka, S.; Itakura, K; Dickerson, R. E. Structure of a B-DNA dodecamer: conformation and dynamics. *Proc. Natl. Acad. Sci. U. S. A.* **1981**, *78*, 2179−2183.

(26) Schütte, C.; Noé, F.; Lu, J.; Sarich, M.; Vanden-Eijnden, E. Markov state models based on milestoning. *J. Chem. Phys.* **2011**, *134*, 204105.

(27) De Sancho, D.; Best, R. B. What is the timescale for alpha-helix nucleation? *J. Am. Chem. Soc.* **2011**, *133*, 6809−6816.

(28) Lane, T. J.; Bowman, G. R.; Beauchamp, K.; Voelz, V. A.; Pande, V. S. Markov state model reveals folding and functional dynamics in ultra-long MD trajectories. *J. Am. Chem. Soc.* **2011**, *133*, 18413−18419.