JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

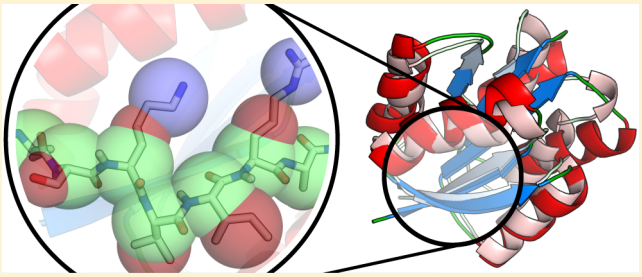
PaLaCe: A Coarse-Grain Protein Model for Studying Mechanical Properties

Marco Pasi, Richard Lavery,* and Nicoletta Ceres

Bases Moléculaires et Structurales des Systèmes Infectieux, Univ. Lyon I/CNRS UMR 5086, IBCP, 7 Passage du Vercors, 69367 Lyon, France

Supporting Information

ABSTRACT: We present a coarse-grain protein model PaLaCe (Pasi–Lavery–Ceres) that has been developed principally to allow fast computational studies of protein mechanics and to clarify the links between mechanics and function. PaLaCe uses a two-tier protein representation with one to three pseudoatoms representing each amino acid for the main nonbonded interactions, combined with atomic-scale peptide groups and some side chain atoms to allow the explicit representation of backbone hydrogen bonds and to simplify the treatment of bonded interactions. The PaLaCe force field is composed of physics-based terms, parametrized using Boltzmann inversion of conformational probability distributions derived from a protein structure data set, and iteratively refined to reproduce the experimental distributions. PaLaCe has been implemented in the MMTK simulation package and can be used for energy minimization, normal mode calculations, and molecular or stochastic dynamics. We present simulations with PaLaCe that test its ability to maintain stable structures for folded proteins, reproduce their dynamic fluctuations, and correctly model large-scale, force-induced conformational changes.



1. INTRODUCTION

Our aim in developing a new coarse-grain protein model was primarily to investigate the mechanical properties of proteins. The idea for this approach grew out of our earlier studies of mechanical properties using elastic network models associated with a residue-dependent coarse-grain representation. Such models can be very successful in describing deformations within the energy basin around a single conformation,[1,2] and indeed we found them to be very useful in describing enzyme active site properties[3,4] and even in treating the impact of point mutations.[5] However, a simple harmonic approach cannot deal with transitions between energy basins, or, more generally, with large-scale conformational changes.

We therefore began to design a more refined coarse-grain model (using a bottom-up approach, in contrast to the top-down methods for coarse-graining atomic representations),[6,7] with three main criteria. First, we wanted to maintain the level of coarse-graining that had worked well in our elastic network studies (one to three pseudoatoms per amino acid).[8] Second, in order to study large conformational changes, it was necessary to avoid introducing any restraints on secondary structure (common in other coarse-grain protein models).[9] Third, we wanted to make parameter optimization fast and automatic, so that the force field, or the protein representation, could be changed at will without much effort.

Although we effectively maintained a pseudoatom representation of amino acids, the second of our design criteria meant that we required an explicit representation of peptide backbone hydrogen bonds. This led use to choose a two-tier representation, where the side chain and Cα pseudoatoms were complemented by an atomic-scale peptide backbone.[10−13] Explicit peptide groups also enabled us to fit backbone torsions simply, without the complex coupling associated with effective Cα−Cα bonds.[14,15] We choose to maintain this advantage for the side chains by using an explicit Cβ center (and some additional atomic construction centers for longer side chains) again to avoid complex torsional coupling.

Using this two-tier representation, we developed a force field describing bonded interactions within the peptide backbone and side chains, peptide hydrogen bonds, and nonbonded repulsion/dispersion and electrostatic interactions between the pseudoatoms. To avoid having to treat solvation effects implicitly via nonbonded two-body terms, we developed an explicit one-body solvation contribution for each pseudoatom.[16] Despite a force field composed of terms with a physics-based formulation, the third of our design criteria led us to choose to parametrize the terms in a "knowledge-based" manner using data from known protein structures. Coupling this approach with an iterative refinement procedure effectively made complete parametrization possible in roughly one week of computation.

The resulting method, termed PaLaCe (named after its three developers), is presently implemented in MMTK (Molecular Modeling ToolKit),[17] enabling it to be used for normal mode calculations, energy minimization, and molecular or stochastic

Received: September 11, 2012
Published: November 9, 2012

   dx.doi.org/10.1021/ct3007925 | J. Chem. Theory Comput. 2013, 9, 785−793

dynamics. This article describes the PaLaCe coarse-grain model and presents some preliminary tests of its performance in preserving the folded structure of soluble proteins, in describing their dynamic fluctuations, and in modeling large-scale deformations created by external forces.

2. MAPPING

The PaLaCe coarse-grained protein model, illustrated in Figure 1, uses a two-tier representation of polypeptide chains. The first

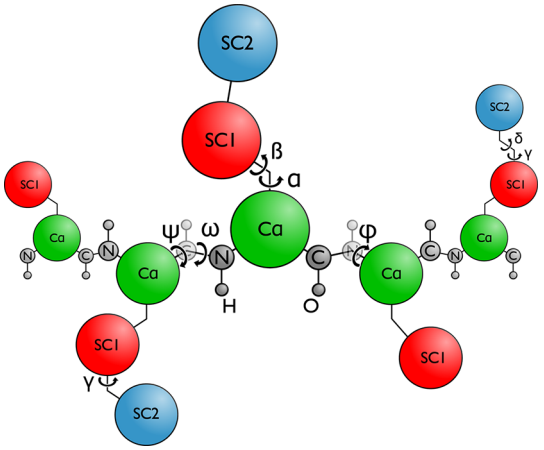


Figure 1. Schematic PaLaCe model of a pentapeptide (Val–Met–Glu–Ile–Lys). Colored spheres show the first-tier pseudoatoms used for nonbonded interactions (Cα, SC1, and SC2). Second-tier atoms are shown in stick representation. Peptide hydrogen bonding atoms are colored gray. Arrows indicate torsional degrees of freedom. Side chains (except Gly) include Cβ, and some side chains (Met, Ile, Lys, Arg) use further atomic centers (see text).

tier is used for nonbonded interactions, other than main chain hydrogen bonds, and involves a pseudoatom representation of the amino acids with one pseudoatom for the Cα backbone and one or two pseudoatoms (with the exception of glycine) for the side chain. This mapping follows the scheme proposed by Zacharias for the modeling of coarse-grain protein–protein interactions.[8] The side chains of the smaller amino acids (Ala, Cys, Asp, Leu, Asn, Pro, Ser, Thr, Val, and Ile) are mapped to a single bead (SC1) placed at the geometrical center of the side chain heavy atoms. Larger side chains (Tyr, Phe, His, Glu, Gln, Trp, Lys, Arg, Met) are modeled with a first pseudoatom placed midway along the Cβ-Cγ bond (or at the Cβ for Phe, His, and Tyr to avoid β torsions with three aligned atoms, see Figure 1). A second pseudoatom (SC2) is located at the geometrical center of the remaining heavy atoms.

The second tier is used for bonded interactions and for backbone hydrogen bonds. It involves atomic beads for N, C′, and Cα. Backbone oxygen and amide hydrogen positions, used for hydrogen bonding, are geometrically constructed on the basis of the Cα, N, and C′ positions, avoiding any additional degrees of freedom. The choice of adding explicit peptide groups to the more common Cα-only backbone model[7,9,14,18−20] not only allows us to treat backbone hydrogen bonds explicitly but also simplifies the treatment of backbone valence angle and torsion terms (which, in the case of a Cα-only model, lead to complex coupling between the resulting pseudobonds).[14,15] This approach additionally allows an explicit treatment of peptide *cis*/*trans* isomerization.

The second tier representation is also used for bonded interactions within the side chains, where the introduction of extra atomic positions again avoids the complex angle and torsion coupling that would otherwise result from directly using pseudobonds (see Supporting Information Figure S1). This involves creating a Cβ atom for all residues except Gly and Ala and adding either two or three atomic centers to Met, Lys, Arg, and Ile side chains (see Figure 1).

3. ENERGY FUNCTION

The PaLaCe force field is composed of bonded (E_b) and nonbonded (E_{nb}) energy terms.

3.1. Bonded Interactions.

$$E_b = E_{bon} + E_{ang} + E_{tor}$$

Bond lengths (r) and valence angles (θ) are constrained using harmonic potentials with respect to their ideal values (r_0 and θ_0, respectively) with force constants k_r and k_θ:

$$E_{bon} = \sum_{bonds} k_r(r - r_0)^2 \tag{1}$$

$$E_{ang} = \sum_{angles} k_\theta(\theta - \theta_0)^2 \tag{2}$$

Torsion angles (τ) are modeled using sums of cosine terms:

$$E_{tor} = \sum_{(\alpha,\beta,\omega)} E_{tor}(\tau) + \sum_{(\alpha,\beta),(\phi,\psi)} E_{tor}(\tau_1, \tau_2) \tag{3}$$

$$E_{tor}(\tau) = \sum_{i=1}^{n} w_i \cos(i\tau + \varphi_i) + c \tag{4}$$

$$E_{tor}(\tau_1, \tau_2) = E_{tor}(\tau_1) + E_{tor}(\tau_2) + \sum_{i=1}^{m}\sum_{j=1}^{n} w_{ij} \cos(i\tau_1 + \varphi_{ij}^{(1)}) \cos(j\tau_2 + \varphi_{ij}^{(2)}) + c \tag{5}$$

with coefficients w_i and phases φ_i, and a constant c ensuring that $E_{tor} \geq 0$.

Equation 5 is only used to model adjacent torsions showing significant coupling, namely the backbone dihedrals ψ and ϕ (m,n = 3), and adjacent side chain torsions α/β of Glu, Gln, Trp, Lys, Arg, Met, and Ile and γ/δ of Arg and Lys (m,n = 3, see Figure 1).

3.2. Nonbonded Interactions.

$$E_{nb} = E_{CG} + E_{HB} + E_{sol}$$

The first term (E_{CG}) is the interaction energy between any two pseudoatoms i and j (Cα, SC1, or SC2) separated by more than three consecutive bonds.

$$E_{CG} = \sum_{pairs} \varepsilon_{ij}^{(CG)} \left[3\left(\frac{\sigma_{ij}^{(CG)}}{r_{ij}}\right)^8 - 4\left(\frac{\sigma_{ij}^{(CG)}}{r_{ij}}\right)^6 \right] + \frac{q_i q_j}{4\pi\varepsilon_0 \varepsilon_{CG}(r_{ij}) r_{ij}} \tag{6}$$

where $\sigma_{ij}^{(CG)}$ and $\varepsilon_{ij}^{(CG)}$ characterize a Lennard-Jones-like interaction using a softer repulsive term than atomic force fields that better suits the lower-resolution pseudoatoms.[8] Unit charges on the terminal pseudoatoms of charged amino acids,

dx.doi.org/10.1021/ct3007925 | J. Chem. Theory Comput. 2013, 9, 785−793

damped by a distance-dependent dielectric constant $\varepsilon_{CG}(r) = k_e r$ describe electrostatic interactions.

The backbone hydrogen bond potential, E_{HB}, is similarly composed of Lennard-Jones and electrostatic contributions:

$$E_{HB} = \sum_{i,j} \left\{ \varepsilon_{ij}^{(HB)} \left[\left(\frac{\sigma_{ij}^{(HB)}}{r_{ij}} \right)^{12} - 2 \left(\frac{\sigma_{ij}^{(HB)}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4\pi\varepsilon_0 \varepsilon_{HB}(r_{ij}) r_{ij}} \right\} \quad (7)$$

where the sum runs over the backbone atoms N, H, C, and O of nonadjacent residues, $\sigma_{ij}^{(HB)}$ are equilibrium distances, and q_i is the partial charge of atom i.

The $\varepsilon_{HB}(r)$ dielectric constant has a linear distance-dependence below a critical distance r_l and is quadratic above this distance:

$$\varepsilon_{HB}(r) = \begin{cases} k_q r^2 + 1, & r < r_l \\ 2k_q r_l r - k_q r_l^2 + 1 & r \geq r_l \end{cases} \quad (8)$$

This functionality provides long-range electrostatic damping without destabilizing short hydrogen bond interactions. It is a simplified version of the sigmoidal dielectric function developed earlier.[21−23]

Finally, E_{sol} is a one-body solvation term based on circular variance (CV), a simple measure used in directional statistics to quantify the distribution of vectors around a point. It is used here to describe the occlusion of a pseudoatom within a protein structure,[24] and in contrast to surface accessibility, it has the advantage of changing smoothly from 0 to 1 as a pseudoatom becomes buried.[16]

$$CV_i = 1 - \frac{1}{n_i} \left| \sum_{j \neq i, r_{ij} \leq r_c}^{n_i} \frac{r_{ij}}{|r_{ij}|} \right| \quad (9)$$

where the sum runs over the n_i nonbonded neighbors of pseudoatom i within a cutoff distance r_c.

More details of the amino acid dependence of the force field parameters are provided in section S1 of the Supporting Information.

4. PARAMETER OPTIMIZATION

Parameter optimization of the PaLaCe force field relies on the large body of structural information on soluble, globular proteins in their native state present in the Protein Data Bank (PDB).[25] This approach has two main advantages: first, the resulting parameter set depends only on experimental data and not on other parametrized models such as all-atom force fields; second, using data from a large number of proteins should lead to parameters that are applicable to a wide variety of protein structural classes.

The procedure we employ involves setting up an appropriate structural data set and then extracting the statistics of the interactions we include in our potential energy function in the form of probability distributions $p_N(x)$, where x is most commonly a distance between atoms or pseudoatoms, but can also be torsion angles, or the circular variance of a pseudoatom in the case of the solvation term. If the interactions represented by these distributions were independent and obeyed Boltzmann statistics, the free energy $F(x)$ could be calculated exactly from the observed probabilities using the inverse Boltzmann relation:

$$F(x) = -k_B T \ln \left[\frac{p_N(x)}{p_0(x)} \right] \quad (10)$$

where k_B is the Boltzmann constant, T is the temperature (300 K was used in this work), and $p_0(x)$ is the probability distribution of x in a reference state, where the corresponding interaction is zero. The fact that the assumptions do not hold for folded proteins[26,27] makes the definition of the reference state harder and hinders the accuracy of the resulting free energies.[20] For this reason, the $F(x)$'s obtained using eq 10 are only used as a first guess for the corresponding interaction potentials, which then undergo an iterative refinement procedure (see section 4.2). Nevertheless, considerable thought was given to defining the reference state, to improve the quality of the initial parameter set and therefore the convergence of the subsequent optimization. In particular, for nonbonded pseudoatom distance distributions, we used a freely jointed-chain polymer that allowed us to take into account the effects of chain length and connectivity, while for one-body CV distributions we used a random mixing model averaged over all structures and pseudoatom types (see section S2 of the Supporting Information for details).

4.1. Data Set. Our starting point for creating a structural data set was a set of high quality PDB structures (at least 2 Å resolution and R-factor smaller than 0.25) sharing less than 20% sequence identity, culled using the web-server PISCES.[28] Transmembrane proteins, identified by comparison with the PDBTM database,[29] were removed, as well as nonglobular proteins. The latter are defined as proteins with a radius of gyration R_g at least 15% greater than the expected value for a globular protein with a corresponding chain length.[30] This selection led to a data set of 2830 structures. Given the requirements of CV parametrization, we also excluded proteins with incomplete residues, or gaps, resulting in 1202 structures. These structures were converted into our pseudoatom representation prior to extracting distance, torsion or CV probability distributions.

Within this data set, filtering was carried out to remove the impact of the secondary structural features (α-helices, β-sheets, and β-turns) that would otherwise dominate the probability distributions involving backbone hydrogen bonds, backbone ϕ/ψ torsions, and interpseudoatom distances. Removing these regular features avoids overcounting the effects of the various interactions that contribute to their stabilization and effectively uncouples the energy contributions, leading to unbiased statistics for the conformational properties of the polypeptide chain. Probability distributions associated with ϕ and ψ were therefore computed excluding residues taking part in, or adjacent to, 2-, 3-, 4-, or 5-turns or bridges, as defined by the DSSP algorithm.[31] Similarly, interparticle distance probability distributions excluded residues within the same secondary structure (α-helix, β-strand, or β-sheet, as defined by DSSP).

4.2. Optimization Procedure. The target of the optimization is an ensemble of potentials $E(x)$ that, when employed in dynamic simulations for a set of proteins, yield probability distributions matching those deduced from the experimental structures, $p_N(x)$, to an acceptable accuracy. According to a strategy proposed by Thomas and Dill,[32] and subsequently employed by others,[33,34] we use the differences

dx.doi.org/10.1021/ct3007925 | J. Chem. Theory Comput. 2013, 9, 785−793

between $P_i(x)$, extracted from $E_i(x)$, and the experimental $p_N(x)$ to drive an iterative procedure that refines the approximate potential $E_i(x)$ at iteration i according to

$$\Delta E_i(x) = -k_B T \ln\left[\frac{p_N(x)}{p_i(x)}\right]$$
$$E_{i+1}(x) = E_i(x) + w(x)\Delta E_i(x) \quad (11)$$

where $w(x)$ is a weighting function employed to reduce the noise generated by low counts typically found at the edges of the probability distributions.[35] Each iteration involved running simulations on a training set of 48 globular, soluble, single-domain, monomeric proteins selected from the PISCES data set described in section 4.1. The training set includes architectures and topologies of the three major classes of proteins (mainly α, mainly β, and α/β), according to the CATH classification.[36] A list of these proteins is given in Supporting Information Table S1.

For each protein in the training set, we first perform an energy minimization (5000 steps of steepest descent followed by 500 steps of conjugate gradient) and then gently heat the system to 300 K over 500 ps, before carrying out 1 ns of Langevin dynamics (LD) with T = 300 K and collision frequency γ = 1 ps^{-1}, in the canonical ensemble with a time step of 5 fs, using the Molecular Modeling ToolKit (MMTK).[17] Note that longer simulations would be detrimental, particularly in the early stages of the optimization, because insufficiently refined potentials could lead to significant structural deformation. The simulated probability distributions for step i, $p_i(x)$, are then extracted using snapshots sampled every picosecond during the final 250 ps of the simulation.

This procedure is used to simultaneously optimize all nonbonded energy terms (E_{CG}, E_{HB}, and E_{sol}). Simultaneous optimization, as opposed to a sequential approach,[33] considerably speeds up the procedure and avoids any dependence on the order of optimization and notably overfitting of the terms optimized first.

The optimization is initialized using $E_0(x) = F(x)$ from eq 10 and then run in a fully automated fashion until the convergence criterion is met. Given the speedup of the potential energy calculation resulting from the simplified PaLaCe protein model, and the parallel treatment of all the proteins in the training set, a full parameter optimization (typically involving 20 iterative cycles) requires roughly one week, using one processor per protein in the training set.

Equation 11 shows that when the potential $E_i(x)$ at iteration i reproduces $p_N(x)$, then the corrector $\Delta E_i(x)$ becomes zero. In practice, convergence is limited by various assumptions, and, notably, by the functional form of the terms constituting $E(x)$.

To quantitatively measure convergence, we therefore define a merit function f_i as the sum of the integrals of the correctors $\Delta E_i^{(k)}$ for each interaction k, grouped by type:

$$f_i = \frac{1}{I_t}\sum_k^{I_t}\int |\Delta E_i^{(k)}(x)|\, dx \quad (12)$$

where the sum runs over all I_t interactions of a given type t (e.g., E_{CG}, E_{HB}) and the integral is over the entire range of x. f_i should decrease as the optimization proceeds to some limiting value. This is indeed observed (see Supporting Information Figure S2), enabling us to stop the optimization of each energy term when the corresponding f_i reaches a stable plateau value. For the present version of PaLaCe, this took between 12 (for E_{HB}) and 18 (for E_{sol}) iterations. The final set of parameters that will be discussed below results from iteration 20: $E_{20}(x)$. These parameters are provided as part of the Supporting Information.

5. PERFORMANCE TESTS

The optimized potentials following 20 iteration cycles, $E_{20}(x)$, were used to assess the ability of the PaLaCe force field to reproduce the structural, mechanical, and dynamic properties of proteins derived from experimental structures or from simulations using all-atom force fields. The results show that, even though the parametrization procedure was based on matching structural properties, the resulting parameter set can reliably be used to gain insight into both protein mechanics and dynamics. Furthermore, although a limited number of protein structures were used for our training data set during iterative optimization, we show that the resulting potentials are transferable and perform well on proteins outside the training set.

5.1. Structural Properties. The first set of tests is aimed at assessing the ability of the PaLaCe force field to preserve the native conformations of proteins during equilibrium LD simulations at 300 K. The simulation protocol is the same as that described in section 4.2, with production runs of either 10 or 100 ns. The simulations were run on two independent sets of proteins, the training set of 48 proteins (see section 4.2), and an independent test set of the same size (test set 1, see Supporting Information Table S2), selected with the same criteria used for the training set. Production runs of 100 ns were carried out on a third set composed of four proteins selected from the training set, four from the test set 1, and two further proteins previously studied for their dynamic properties,[37] homologous psychrophilic and mesophilic trypsins from *Salmo salar*.[38,39] The resulting set of 10 proteins constitutes test set 2 (see Supporting Information Table S3). The equilibrium simulations on the training set and the two test sets were compared to the starting experimental structures using a variety of metrics. In what follows, the equilibrium ensemble is defined as structures sampled every 1 ps during the last 1 ns of stochastic dynamics simulations.

Figure 2 shows the results of this comparison. The backbone RMSD and TM-score[40] were computed with respect to the

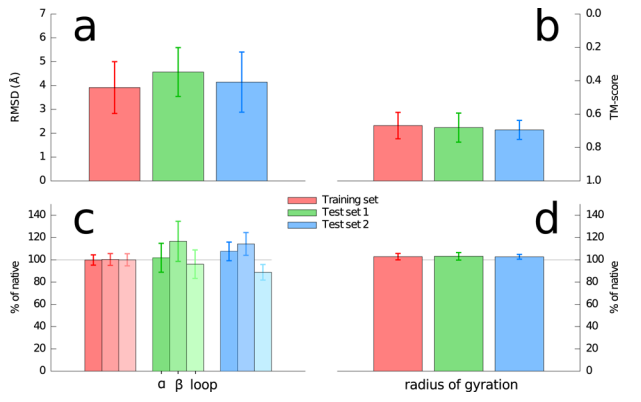


Figure 2. Matching structural properties. The equilibrium ensemble averages are averaged over each set of proteins and plotted as histograms, with error bars representing one standard deviation around the mean. See text for details of the plotted values.

dx.doi.org/10.1021/ct3007925 | J. Chem. Theory Comput. 2013, 9, 785−793

crystallographic structure and averaged over the equilibrium ensemble. The TM-score and RMSD both evaluate the structural similarity between two given conformations, but TM-score (commonly used in assessing protein fold predictions) is less sensitive than RMSD to local structural differences. TM-scores range from 0 to 1: a score between 0 and 0.2 corresponds to a comparison between two random structures,[40] while a score above 0.5 suggests that the two structures belong to the same fold.[41] Secondary structures (using DSSP)[31] and radii of gyration were calculated for the crystal structure (native) and for each structure of the equilibrium ensemble.

The results in Figure 2a show that the equilibrium trajectories remain in the neighborhood of the native structure, with average RMSD values between 3.9 and 4.6 Å. This demonstrates that the optimized PaLaCe force field is capable of preserving the native fold on time scales up to 100 ns. This is confirmed by the high TM-scores (average values between 0.67 and 0.69, Figure 2b). Further structural information is provided by the comparison of the secondary structure content and the radii of gyration (Figure 2c and d). Figure 3 illustrates the

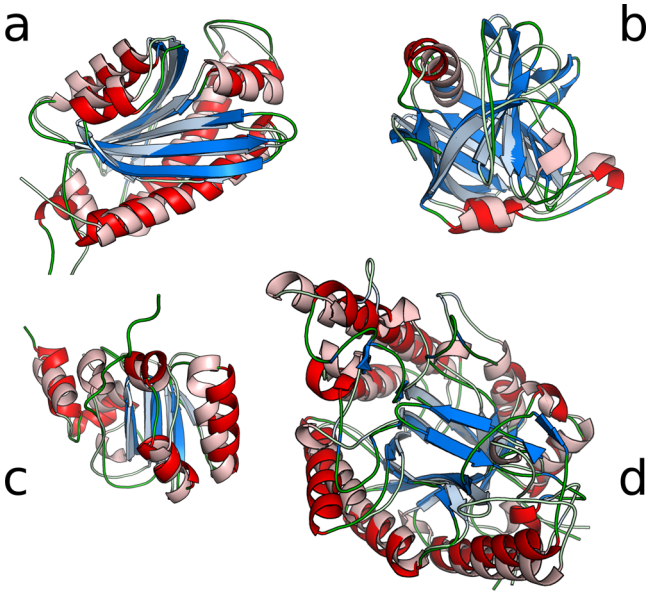


Figure 3. Superposition of the final PaLaCe structure after 10 ns simulation and the corresponding crystal structure for four proteins: (a) chemotaxis protein CheC, PDB code 1XKR (RMSD 2.5 Å); (b) cyclophilin, PDB code 1XO7, which shows one of the largest secondary structure content variations (31% increase in β); (c) the small YueI protein from *Bacillus subtilis*, PDB code 2OHW, which shows the largest variation in radius of gyration (11% increase); (d) xylanase B, PDB code 2DEP (RMSD 6.5 Å). Helices are colored red, strands blue, and loops green for the simulated structures, with paler colors for the crystal structures.

results visually for four of the proteins, via the superposition of the final structures of 10 ns simulations with the corresponding crystal structures. Structural variations are generally small (sometimes below 3 Å, as in Figure 3a) and typically involve minor rearrangements of loops and helices, most notably at flexible N- or C-termini.

Overall, the explicit description of hydrogen bonds in PaLaCe and an appropriate balance of its various energy terms contribute to well-preserved native folds and secondary structure content. Although there is a small tendency to increase β-strand content (particularly in dominantly β-fold proteins, see Figure 3b), one should note that using the same measures as in Figure 2b, DSSP assignments for α-helices and β-strands typically differ by 10% among the structures of a single NMR ensemble.[42]

Average differences in radius of gyration are between 2.7% and 3.1%, corresponding to differences of about 0.5 Å for a globular protein of 200 residues. These differences are very small, given the sensitivity of radii of gyration to conformational fluctuations in small proteins, and suggest that PaLaCe maintains good packing of the polypeptide chains and a correct balance between the internal and solvent forces acting on the structures. We remark in passing that a transition from a short helix to an extended conformation at the C-terminal of the small YueI protein from *Bacillus subtilis* (Figure 3c) is the main cause of the observed 11% increase in the radius of gyration. It is also worth noting that simulations of one of the largest proteins in test set 1, xylanase B from *Clostridium stercorarum* (340 aa, Figure 3d), preserve its β_8/α_8-barrel architecture very well. The apparently large RMSD with the experimental structure (6.5 Å) is mainly due to small changes in the relative orientation of some of the α-helices and some rearrangements of loops (which may be expected to be more flexible).

5.2. Trajectory Stability and Transferability. The results presented in the previous section illustrate the performance of PaLaCe in matching the native structural properties of a large set of proteins (98 including the training and test sets) of different sizes and architectures, on time scales up to 100 ns. Figure 4a shows the time course of the backbone RMSD for three of the proteins belonging to test set 2 and representative of the three main CATH classes: mainly α (Figure 4a), mainly β (Figure 4b), and α/β (Figure 4c), along with the superposition of the crystal and final simulated structures. The trajectories show occasional transitions between relatively long-lived conformational substates separated by about 1 Å RMSD. In each case, the final structures remain very close to the experimental conformations and show the same conformational features (conservation of secondary structure content and radius of gyration) described in the previous section.

5.3. Dynamic Properties. To test the dynamic fluctuations observed during LD simulations performed with PaLaCe, we have compared the root-mean-square fluctuations (RMSF) of the backbone atoms with crystallographic B-factors for a mainly-α protein and a large α/β protein (Figure 5). We have also compared RMSF values with those derived from concatenated equilibrium trajectories using atomistic molecular dynamics simulations in explicit solvent (ATMD)[37] for two homologous mainly-β proteins (Figure 6). Both comparisons show well-matched flexibility peaks (despite the fact that the B-factors plotted in Figure 5 can be affected by factors other than internal fluctuations).[43]

5.4. Response to Applied Force. The final test assesses the ability of PaLaCe to model the mechanical properties of proteins. We chose the immunoglobulin-like (Ig-like) domain of the giant protein titin, a key component of the contractile unit of muscle (the sarcomere) and the subject of a large body of both single-molecule experimental results and constrained all-atom molecular dynamics.[44] As first highlighted by single-molecule force spectroscopy studies,[45−48] the Ig-like domains of titin contribute to the passive stiffness of the sarcomere during muscle extension through sequential rupture of their secondary structure elements at a loading force of roughly 100 pN on a millisecond time scale. The atomistic details of Ig-like

dx.doi.org/10.1021/ct3007925 | J. Chem. Theory Comput. 2013, 9, 785−793

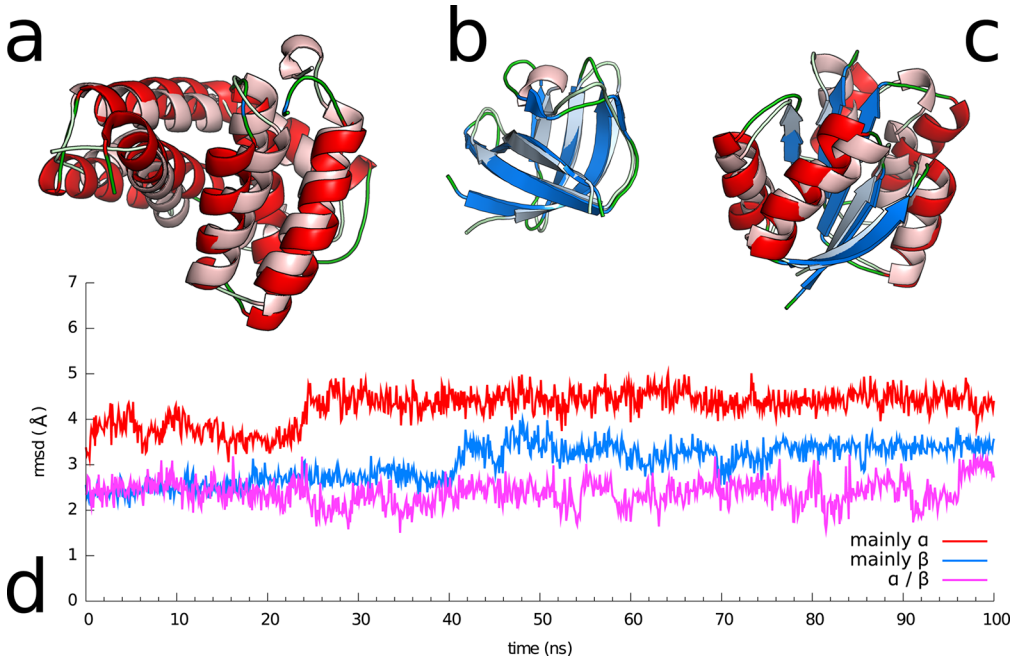


Figure 4. Trajectory stability. The backbone RMSD time course (d) is plotted for three proteins belonging to test set 2 and having mainly α, mainly β, or mixed α/β folds. The final structures from the simulations are superposed on the corresponding crystal structures: (a) antitermination factor NusB, PDB code 1TZV; (b) cold shock protein CspA, PDB code 1MJC; and (c) chemotaxis protein CheY, PDB code 1TMY. Helices are colored red, strands blue, and loops green for the simulated structures, with paler colors for the crystal structures.

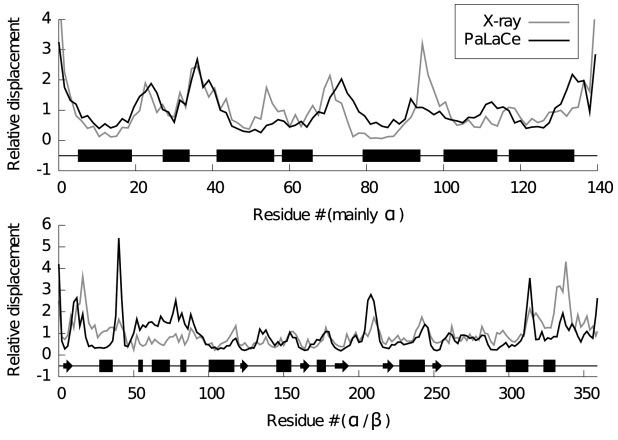


Figure 5. Comparison of PaLaCe backbone fluctuations with crystallographic B-factors. Backbone RMSF profiles, computed from the final 2.5 ns of 10 ns simulations, were converted to B-factors. The profiles for antitermination factor NusB, PDB code 1TZV (above), and histone deacetylase, PDB code 1C3P (below), have been scaled to yield the same mean displacements and therefore represent the relative values. The DSSP secondary structures derived from the crystallographic data are shown schematically as black boxes (α-helices) and arrows (β-strands).

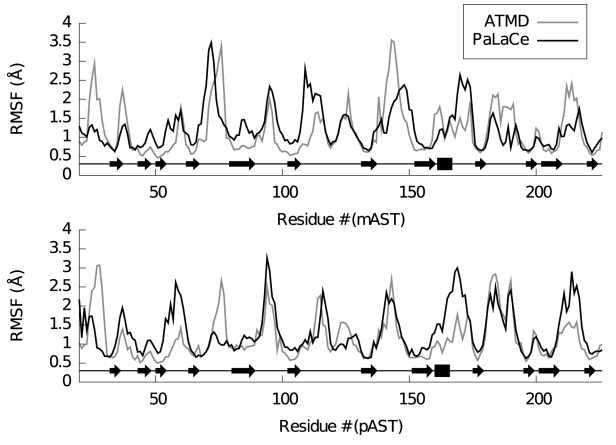


Figure 6. Backbone fluctuations for PaLaCe and for all-atom simulations. Cα RMSF profiles computed from the final 25 ns of 100 ns PaLaCe simulations for the two homologous trypsins: mesophilic mAST, PDB code 1A0J (above), and psychrophilic pAST, PDB code 2TBS (below). The all-atom profiles were computed from concatenated trajectories of 31.7 and 33.5 ns for the mesophilic (mAST) and psychrophilic (pAST) trypsins. The DSSP secondary structures derived from the crystallographic data are shown schematically as black boxes (α-helices) and arrows (β-strands).

domain unfolding have been unraveled using steered molecular dynamics (SMD), applying an external force along the N-to-C-terminal vector of a single immunoglobulin domain (I27).[49−52]

To test PaLaCe, we carried out coarse-grain, constant-velocity, steered Langevin dynamics (SLD) simulations[53] to induce I27 unfolding. Exploiting the relationship between pulling speed ν_s and rupture force $\overline{F}$ (the maximum force along the unfolding process),[54] we also estimate the inherent acceleration of the PaLaCe model due to the smoothing of the energy hypersurface linked with a coarse-grain representation and an implicit solvent model.

SLD simulations were run following Lee et al.[55] (see section S4 and Table S4 of the Supporting Information), using a range of pulling speeds encompassing more than 5 orders of magnitude, from 0.00005 to 7.0 Å/ps, with between one and 40 independent runs at each speed (resulting in a total of more than 250 independent simulations and almost 6 μs of trajectories; a single trajectory at the slowest pulling speed requiring 2 μs).

dx.doi.org/10.1021/ct3007925 | J. Chem. Theory Comput. 2013, 9, 785−793

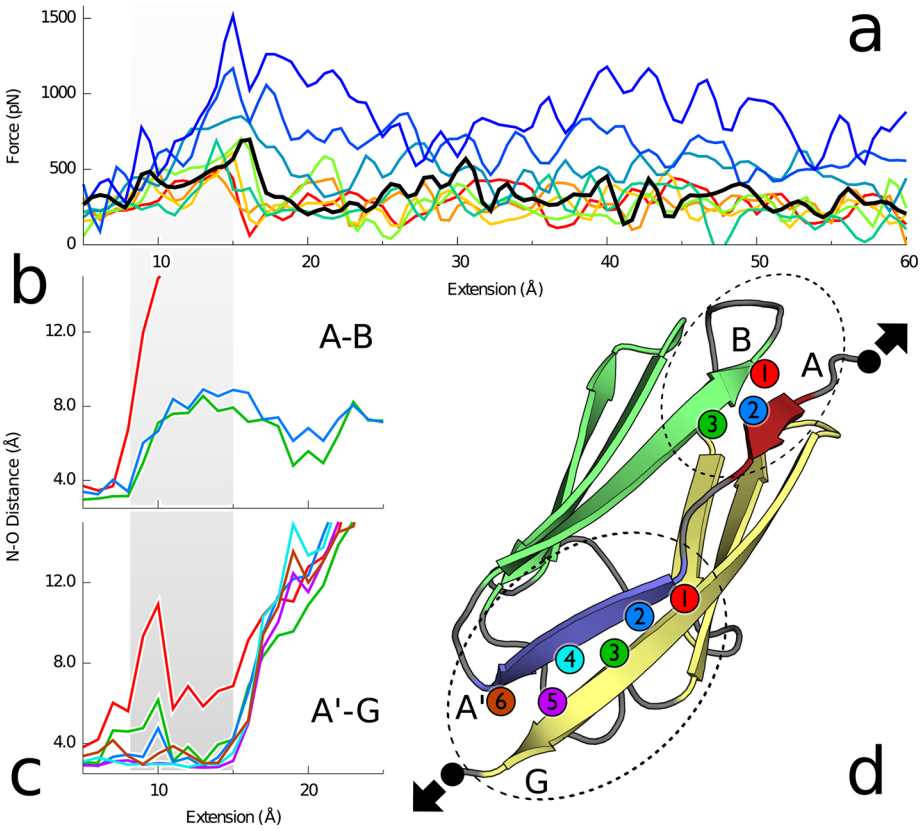


Figure 7. Mechanism of I27 forced unfolding. (a) Force−extension curves colored according to pulling speed, from red (ν_s = 0.00005 Å/ps) to blue (ν_s = 2.0 Å/ps); the thick black curve (ν_s = 0.1 Å/ps) indicates the simulation used to derive the data presented in the rest of the figure. (b and c) N−O distances for the three hydrogen bonds at the A−B interface (b) and the six hydrogen bonds at the A′−G interface (c); the gray shading highlights the partially unfolded intermediate. (d) Cartoon representation of the crystal structure of I27: the two β-sheets composing the sandwich are colored green (strands B to D) and yellow (strands E to G), while the A and A′ strands are colored dark red and blue, respectively; the colored dots correspond to the backbone hydrogen bonds whose N−O distances are plotted in b and c (see Supporting Information Table S5 for details), and the black dots represent the terminal residues.

The resulting force−extension profiles are shown in Figure 7a. The profiles indicate that the force reaches a maximum at an end-to-end extension of ≈15 Å independently of ν_s, in accord with the distances observed in all-atom SMD simulations. The β-sandwich structure of I27 (Figure 7d) is composed of two β-sheets, each composed of three antiparallel strands. Two further strands (A and A′), which both belong to the N-terminal segment of the protein, lock the structure together by linking the two sheets. The A strand establishes three hydrogen bonds with the B strand of one β-sheet (forming the A−B interface), while the A′ strand establishes six hydrogen bonds with the G strand of the other β-sheet (the A′−G interface; see Supporting Information Table S5). An analysis of the PaLaCe unfolding trajectories confirms that the main rupture event (responsible for the peak force) is identical in all SLD simulations and involves the nearly synchronous cleavage of the six hydrogen bonds between the A′ and G strands (Figure 7c). Furthermore, as also seen in the all-atom simulations, the rupture event is preceded by a so-called "pre-burst" phase, involving breaking the three hydrogen bonds at the A−B interface (Figure 7b). This leads to the formation of a relatively stable unfolding intermediate, whose existence has been confirmed experimentally.[56]

The PaLaCe results as a function of pulling speed ν_s can be compared more quantitatively with both experimental and simulation data using the statistical mechanics model developed by Dudko, Hummer, and Szabo (DHS).[57−59] As shown in Figure 8, this model can be used to place the slow-pulling AFM experimental data and the fast-pulling all-atom simulation data on a single $\overline{F}$ versus ν_s curve.[55] If we add the corresponding data obtained using PaLaCe to the same plot, we see that, although it has the same functional form as the all-atom data, it

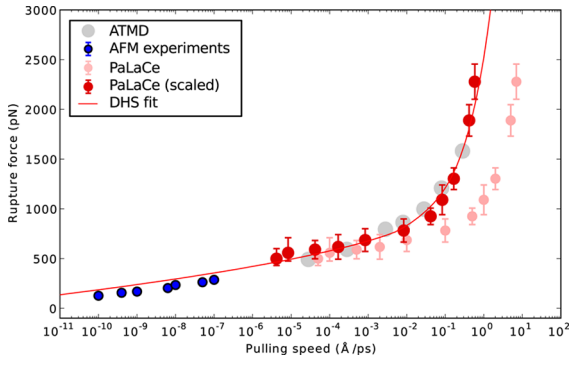


Figure 8. Matching PaLaCe simulations of I27 stretching with experimental values and with all-atom simulations. The pulling speed dependence of the rupture force derived from AFM experiments[45] and from all-atom SMD simulations (ATMD) can be grouped together[55] using the DHS model[57] (red line). The data from the PaLaCe SLD simulations (pink circles) fit this curve well after scaling the pulling speed by a factor of 12 (red circles). Error bars show the standard deviations for multiple PaLaCe trajectories at a given pulling speed.

dx.doi.org/10.1021/ct3007925 | J. Chem. Theory Comput. 2013, 9, 785−793

is shifted right along the pulling speed axis. This corresponds to time effectively running faster for the coarse-grain simulations. This effect is linked to a smoothing of the energy hypersurface and has been observed with other coarse-grain models, although it has generally been difficult to make a quantitative analysis of this effect.[60] The data in Figure 8 however allow us to obtain a quantitative estimate of the speedup obtained with PaLaCe. Fitting the PaLaCe data to the DHS curve using a least-squares procedure requires dividing the PaLaCe pulling speeds by 12, implying that PaLaCe simulation time is 12 times faster than physical time for this system.

6. CONCLUDING REMARKS

The first version of PaLaCe presented here satisfies our three design criteria: (1) We have conserved the residue-dependent level of coarse-graining successfully used in our earlier studies of protein mechanics. (2) We have avoid adding any artificial restraints on secondary structure by introducing an atomic-scale backbone which also allows for explicit backbone hydrogen bonding and reduces torsional coupling. (3) We have developed an automated, iterative procedure for parametrizing the force field on the basis of experimental structural data.

While there is certainly room for improvement, the current version of PaLaCe is capable of maintaining the structure of folded proteins during stochastic dynamic simulations of up to 100 ns, yields backbone fluctuations in line with experimental values and all-atom simulations, and has a force field that is applicable to a wide range of protein families. In line with our initial goal, we have shown that PaLaCe is capable of reproducing the mechanical properties of an immunoglobin domain subjected to stretching, yielding results in accord with all-atom molecular dynamics simulation and single molecule experiments. The latter study also gave us the opportunity to quantify the time acceleration associated with the PaLaCe coarse-grain representation, which, in this case, turns out to be a factor of 12.

PaLaCe is currently implemented in a "test bed" version in MMTK. In its current version it is faster than equivalent all-atom simulations in a vacuum by 50 times and faster than explicit solvent simulations by roughly 1000 times. It will shortly be ported to other molecular dynamics packages and be made freely available to the community.

■ ASSOCIATED CONTENT

Supporting Information

We provide details of the amino acid dependence of the force field parameters, the definition of the reference states, computational methods used for the SLD simulations, and complete lists of the protein data sets and the current PaLaCe force field parameters. This information is available free of charge via the Internet at http://pubs.acs.org/.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: richard.lavery@ibcp.fr.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank the Rhône-Alpes action CIBLE for a grant supporting N.C. and the French supercomputer CINES for a generous allocation of computer time. M.P. acknowledges funding from the ANR Blanc projet EXPENANTIO and thanks E. Papaleo and L. De Gioia for funding an initial research visit to Lyon during which the work on PaLaCe began.

■ REFERENCES

(1) Bahar, I.; Rader, A. J. *Curr. Opin. Struct. Biol.* **2005**, *15*, 586−92.
(2) Eyal, E.; Dutta, A.; Bahar, I. *WIRES:. Comput. Mol. Sci.* **2011**, *1*, 426−439.
(3) Sacquin-Mora, S.; Lavery, R. *Biophys. J.* **2006**, *90*, 2706−17.
(4) Sacquin-Mora, S.; Laforet, E.; Lavery, R. *Proteins* **2007**, *67*, 350−9.
(5) Sacquin-Mora, S.; Sebban, P.; Derrien, V.; Frick, B.; Lavery, R.; Alba-Simionesco, C. *Biochemistry* **2007**, *46*, 14960−8.
(6) Izvekov, S.; Voth, J. A. *J. Phys. Chem. B.* **2005**, *109*, 2469.
(7) Hills, R. D.; Lu, L.; Voth, G. A. *PLoS Comput. Biol.* **2010**, *6*, e1000827.
(8) Zacharias, M. *Protein Sci.* **2003**, *12*, 1271−82.
(9) Monticelli, L.; Kandasamy, S. K.; Periole, X.; Larson, R. G.; Tieleman, D. P.; Marrink, S. J. *J. Chem. Theory Comput.* **2008**, *4*, 819−834.
(10) Simons, K. T.; Kooperberg, C.; Huang, E.; Baker, D. *J. Mol. Biol.* **1997**, *268*, 209−25.
(11) Voegler Smith, A.; Hall, C. K. *Proteins* **2001**, *44*, 344−60.
(12) Maupetit, J.; Tuffery, P.; Derreumaux, P. *Proteins* **2007**, *69*, 394−408.
(13) Bereau, T.; Deserno, M. *J. Chem. Phys.* **2009**, *130*, 235106.
(14) Tozzini, V.; McCammon, J. A. *Chem. Phys. Lett.*. **2005**, *413*, 123−128.
(15) Tozzini, V.; Rocchia, W.; McCammon, J. A. *J. Chem. Theory Comput.* **2006**, *2*, 667−673.
(16) Ceres, N.; Lavery, R. In *Innovations in Biomolecular Modeling and Simulations*; Schlick, T., Ed.; Royal Society of Chemistry: London, 2012; pp 219−248.
(17) Hinsen, K. *J. Comput. Chem.* **2000**, *21*, 79−85.
(18) Levitt, M. *J. Mol. Biol.* **1976**, *104*, 59−107.
(19) Ołdziej, S.; Kozłowska, U.; Liwo, A.; Scheraga, H. A. *J. Phys. Chem. A* **2003**, *107*, 8035−8046.
(20) Májek, P.; Elber, R. *Proteins* **2009**, *76*, 822−36.
(21) Hingerty, B.; Richie, R. H.; Ferrel, T. L.; Turner, J. E. *Biopolymers* **1985**, *24*, 427−439.
(22) Lavery, R.; Zakrzewska, K.; Sklenar, H. *Comput. Phys. Commun.* **1995**, *91*, 135−158.
(23) Mehler, E. L.; Solmajer, T. *Protein Eng.* **1991**, *4*, 903−910.
(24) Mezei, M. *J. Mol. Graphics Modell.* **2003**, *21*, 463−72.
(25) Berman, H. M.; Westbrook, J.; Feng, Z.; Iype, L.; Schneider, B.; Zardecki, C. *Acta Crystallogr., Sect. D* **2002**, *58*, 889−898.
(26) Thomas, P. D.; Dill, K. A. *J. Mol. Biol.* **1996**, *257*, 457−69.
(27) Ben-Naim, A. *J. Chem. Phys.* **1997**, *107*, 3698.
(28) Wang, T.; Wade, R. C. *Proteins* **2003**, *50*, 158−169.
(29) Tusnády, G. E.; Dosztányi, Z.; Simon, I. *Bioinformatics* **2004**, *20*, 2964−72.
(30) Narang, P.; Bhushan, K.; Bose, S.; Jayaram, B. *Phys. Chem. Chem. Phys.* **2005**, *7*, 2364.
(31) Kabsch, W.; Sander, C. *Biopolymers* **1983**, *22*, 2577−2637.
(32) Thomas, P. D.; Dill, K. A. *Proc. Natl. Acad. Sci. U. S. A.* **1996**, *93*, 11628.
(33) Reith, D.; Pütz, M.; Müller-Plathe, F. *J. Comput. Chem.* **2003**, *24*, 1624−36.
(34) Huang, S. Y.; Zou, X. *Proteins* **2008**, *72*, 557−79.
(35) Song, Y.; Tyka, M.; Leaver-Fay, A.; Thompson, J.; Baker, D. *Proteins* **2011**, *79*, 1898−909.
(36) Orengo, C. A.; Bray, J. E.; Buchan, D. W.; Harrison, A.; Lee, D.; Pearl, F. M.; Sillitoe, I.; Todd, A. E.; Thornton, J. M. *Proteomics* **2002**, *2*, 11−21.
(37) Papaleo, E.; Pasi, M.; Riccardi, L.; Sambi, I.; Fantucci, P.; De Gioia, L. *FEBS Lett.* **2008**.
(38) Smalås, A. O.; Heimstad, E. S.; Hordvik, A.; Willassen, N. P.; Male, R. *Proteins: Struct., Funct., Bioinf.* **1994**, *20*, 149−166.

(39) Schroder, H. K.; Willassen, N. P.; Smalas, A. O. *Acta Crystallogr., Sect. D* **1998**, *54*, 780−798.
(40) Zhang, Y.; Xi, Z.; Hegde, R. S.; Shakked, Z.; Crothers, D. M. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101*, 8337−41.
(41) Xu, D.; Williamson, M. J.; Walker, R. C. *Ann. Rep. Comput. Chem.* **2010**, *6*, 1.
(42) Andersen, C. A. F.; Palmer, A. G.; Brunak, S.; Rost, B. *Structure* **2002**, *10*, 175−184.
(43) Hunenberger, P. H.; Mark, A. E.; van Gunsteren, W. F. *Proteins* **1995**, *21*, 196−213.
(44) Tskhovrebova, L.; Trinick, J. *Nat. Rev. Mol. Cell. Biol.* **2003**, *4*, 679−89.
(45) Rief, M.; Gautel, M.; Oesterhelt, F.; Fernandez, J. M.; Gaub, H. E. *Science* **1997**, *276*, 1109−12.
(46) Kellermayer, M. S.; Smith, S. B.; Granzier, H. L.; Bustamante, C. *Science* **1997**, *276*, 1112−6.
(47) Carrion-Vazquez, M.; Oberhauser, A. F.; Fowler, S. B.; Marszalek, P. E.; Broedel, S. E.; Clarke, J.; Fernandez, J. M. *Proc. Natl. Acad. Sci. U. S. A.* **1999**, *96*, 3694.
(48) Grützner, A.; Garcia-Manyes, S.; Kötter, S.; Badilla, C. L.; Fernandez, J. M.; Linke, W. A. *Biophys. J.* **2009**, *97*, 825−34.
(49) Lu, H.; Isralewitz, B.; Krammer, A.; Vogel, V.; Schulten, K. *Biophys. J.* **1998**, *75*, 662−671.
(50) Lu, X. J.; Babcock, M. S.; Olson, W. K. *J. Biomol. Struct. Dyn.* **1999**, *16*, 833−43.
(51) Lu, H.; Krammer, A.; Isralewitz, B.; Vogel, V.; and Schulten, K. In *Elastic Filaments of the Cell*; Granzier, H. L., Pollack, G. H., Eds.; Kluwer: New York, 2000; pp 143−162.
(52) Fowler, S. B.; Best, R. B.; Toca Herrera, J. L.; Rutherford, T. J.; Steward, A.; Paci, E.; Karplus, M.; Clarke, J. *J. Mol. Biol.* **2002**, *322*, 841−849.
(53) Grubmüller, H.; Heymann, B.; Tavan, P. *Science* **1996**, *271*, 997.
(54) Evans, E.; Ritchie, K. *Biophys. J.* **1997**, *72*, 1541−55.
(55) Lee, E. H.; Hsin, J.; Sotomayor, M.; Comellas, G.; Schulten, K. *Structure* **2009**, *17*, 1295−306.
(56) Fowler, S. B.; Best, R. B.; Toca Herrera, J. L.; Rutherford, T. J.; Steward, A.; Paci, E.; Karplus, M.; Clarke, J. *J. Mol. Biol.* **2002**, *322*, 841−849.
(57) Hummer, G.; Szabo, A. *Biophys. J.* **2003**, *85*, 5−15.
(58) Dudko, O. K.; Hummer, G.; Szabo, A. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 15755.
(59) Hummer, G.; Szabo, A. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 21441−6.
(60) Schäfer, L. V.; de Jong, D. H.; Holt, A.; Rzepiela, A. J.; de Vries, A. H.; Poolman, B.; Killian, J. A.; Marrink, S. J. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 1343−8.