Coarse-Grained Representation of β-Helical Protein Building Blocks

David Curcó,*,† Ruth Nussinov,‡,§ and Carlos Alemán*,∥

Departament d'Enginyeria Química, Facultat de Química, Universitat de Barcelona, Martí i Franquees 1, Barcelona E-08028, Spain, Basic Research Program, SAIC-Frederick, Inc. Center for Cancer Research Nanobiology Program, NCI, Frederick, Maryland 21702, Department of Human Genetics Sackler, Medical School, Tel Aviv University, Tel Aviv 69978, Israel, and Departament d'Enginyeria Química, E. T. S. d'Enginyeria Industrial de Barcelona, Universitat Politècnica de Catalunya, Diagonal 647, Barcelona E-08028, Spain

Received: April 11, 2007; In Final Form: June 25, 2007

A general strategy to develop coarse-grained models of β-helical protein fragments is presented. The procedure has been applied to a building block formed by a two-turn repeat motif from *E. coli* galactoside acetyltransferase, which is able to provide a very stable self-assembled tubular nanoconstruct upon stacking of its replicas. For this purpose, first, we have developed a computational scheme to sample very efficiently the configurational space of the building block. This method, which is inspired by a strategy recently designed to study amorphous polymers and by an advanced Monte Carlo algorithm, provides a large ensemble of uncorrelated configurations at a very reasonable computational cost. The atomistic configurations provided by this method have been used to obtain a coarse-grained model that describes the amino acids with fewer particles than those required for full atomistic detail, i.e., two, three, or four depending on the chemical nature of the amino acid. Coarse-grained potentials have been developed considering the following types of interactions: (i) electrostatic and van der Waals interactions between residues i and $i + n$ with $n \geq 2$; (ii) interactions between residues i and $i + 1$; and (c) intra-residue interactions. The reliability of the proposed model has been tested by comparing the atomistic and coarse-grained energies calculated for a large number of independent configurations of the β-helical building block.

Introduction

The successful design of novel nanostructures using complex self-assembled building blocks extracted from biological macromolecules, i.e., DNA, RNA, peptides, and proteins, provides evidence for the tremendous progress in nanotechnology and nanobiology.[1−11] Thus, now the potential biomedical applications of the designed self-assembled nanostructures, such as antiviral drug carriers, scaffolds for tissue repair, and spectroscopic labels for therapeutic and photonic devices, are also becoming an important topic of research.[12−14] Accordingly, design and construction of nanostructures using natural building blocks appears to be a good strategy to modulate and control the supramolecular assemblies toward the desired goal.

Within this context, we have been working on the construction of nanostructures with different shapes using naturally occurring building blocks.[15−17] In particular, we are interested in assemblies derived from protein building blocks,[18−20] which are structural units (protein fragments) that retain a conformation similar to the one they have when embedded in native protein structures.[21] The protein database is populated by an extensive repertoire of building blocks with different shapes, sizes, and chemical properties which can be used in design.[17] The challenge is to manipulate such pre-existing protein foldamers toward the formation of the desired stable nanostructures through a favorable association process.

* Corresponding authors. E-mail: curco@angel.qui.ub.es (D.C.); carlos.aleman@upc.es (C.A.).
† Universitat de Barcelona.
‡ NCI.
§ Tel Aviv University.
∥ Universitat Politècnica de Catalunya.

We recently used this strategy to construct stable nanotubular structures from monomeric naturally occurring β-helical building blocks.[20] For this purpose, we selected 17 building blocks from native left-handed β-helical proteins by slicing β-helices into two-turn repeat units. Four copies of each structural unit were stacked one atop of the other, i.e., with no covalent linkage between them, using computer simulation methods. The stability of these self-assembled tube organizations was investigated through short (20−40 ns) atomistic molecular dynamics (MD) simulations. Constructs able to preserve their organization in the simulation are candidates for experiments. We observed that a structural model based on the self-assembly of a two-turn repeat motif from *E. coli* galactoside acetyltransferase (PDB code 1krr, chain A) produced a very stable nanotube. Furthermore, in order to enhance the stability of this nanoconstruct, synthetic amino acids with conformational tendencies restrained to those of natural amino acids in the most mobile loop regions have been engineered.[22,23] We found that the thermodynamic stability of the β-helical repeat sequence increases when selective substitution of natural residues by synthetic ones eliminates unfavorable electrostatic interactions.

Conventional atomistic computer simulations, in which the detailed chemistry of the systems was kept, have been found to be essential for designing nanotubular constructs. Thus, evaluation of both explicit atom pair interactions and analysis of small structural details was required for the right selection of the most adequate monomeric β-helical building blocks. However, only very short time and length scales can be reached throughout the simulation of atomistic models owing to limitations of computer power. Thus, although the computational power

10.1021/jp072832q CCC: $37.00
Published on Web 08/11/2007

increases 10-fold every 4−5 years, the huge number of degrees of freedom characteristic of atomistic models limits brute force approaches when an investigation of phenomena at the mesoscopic scale is desired. For instance, although structural phenomena on the ∼100 Å scale are of interest, conventional atomistic simulations of nanotubes consisting of more than six stacked β-helical building blocks are now practically impossible. Furthermore, full atomistic detail MD simulations to observe the mesoscopic properties and phenomena of the nanoconstructs (for example, the persistence length and stiffness, the diffusion of small molecules inside the tube, and the mechanistic aspects of the application of nanotubes as drug delivery systems) are completely unfeasible, i.e., on microsecond and millisecond time scales. The detailed treatment of the fast vibration modes would slow down the run time so strongly that the slow modes cannot reach equilibrium. In addition, the atomistic details sometimes obscure the interesting properties. Therefore, new developments that enable an expansion of the time and length scales must be considered.

One way to circumvent these problems is to reduce the degrees of freedom by coarsening the models and keeping only those degrees of freedom that are deemed important for the particular range of interest. In the coarse-graining approach, the detailed chemistry enters only in the derivation of the potential between new interacting particles, hereafter denoted as blobs. In this work, we initiate a project which addresses the problem of representing complex supramolecular nanostructures, such as nanotubes formed by stacked β-helical building blocks, on the mesoscopic metric. Specifically, in the present study, we report a general strategy to develop coarse-grained (CG) models for the simplest constituent of these nanotubular structures: the β-helical building block. The resulting blob-based model, in which each blob represents a group of atoms, is roughly 2 orders of magnitude faster than atomistic models but still contains the essential physics of the system under study. It should be emphasized that, although the procedure has been applied to the building block extracted from 1krr, the strategy can be easily extended to other building blocks involved not only in nanotubes but also in other nanostructures.

This paper is organized as follows. First, the advantages and limitations of the coarse-graining strategy used in this work are briefly outlined. After this, details about the 1krr building block are provided. Next, a very efficient procedure that has been developed to sample the configurational space of the building block at the atomistic level is described. Subsequently, the CG model developed for the 11 amino acids contained in the 1krr building block is presented in detail. Finally, the reliability of the model provided in this work has been confirmed by comparing the atomistic and CG energies calculated for a large set of configurations that were not used in the development of the CG model.

Outlining the Coarse-Graining Strategy: Advantages and Limitations

CG models retain only as much unique and relevant information as needed about the specific system under investigation, using significantly fewer particles than required for full atomistic detail. The coarsened degrees of freedom must be constrained to ensembles of configurations that represent an appropriate average over the microscopic atomic-scale potential energy surface for the fully resolved system. Coarse-graining aims to guarantee that the global conformational relaxation, which takes orders of magnitude more in time than any local move, is achieved. This means that the characteristic time step in a MD simulation increases significantly, i.e., fast degrees of freedom are eliminated, and the short spatial length scale is upscaled. Tremendous effort and progress have been achieved toward developing CG models of melted and amorphous synthetic polymers, membranes, and micelles,[24−36] i.e., systems consisting of a large number of macromolecular chains. Furthermore, a number of schemes based on CG models have also been reported during the past decade to study protein folding.[37−47] It should be noted that, independently of the synthetic or biological nature of the macromolecular systems, the CG models are only valid for the thermodynamical conditions used in their development; i.e., the detailed chemistry of the atomistic models generated under given conditions is reflected in the empirical CG potentials. Thus, it has been demonstrated that atomistic models generated considering different thermodynamical conditions led to very different CG models and potentials.[30]

The coarse-graining strategy proposed in this work is based on the use of atomistic models that describe satisfactorily the configurational space of the building block. In order to reach such generated structures, a very efficient sampling procedure has been developed. This method, which is inspired by a strategy we have recently developed to sample the configurational space of amorphous polymers[48] combined with an advanced Monte Carlo (MC) algorithm,[49] has been applied to generate a large ensemble of uncorrelated configurations for the 1krr building block. It should be noted that the reliability of the CG potentials increases with the size of the ensemble, i.e., with the quality of the sampling.

The generated atomistic microstructures have been used to develop a CG model for the 11 amino acids contained in the 1krr building block. In this model, the amino acids have been represented by a small set of blobs (two, three, or four blobs depending on the chemical nature of the amino acid) rather than by explicit atoms. The number and position of these blobs have been selected by looking to the best fitting between the atomistic and CG energies. On the other hand, the CG potentials for the 1krr building block have been obtained by dividing the interactions into types of contributions, electrostatic and van der Waals, between residues i and $i + n$ with $n \geq 2$; between residues i and $i + 1$; and intra-residue interactions. These interactions have been treated independently, and the procedures and strategies followed to derive the corresponding CG potentials differ accordingly. Nonbonding interactions between residues i and $i + n$ with $n \geq 2$ have been found to depend strongly on both the chemical nature of the amino acids and the separation between pairs of amino acids (n). Therefore, specific electrostatic and van der Waals CG potentials were explicitly developed for each of the 561 different pairs of residues found in the 1krr building block. Inter-residue interactions between consecutive residues have been described using geometrical variables (distances, angles, and dihedrals) that include the spatial orientation of the amino acids. In spite of the fact that the distance between two consecutive residues is very small, which is usually a handicap for coarse-graining developments, the results provided by this strategy are excellent. Intra-residue potentials, which also involve geometrical variables, have provided the largest errors. However, this energy contribution is the least important not only for the stability of the β-helical building block but also for the nanotubular structure formed by stacked building blocks, which is our final objective.

The 1krr β-Helical Building Block

In a previous work,[20] left-handed β-helices were selected for nanotubular structural design according to the following crite-

TABLE 1: Sequences of the Wild Type 1krr Building Block Used to Construct the Self-Assembled Nanotubular Constructs

PDB	protein name	residues	sequence
1krr	galactoside acetyltransferase from *E. coli*	131−165	PITIGNNVWIGSHVVINPGVTIGDNSVIGAGSIVT

ria: (i) they contain highly repetitive, symmetrical building blocks allowing formation of nanofibers without performing many structural manipulations;[50] (ii) they mostly occur in or near active or binding sites, and are thus likely to retain the functional importance of β-helical proteins;[51−53] and (iii) as compared with right-handed β-helices, left-handed β-helices exhibit limited variability in shape, size, and sequence.[54] Left-handed β-helices display an equilateral triangular shape and highly repetitive sequence, while right-handed β-helices are less regular. We found that a nanotube constructed of four stacked replicas of the left-handed β-helix formed by residues 131−135 of 1krr exhibited remarkable stability under various simulated conditions, including temperature increase and addition of ions.[20] Thus, nanoconstructs formed by this repeat are good systems to test the new procedure for developing CG potentials. A description of the sequence used to create the model based on 1krr is provided in Table 1. This sequence contains a repetitive helical strand−loop motif, where the peptide backbone alternates between β-strands and loops. The X-ray crystal structure of 1krr shows that this protein fragment has an almost perfect equilateral triangular shape, with each side being ∼18 Å.

Sampling the Configurational Space at the Atomistic Level

Generation of a complete ensemble of uncorrelated atomistic configurations using conventional simulation procedures is inefficient due to the long relaxation times characteristic of large molecular systems, which involve a huge amount of computational power. Recently, we addressed this problem in studies of synthetic polymers in the amorphous state.[48,55,56] Thus, we developed a computational strategy based on a random search of energy minima, which is able to provide a large number of representative and relaxed independent atomistic configurations by combining a powerful generation algorithm[48,55] and a very efficient minimization strategy based on advanced MC methods.[56] This method has been used as a starting point to design a new procedure for sampling the configurational space of the 1krr building block. The new procedure is also organized in a two-step strategy: generation and relaxation.

The internal coordinates needed by the generation algorithm are taken from a configuration previously equilibrated by conventional MD, which is used as the starting structure. The main parts of the generation algorithm can be summarized as follows:

(1) New values are provided to the backbone dihedral angles by introducing random distortions to those values extracted from the starting structure. Thus, a value between $\kappa_0 + \Delta$ and $\kappa_0 - \Delta$ is randomly assigned to each dihedral angle, where κ_0 is the value of such dihedral in the starting equilibrated configuration and Δ is a value fixed by the user. After several trials, we found that suitable configurations for the development of CG models are generated when the value of Δ is 15° for the flexible dihedral angles $\{\varphi,\psi\}$ and the peptide bonds $\{\omega\}$ are not allowed to distort more than 10° with respect to the ideal value (180°); that is, for peptides bonds, Δ is a variable that depends on the starting structure.

(2) The positions of all of the backbone atoms are generated using the geometric parameters, i.e., bond lengths and angles, extracted from the starting structure and the dihedral angles assigned in the previous step. The coordinates of all of the atoms directly attached to those contained in the backbone are then generated considering that their positions are automatically defined by the chirality of the C^{α}, the hybridization of the backbone atoms, and the geometric parameters extracted from the starting structure. Overlaps between atoms separated by more than three chemical bonds are examined using the corresponding van der Waals radii, which have been taken from libraries of the Amber force field.[57] If atomic overlaps are identified at atom *i*, the backbone is rebuilt by assigning new dihedral angles (see step 1) from atom *i*-*ip*, where *ip* increases with the number of failures.

(3) The side chains are generated atom by atom and residue by residue using randomly generated dihedral angles. If there is no overlap with atoms separated by more than three chemical bonds, the dihedral is accepted; otherwise, the whole side chain of the residue is reconstructed. The backbone conformation is rejected if atomic overlaps remain after several trials, i.e., typically eight, and the whole process starts again from step 1. In all cases, the generated positions for the side chain atoms fulfill the internal geometry restrictions, i.e., bond distances and angles extracted from the starting structure. The planarity and cyclic geometry of the aromatic rings contained in the side chains of some amino acids were maintained in all cases. Thus, after generating the positions of the first three atoms of the ring, the coordinates of the remaining atoms were automatically generated using the bond distances, bond angles, and dihedral angles provided by the starting structure.

The generated structures are relaxed by applying the configurational bias (CB) MC algorithm.[49] This method has been exclusively applied to the side chains, with the backbone not affected by the relaxation. Thus, the backbone, which was generated by introducing small distortions in the starting structure, must belong to the ensemble of configurations that describes the desired building block. Furthermore, as the value of Δ is relatively small (see above), the energy penalty associated with such distortions is expected to be low. Accordingly, it is highly desirable to adapt the side chain conformations to the backbone that describes such a building block by relaxing unfavorable interactions. The CB-MC method, whose efficacy and reliability has been extensively described in the literature,[49,58−60] can be briefly summarized as follows: (i) a side chain is selected at random; (ii) the side chain is cut at a random position; and (iii) the side chain is sequentially regrown bond by bond by examining a number of randomly chosen positions (M), the new position being chosen with a weight proportional to its normalized Boltzmann factor. In this work, M was fixed at 20 in all calculations. It is well-known that the efficiency of CB-MC decays rapidly when applied to molecular segments with pending groups.[59,60] However, in this case, a remarkable efficiency has been achieved because, as mentioned above, the backbone conformation has not been included in the relaxation process.

In order to develop CG potentials to describe the 1krr building block, 3×10^4 atomistic configurations have been generated and relaxed for a single building block using the procedure described above and considering 15 different starting structures, i.e., 2000 configurations per starting structure. Series of different consecutive rounds of short MD runs were performed to obtain the starting structures. Thus, after minimizing the potential

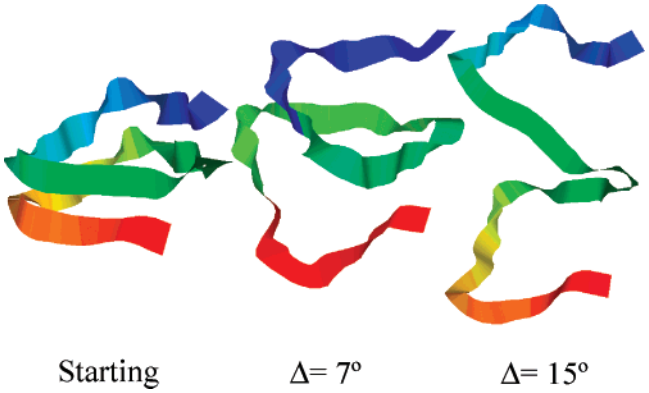

Figure 1. Starting structure (left) of the 1krr building block, which was obtained by MD, and representative atomistic configurations derived from the generation−relaxation algorithm proposed in this work (see text) when the flexible dihedrals are distorted, Δ = 7.0° (middle) and 15.0° (right).

energy of the system and thermal relaxation of the solvent molecules, the heating and equilibration short runs consisted of κ·100 ps of steady heating until the target temperature was reached (298 K), μ·300 ps of *NVT*-MD at 298 K (thermal equilibration) followed by μ·300 ps of density relaxation (*NPT*-MD). Fifteen sets of $\{\kappa,\mu\}$ values, which are random numbers between 0.50 and 1.00, were used to provide the 15 starting structures. The simulated system was formed by the 1krr building block placed in the center of an orthorhombic simulation box filled with explicit water molecules, which were represented using the TIP3 model.[61] A positively charged sodium atom was added to the simulation box in the required amount to reach electric neutrality (the 1krr building block has a negative net charge at neutral pH). The energy was calculated by using the Amber force field[57] with the required parameters taken from the Amber libraries. Atom pair distance cutoffs were applied at 14 Å to compute nonbonding interactions. Both temperature and pressure were controlled by the weak coupling method, the Berendsen thermo-barostat[62] using a time constant for heat bath coupling and a pressure relaxation time of 1 ps. Bond lengths were constrained using the SHAKE algorithm,[63] with numerical integration steps of 2 fs. All MD simulations were performed using the NAMD program.[64]

Figure 1 compares two configurations generated and relaxed using the procedure discussed above with the corresponding starting structure. As can be seen, the configurations produced by applying Δ values of 7° and, especially, 15° are significantly distorted with respect to the starting structure, which presents a well-defined β-helical conformation. On the other hand, analysis of the different energy terms calculated for the 3×10^4 relaxed configurations reveals interesting features. Figure 2 represents the total, 1−4, van der Waals and electrostatic energy terms computed for each relaxed configuration against the value of Δ used for its generation. As expected, the total energy increases with Δ, indicating that the destabilization of the β-helix increases with the degree of distortion (Figure 2a). The loss of stability is mainly due to nonbonding energy contributions, which become less attractive when Δ grows. However, the tendencies displayed by the van der Waals (Figure 2c) and electrostatic (Figure 2d) energies are clearly different. Thus, the van der Waals contribution presents an almost linear tendency; i.e., the energy becomes more repulsive for the whole range of Δ values, whereas the electrostatic energy grows when Δ increases from 0 to 4.5° but becomes almost constant for higher values of Δ. Inspection of the 1−4 energies (Figure 2b), which include all of the contributions associated with the interactions between atoms separated by three bonds (torsional, electrostatic, and van der Waals), reveals that, in this case, the distortion introduces a stabilizing effect. The CG model developed in the next sections should capture the complex behavior of all of these short- and long-range interactions.

It should be noted that the role of the solvent is neglected in this procedure. Thus, explicit solvent molecules were used in atomistic MD simulations to investigate the stability of both the β-helix building blocks and the self-assembled nanoconstructs. However, atomistic solvent molecules are not compatible with the CG model of the building block, since the purpose of the latter is to increase significantly the time and length scales; i.e., fast degrees of freedom and short spatial phenomena are upscaled in CG models. In this work, atomistic configurations

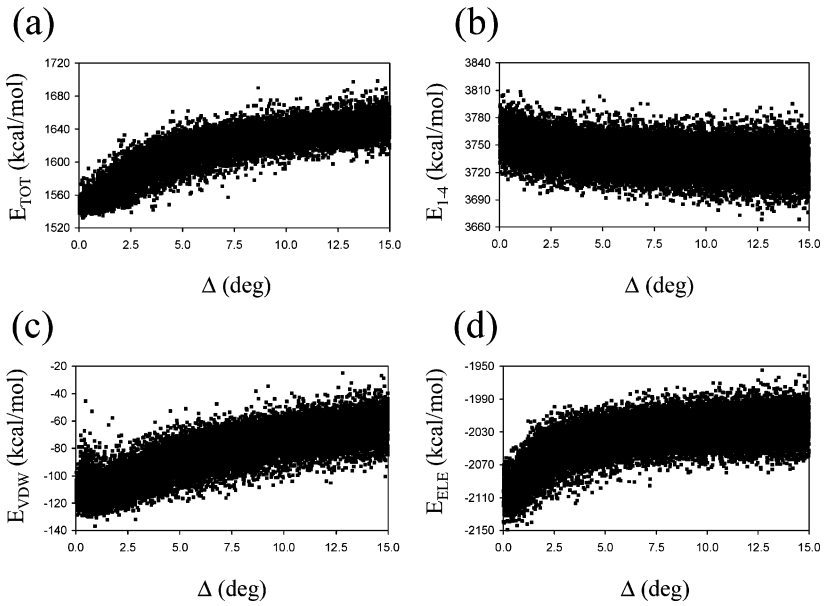

Figure 2. Graphical representation of the total (a), 1−4 (b), van der Waals (c), and electrostatic (d) energies (in kcal/mol) against Δ for the 3×10^4 atomistic configurations provided for the 1krr building block by the generation−relaxation sampling procedure proposed in this work. The influence of the different energy contributions of the distortion degree on the flexible dihedral angles during the generation of the configurations is reflected.

of the building block were generated without considering the solvent, even though starting structures were taken from simulations in aqueous solution. However, it should be mentioned that we plan to develop a coarse-grained model of the solvent in a near future.

Definition of the Coarse-Graining Model

The goal of the present work is to develop a generic CG model for the β-helical building blocks used to construct nanotubular structures. For this purpose, the residues contained in such building blocks have been represented using blobs able to reproduce the energy contributions derived from the atomistic models. The number and position of the blobs used to describe each residue were defined by adjusting atomistic energies. Specifically, the following energy contributions were considered in the development of the CG model:

(1) electrostatic and van der Waals energies associated with the interaction between residues i and $i + n$ with $n \geq 2$, i.e., nonbonding energies calculated between residues separated by at least two positions within the building block ($E_{\mathrm{el}}^{i,i+n}$ and $E_{\mathrm{vdW}}^{i,i+n}$);

(2) sum of the torsional, electrostatic, and van der Waals energies derived from the interaction between two consecutive residues (i and $i + 1$), i.e., total energy associated with the 1−2 inter-residue interactions ($E_{\mathrm{inter\text{-}res}}^{i,i+1}$);

(3) total internal energy of each residue contained in the building block, i.e., sum of the torsional, electrostatic, and van der Waals energies calculated considering all of the atoms included in residue i ($E_{\mathrm{internal}}^{i}$); and

(4) sum of $E_{\mathrm{inter\text{-}res}}^{i,i+1}$ and $E_{\mathrm{internal}}^{i}$, i.e., total energy associated with residues i and $i + 1$ ($E_{\mathrm{tot}}^{i,i+1}$).

The justification of this wide set of energy contributions will be detailed in the next section, in which the CG potentials will be presented. It should be noted that stretching and bending energy contributions were not taken into account within this procedure. This is because, in the procedure used to generate and relax atomistic configurations, bond lengths and bond angles were kept fixed at the values of the starting structures, i.e., those derived from MD simulations.

The energy contributions $\{E_{\mathrm{el}}^{i,i+n}, E_{\mathrm{vdW}}^{i,i+n}, E_{\mathrm{inter\text{-}res}}^{i,i+1}, E_{\mathrm{internal}}^{i}, E_{\mathrm{tot}}^{i,i+1}\}$ were calculated for the 3×10^4 relaxed atomistic configurations of the 1krr building blocks. Adjustments to define the CG model for the 11 different residues contained in the sequence of the wild type 1krr building block were performed under the following considerations: (i) each residue must be satisfactorily described using as few blobs as possible and (ii) the number and position of the blobs may depend on the chemical nature of the amino acid. After some trials, we rapidly concluded that unfortunately no residue can be described using only one blob; i.e., the sample variances were significantly high when the fittings were performed using such a simple CG model. On the other hand, the positions of the blobs are a crucial decision in the development of a CG model. The most obvious alternative is to localize all of the blobs at physically reliable positions like the nuclei. However, this alternative did not provide satisfactory results, since the quality of the fittings, which were defined using the sample variance, was in many cases very poor. In order to overcome this difficulty, a systematic analysis was performed to define the preferred location of blobs using standard optimization routines,[65] so as to maximize the quality of the fitting between the atomistic and CG models. In order to locate the blobs at chemically intuitive positions, the resulting locations were subsequently redefined by introducing small controlled modifications through a number of systematic trials. This procedure led to a very satisfactory CG model with some blobs located at the nuclei and the others at chemically intuitive vectors. Specifically, three types of residues (Figure 3), which differ in the number of blobs used for their representation, were defined for the 1krr building block:

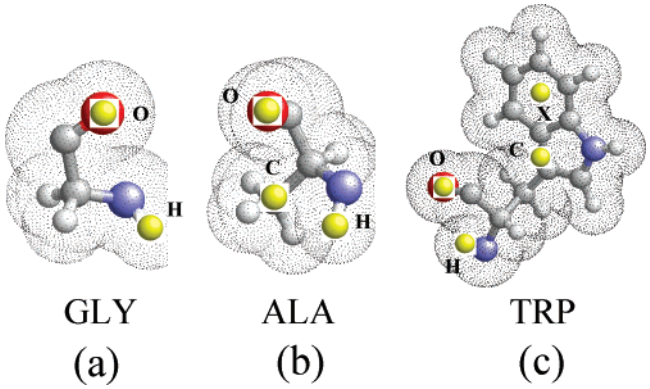

Figure 3. CG model for the three types of amino acids considered in this work. (a) Type 1, which only involves GLY, is represented by two blobs located at the positions of the amide hydrogen and oxygen atoms (H and O, respectively). (b) Type 2 contains ALA and VAL, the former being specifically displayed in the figure. In this case, the CG model involves three blobs: H, O, and the third one, which is denoted as C, located at the geometric center of the side chain. (c) All of the remaining amino acids correspond to type 3, which have been illustrated with TRP. These amino acids are represented using four blobs: H, O, C, and a fourth one, which is denoted as X, whose position depends on the chemical nature of the amino acid (see text).

(1) Type 1. GLY, which is the only amino acid contained in this category, is represented with two blobs centered at the positions of the amide hydrogen and oxygen atoms (Figure 3a). Hereafter, the blobs located at these positions will be denoted as H and O, respectively.

(2) Type 2. This category contains ALA and VAL. In this case, three blobs are required to describe each residue: two at the same positions found for the H and O blobs of GLY and the third located at the geometrical center of the side group (Figure 3b), i.e., methyl (ALA) and isopropyl (VAL). The latter blob has been denoted as C.

(3) Type 3. All of the remaining amino acids are included in this category. In this case, a fourth blob is added to those three considered for VAL and ALA (Figure 3c). The position of the fourth blob depends on the chemical nature of the amino acid. Thus, for the residues contained in the 1krr building block (Table 1), the position of the fourth blob, which has been denoted as X, coincides with that of ASN, the nitrogen atom of the side chain; ASP, the carbon atom of the side carboxylate group; HIS, the geometric center of the ring; ILE, the intermediate distance between the two methyl groups of the side chain; PRO, the C^{γ} atom of the cycle; SER, the middle of the O−H bond; THR, the intermediate distance between the methyl group and the hydrogen atom of the hydroxyl; TRP, the geometric center of the six-membered ring.

This CG description provides the best fitting between atomistic and CG energies. On the other hand, the only parameters required to define the CG model of each residue are the number and positions of the blobs. Thus, these blobs should be considered as simple geometric points rather than as particles with a well-defined size and hardness, like the atoms and pseudoatoms usually employed in conventional MD simulations. Accordingly, the interaction of each blob with the others will depend only on their relative spatial orientation, which will be defined using distances, angles, and dihedrals (see below).

Coarse-Grained Interactions

In this section, we derive the CG potentials used to describe the energy of the β-helical building blocks. Our attention focuses on the degree of reliability and accuracy of the developed CG potentials for the 11 amino acids contained in the sequence of the 1krr building block. First, the following important considerations are taken into account:

(1) Interactions are categorized according to the three types of energy contributions mentioned above: nonbonding interactions between residues i and $i+n$ with $n \geq 2$, which are separated into electrostatic and van der Waals; interactions between residues i and $i+1$; and internal interactions of residue i. All of these interactions have been treated separately, and the complete CG potential for a given pair of interacting residues is defined as the sum of the analytical expressions obtained for each class of interaction.

(2) Although the maximum number of blobs used to describe amino acids different from GLY, VAL, and ALA is four, only three of these are needed to describe some of the interactions mentioned above, i.e., electrostatic and van der Waals interactions between residues i and $i+n$ with $n \geq 2$. Thus, for each residue of an interacting pair, the three blobs that provide the best adjustment between the atomistic and CG energies are selected to define the potentials that describe these energy contributions. Moreover, the selected blobs do not necessarily need to be the same for the electrostatic and van der Waals contributions. Accordingly, nonbonding interactions that involve nonconsecutive residues of type 3 (see previous section) are actually described by a *fluctuating* three-blob model rather than by a *fixed* four-blob model. Thus, each nonbonding contribution is described using three blobs selected from a set of four blobs through a statistical criterion, i.e., lowest sample variance.

Nonbonding Interactions between Residues i and $i+n$ with $n \geq 2$. As mentioned above, nonbonding interactions between residues that are not directly linked were separated into two contributions: electrostatic and van der Waals. We observed that nonbonding interactions, particularly the electrostatic contribution, between residues i and $i+n$ with $n \geq 2$ strongly depend on n, even when the interacting residues are chemically identical, e.g., GLY(5)···GLY(11) and GLY(5)···GLY(19) in the building block of 1krr. In order to overcome this difficulty, we decided to consider different analytical expressions for the electrostatic contribution, with the function able to provide the best fitting with the atomistic energies being selected to define the CG electrostatic potential of the interacting residues. Thus, within the same building block, the analytical expressions used to represent the electrostatic interactions between a given pair of nonconsecutive residues may be completely different from that employed to represent the same energy contribution of the next interacting pair. Moreover, pairs of interacting residues that are chemically equivalent, e.g., SER(12)···ALA(30), SER(26)···ALA(30), and ALA(30)···SER(32) in 1krr, may be described using different potentials.

Specifically, the three analytical expressions considered for the CG electrostatic potential between residues i and $i+n$ with $n \geq 2$ were

$$V_{\mathrm{el}}^{i,i+n}(i,j) = \frac{\lambda_1}{d_{i,j} - \delta_1} + \frac{\lambda_2}{(d_{i,j} - \delta_2)^2} \quad (1)$$

$$V_{\mathrm{el}}^{i,i+n}(i,j) = \lambda_1 e^{[\alpha_1 (d_{ij} - \delta_1)^2]} e^{\beta_1 / d_{ij}} + \lambda_2 e^{[\alpha_2 (d_{ij} - \delta_2)^2]} e^{\beta_2 / d_{ij}} \quad (2)$$

$$V_{\mathrm{el}}^{i,i+n}(i,j) = \lambda_1 e^{[\alpha_1 (d_{ij} - \delta_1)^2]} e^{\beta_1 d_{ij}} + \lambda_2 e^{[\alpha_2 (d_{ij} - \delta_2)^2]} e^{\beta_2 d_{ij}} \quad (3)$$

where $j = i + n$, $\{\lambda_1, \lambda_2, \delta_1, \delta_2, \alpha_1, \alpha_2, \beta_1$, and $\beta_2\}$ are adjustable parameters, and d_{ij} is the distance between the two interacting blobs. As can be seen, there is a clear resemblance between the classical Coulombic expression typically used in atomistic force fields to evaluate the electrostatic energy and eq 1, the latter involving only four adjustable parameters. In spite of this resemblance, it should be noted that the analytical expressions typically used for CG potentials are physically unmeaning, the shape of these functions being selected to match satisfactorily the atomistic energies. On the other hand, visual inspection of the graphical representations d_{ij} vs atomistic electrostatic energies reveals that, depending on the chemical nature of the interacting residues and n, very high and low values of $V_{\mathrm{el}}^{i,i+n}(i,j)$ may appear. In order to facilitate the fitting of such values, expressions with Gaussian-like functions have been considered in eqs 2 and 3. Furthermore, the quality of the fitting provided by such analytical expressions has been improved by introducing eight different adjustable parameters.

For each pair of interacting residues, the strategy for selecting the most appropriate expression consists of the following steps. First, an optimization routine was used to provide the best adjustable parameters for the three analytical expressions considering all possible CG models. As was mentioned above, GLY was always represented using a two-blob model, ALA and VAL were represented by a fixed three-blob model, and the remaining residues (those of type 3) have been described using a fluctuating three-blob model, i.e., a three-blob model in which the positions of the CG particles are chosen from a set of four blobs. This means that, for all of the residues with the exception of GLY, ALA, and VAL, four different three-blob models can be defined [(H,O,C), (H,O,X), (H,C,X), and (O,C,X)], which significantly increase the number of functions to be optimized. For example, if we consider the interaction of one residue of type 3 with one residue of type 2, SER with ALA, the number of fittings that needs to be optimized is 4(models of SER) $\times$ 1(model of ALA) $\times$ 3(analytical expressions) $=$ 12. Interestingly, we found that in all cases the best fittings were obtained when the model includes the H and O blobs, and the third blob fluctuates between C and X depending on both n and the chemical nature of the residues. Obviously, the objective of the optimization routine is to minimize the error between the CG and atomistic energies (E^{CG} and E^{at}, respectively):

$$\mathrm{err} = \sum_{m=1}^{30\,000} (E^{\mathrm{CG}}(m) - E^{\mathrm{at}}(m))^2 \quad (4)$$

where m is the number of generated and relaxed atomistic configurations for the 1krr building block. In general, the adjustment with the lowest error was used to define the shape of the analytical expression that describes the CG electrostatic potential and the three-blob model. However, if the difference between the errors provided by the best adjustment obtained using eq 1 and the best adjustment derived from eqs 2 and 3 is lower than 10%, the CG electrostatic potential is defined using eq 1. This particular condition ensures a compromise between the quality of the potential and the computational resources required to evaluate the CG energies during the simulations. This procedure was applied to obtain the 561 different electrostatic potentials that are needed to evaluate at the CG level the electrostatic interactions between residues i and $i+n$ with $n \geq 2$ for the 35 residues of the 1krr building block. These potentials are available upon request to the authors.

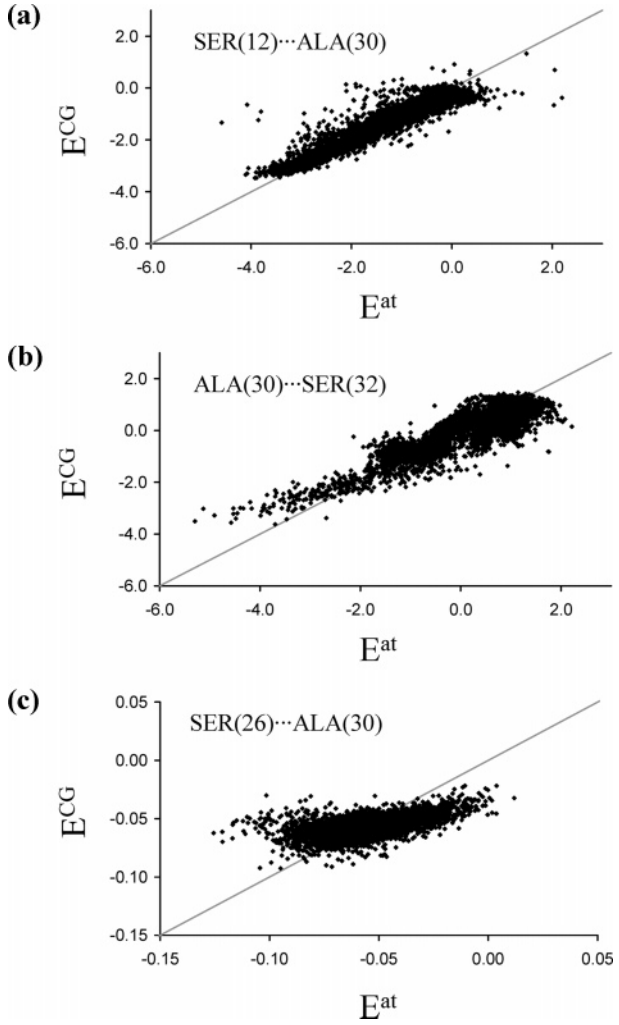

Figure 4. Graphical representation of the atomistic electrostatic energies (E^{at}; in kcal/mol) vs the CG electrostatic energies (E^{CG}; in kcal/mol) for the (a) SER(12)···ALA(30), (b) ALA(30)···SER(32), and (c) SER(26)···ALA(30) pairs of interacting residues. The 3×10^4 atomistic microstructures generated and relaxed for the 1krr building block have been considered. CG energies were calculated using the procedure and potentials explained in the text.

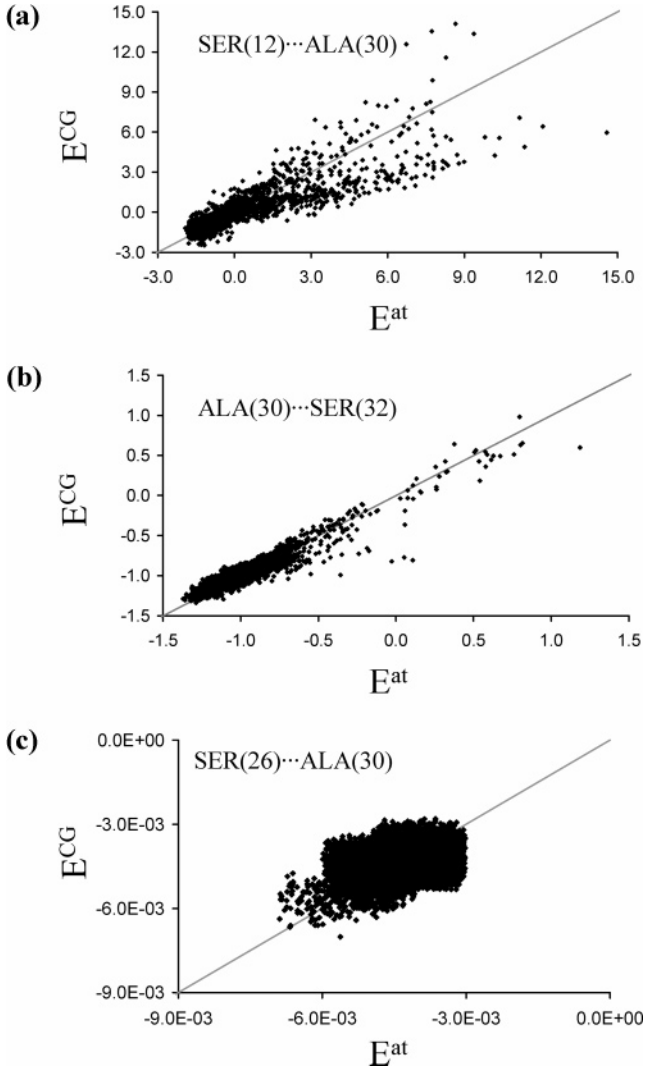

Figure 5. Graphical representation of the atomistic van der Waals energies (E^{at}; in kcal/mol) vs the CG van der Waals energies (E^{CG}; in kcal/mol) for the (a) SER(12)···ALA(30), (b) ALA(30)···SER(32), and (c) SER(26)···ALA(30) pairs of interacting residues. The 3×10^4 atomistic microstructures generated and relaxed for the 1krr building block have been considered. CG energies were calculated using the procedure and potentials explained in the text.

The strong dependence of the CG potential on n is illustrated in Figure 4, which represents the atomistic vs CG energies for three different ALA···SER interactions. Figure 4a represents the energies obtained for the interaction SER(12)···ALA(30), i.e., $n = 18$ (see Table 1). In this case, electrostatic interactions were described at the CG level using the analytical expression displayed in eq 2 and considering that the third blob of SER(12) is located at the geometrical center of the side group, which provided the best fitting of the atomistic electrostatic energies. As can be seen, the correspondence between the energies calculated using the atomistic Coulombic expression and the complex CG potential was excellent. However, this correspondence becomes slightly worse when the interaction between ALA(30) and SER(32), i.e., $n = 2$, is considered. This interaction has been described at the CG level using the analytical expression displayed in eq 3 and considering that the third blob is located in the middle of the O−H bond of SER(32). This is reflected in Figure 4b, which represents the atomistic vs CG energies. As can be seen, in this case the weaker interactions, i.e., those slightly attractive or repulsive, are well reproduced at the CG level. However, the quality of the adjustment is clearly unsatisfactory for the more attractive interactions. In spite of this, we consider that the interaction ALA(30)···SER(32) is well reproduced at the CG level, since strong attractive interactions between ALA(30) and SER(32) are relatively rare, as evidenced by the small number of points ranging from −3.0 to −5.0 kcal/mol (Figure 4b). Finally, Figure 4c presents the electrostatic energies for the interaction SER(26)···ALA(30), i.e., $n = 4$. As can be seen, the correspondence between the atomistic and CG models is very poor in this case. Indeed, the three analytical expressions listed above were not able to provide a good fitting. However, it should be noted that very weak electrostatic interactions are involved in this pair of residues, i.e., the highest and lowest values of the atomistic energy are 0.01 and −0.12 kcal/mol, respectively. Thus, the range of variation is so small that the contribution provided by this pair of residues should be considered negligible with respect to those with $n = 18$ and 2. For this reason, eq 1 was used to define the CG potential of SER(26)···ALA(30), even though the error is slightly higher than those provided by eqs 2 and 3. However, the simplicity of this analytical expression will accelerate the calculations during the CG simulations.

Regarding the van der Waals interactions, we noted that their contribution is less significant than that of the electrostatic ones. The following analytical expression, which contains four

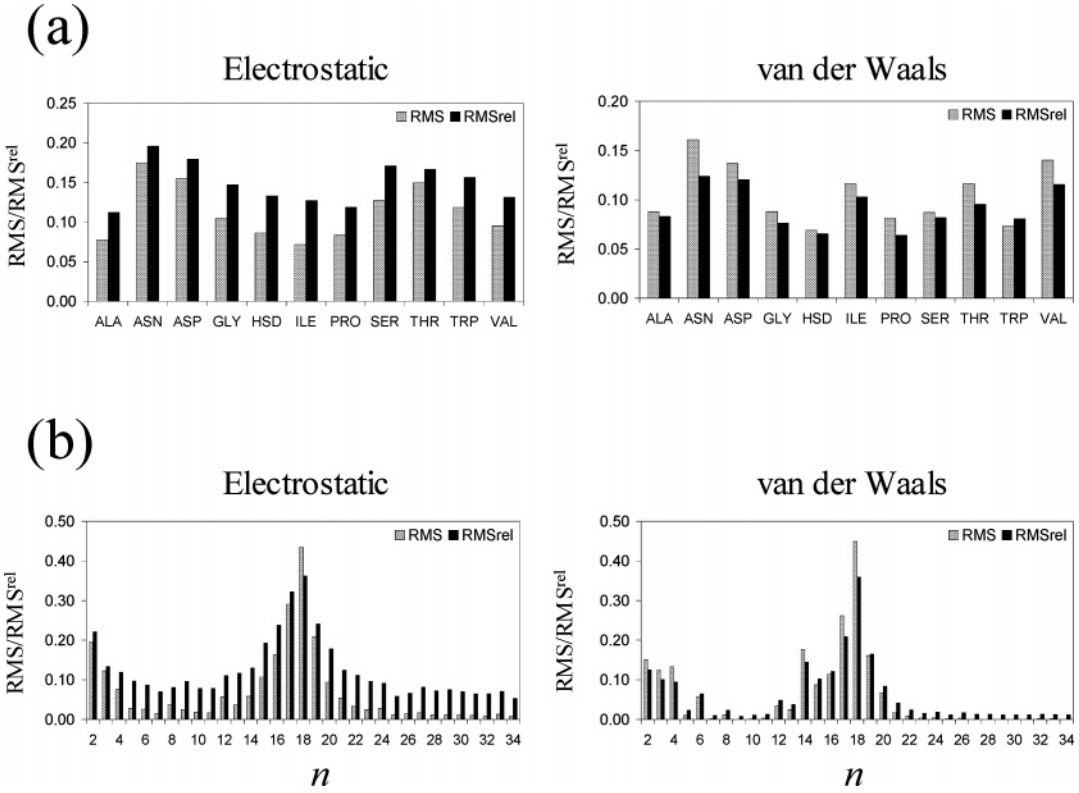

Figure 6. rms and rms^{rel} of the energies associated with the electrostatic and van der Waals interactions between residues i and $i + n$ with $n \geq 2$ for the different amino acids (a) and the values of n (b) found in the 1krr building block.

adjustable parameters $\{\sigma_1, \sigma_2, \kappa_1, \kappa_2\}$, was used to describe the van der Waals potential between residues i and $i + n$ with $n \geq 2$ at the CG level:

$$V_{\mathrm{vdW}}^{i,i+n}(i,j) = \frac{\sigma_1}{(d_{i,j})^{\kappa_1}} + \frac{\sigma_2}{(d_{i,j})^{\kappa_2}} \tag{5}$$

where $j = i + n$ and d_{ij} is the distance between the two interacting blobs. As above, residues different from GLY, ALA, and VAL were described using three blobs that were chosen from the four different possibilities, i.e., the fluctuating three-blob model. An optimization routine was used to determine the four parameters that minimize the sum of the quadratic differences between the CG and atomistic van der Waals energies (eq 4).

The results of the fittings are illustrated in Figure 5, which represents the atomistic vs CG energies obtained for interactions associated with SER(12)···ALA(30), ALA(30)···SER(32), and SER(26)···ALA(30). As can be seen, the strength of the van der Waals interactions also depends strongly on n. However, the adjustment was excellent in all cases, even though only one relatively simple analytical expression was considered for the CG potential. It is worth noting that the strength of the van der Waals interaction for the ALA···SER residues follows the same order that was previously found for the electrostatic interactions: the strongest interactions are those with $n = 18$, while those between residues with $n = 4$ are almost negligible. This correspondence between van der Waals and electrostatic interactions was also detected in all cases formed by pairs of residues with identical chemical structure but different n. On the other hand, as was also found for the electrostatic interactions, the best fittings always involve blobs centered at the positions of the amide hydrogen and oxygen atoms, while the position of the third blob depends on both the chemical nature of the interacting residues and n. For the interactions between ALA and SER (Figure 5), the third blob of the latter was located at the geometrical center of the side group for $n = 18$ and $n = 4$, while for $n = 2$ it was centered in the middle of the O−H bond.

The statistical evaluation of the similarity between the nonbonding energies calculated using atomistic and CG potentials was carried out considering (i) the root-mean-square deviation (rms) and (ii) the relative root-mean-square deviation (rms^{rel}):

$$\mathrm{rms} = \sqrt{\frac{\sum_i (E_i^{\mathrm{at}} - E_i^{\mathrm{CG}})^2}{N}} \tag{6}$$

$$\mathrm{rms}^{\mathrm{rel}} = \sqrt{\frac{\sum_i (E_i^{\mathrm{at}} - E_i^{\mathrm{CG}})^2}{N \cdot E_{\mathrm{max}}^i}} \tag{7}$$

where E^{at} and E^{CG} are the energies calculated using the atomistic and CG potentials, respectively, and E_{max} corresponds to the highest value chosen between $|E^{\mathrm{at}}|$ and $|E^{\mathrm{CG}}|$. These statistical parameters reflect the high quality of the nonbonding potentials used to describe the electrostatic and van der Waals interactions between residues i and $i + n$ with $n \geq 2$ through the CG model presented in the previous section.

Figure 6a shows the averaged rms and rms^{rel} values of the energies associated with the electrostatic and van der Waals interactions between residues i and $i + n$ with $n \geq 2$ for the 11 different amino acids contained in the 1krr building block. As can be seen, both the rms and rms^{rel} values were remarkably low for the two energy contributions, especially for the van der Waals one. Interestingly, the description of the nonbonding interactions at the CG level seems to depend on the chemical

nature of the amino acids. Specifically, polar (ASN, ASP, SER, and THR) and aromatic (TRP) amino acids provided worse results for the electrostatic interactions, while ASN, VAL, and ASP provided the poorer adjustments of the van der Waals energies. In spite of this, it should be emphasized that the proposed CG model is able to provide very satisfactory results for the 11 amino acids.

Figure 6b represents the averaged rms and rms^{rel} values of the nonbonding energies between residues i and $i + n$ with $n \geq 2$ against n. Again, the two statistical parameters present very low values, indicating that nonbonding interactions, especially the van der Waals one, between residues that are not directly attached can be accurately reproduced at the CG level. Analysis of the results reveals that the nonbonding interactions between residues separated by about 15−20 positions show the worst fittings. These residues typically correspond to those arranged one in front of the other but located in different turns of the helix. Thus, the strength of the interaction between these residues is high because they are close in space, even though they are far in the sequence. Accordingly, it is very reasonable that the statistical parameters increase with the strength of the interaction.

Interactions between Residues *i* and *i* + 1. Atomistic electrostatic, van der Waals, and torsional interactions between residues i and $i + 1$ have been joined together at the CG level. Thus, a simple linear function involving spatial geometric parameters, which allows description of the relative orientation between the two interacting residues, has been used to capture such three atomistic energy contributions. The spatial geometric parameters introduced in the CG expression consist of distances, angles, and dihedrals between blobs that usually but not necessarily involve the two residues. This is a striking feature because the aim of this CG potential is to describe the inter-residue atomistic interactions. However, we found that the fitting between the sum of the inter-residue atomistic energy contributions and the CG energies improves significantly when the geometric parameters defined by blobs located at the same residue are also included in the analytical expression of the CG potential.

Another important consideration is that all of the blobs defined above for residues of type 1, 2, and 3 were initially included in the adjustment. Thus, residues different from GLY, ALA, and VAL have been described using the four-blob model rather than the fluctuating three-blob model employed for the nonbonding interactions between residues i and $i + n$ with $n \geq 2$. It should be noted that a very large number of distances, angles, and dihedrals are defined by combining the blobs contained in two residues. However, in order to simplify the shape of the analytical expression, only the most essential geometric parameters are included in the CG potential. In practice, only the geometric parameters that improve the correlation coefficient are used in the adjustment, whereas those that do not affect the quality of the fitting are omitted. For instance, the CG potential needed to describe the interaction between PRO(1) and ILE(2) may involve $4 + 4 = 8$ different blobs, which may be combined to form a significantly high number of pairs, triads, and tetramers able to define distances, angles, and dihedrals, respectively. However, only three distances, two angles, and two dihedrals appear in the analytical expression developed for the CG potential:

$$V^{i,i+1}_{\mathrm{PRO-ILE}} = -385.97 + 7.21\varphi_{\mathrm{C1-O1-H2-O2}} + 5.10(\alpha_{\mathrm{O1-H2-O2}} - 1.73)^2 + 2.15\varphi_{\mathrm{C1-H1-O1-H2}} + 0.57d_{\mathrm{C2-H1}} + 111.46d_{\mathrm{O1-H2}} + \frac{1.00}{\alpha_{\mathrm{O1-H2-C2}}^2} + 1.05\alpha_{\mathrm{X1-C1-H1}} + 1.57d_{\mathrm{H2-O2}} \quad (8)$$

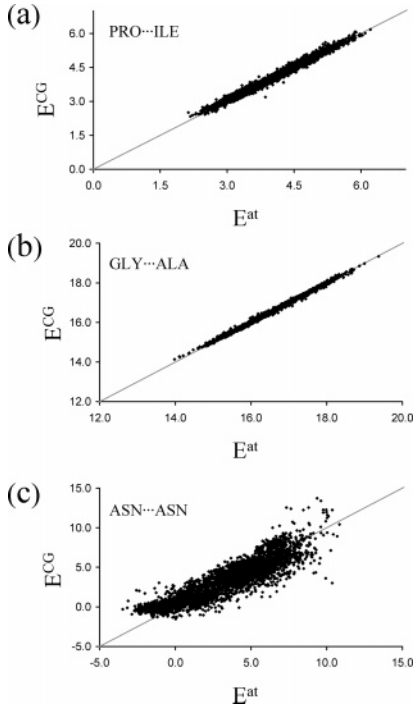

Figure 7. Graphical representation of the atomistic i, $i + 1$ inter-residue energies (E^{at}; in kcal/mol) vs the CG i, $i + 1$ inter-residue energies (E^{CG}; in kcal/mol) for the (a) PRO···ILE, (b) GLY···ALA, and (c) ASN··· ASN pairs of residues. The 3×10^4 atomistic microstructures generated and relaxed for the 1krr building block have been considered. CG energies were calculated using the procedure and potentials explained in the text.

where d, α, and φ refer to the distance, angle, and dihedral formed by two, three, and four blobs, respectively; O, H, and C correspond to the blobs centered at the oxygen atom, amide hydrogen atom, and the geometrical center of the side group, respectively, while X is the fourth blob (centered at the C^{γ} atom of the cycle and at the intermediate distance between the two methyl side groups for PRO and ILE, respectively); and the labels 1 and 2 indicate that a given blob belongs to PRO and ILE, respectively. However, in spite of its simplicity, eq 8 provides an excellent adjustment (Figure 7a).

The procedure used to develop the CG potentials able to describe interactions between consecutive residues presents another remarkable difference with respect to that of nonbonding interactions between residues i and $i + n$ with $n \geq 2$. The potentials have been developed taking into account the chemical nature of the interacting residues and their relative order along the sequence but not their global position in the sequence. For example, the interaction between ILE(4)···GLY(5) is described by the same potential as the interactions ILE(10)···GLY(11), ILE(22)···GLY(23), and ILE(28)···GLY(29), whereas the interactions GLY(29)···ALA(30) and ALA(30)···GLY(31) are represented by two different potentials. This procedure allows reducing the number of CG potentials required to describe the 1krr building block, which contains 35 residues, from 35 to 27. Although this is not a drastic simplification, this strategy would be very useful for building blocks with repetitive sequences. Unfortunately, this advantage was not possible for the nonbonding interactions between residues i and $i + n$ with $n \geq 2$, since the dependence of the CG potential on n was very high.

Suitable routines were programmed to find automatically the linear function that provides the best fitting in every case. Thus, the effect of each geometric parameter on the correlation coefficient was examined, with the variables that were necessary for the goodness of the fitting being the only ones included in

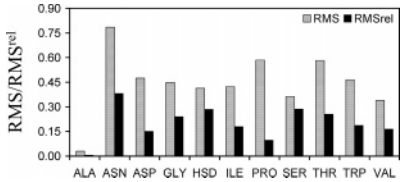

Figure 8. rms and rms[rel] of the energies associated with the $i, i+1$ inter-residue interactions for the different amino acids contained in the 1krr building block.

the final analytical expression. In addition, several shapes were considered for each geometric parameter, e.g., for the distances (d) we considered: d, $1/d$, $1/d^2$, $1/\sqrt{d}$, d^2, etc. Parts b and c of Figure 7 show the adjustments between atomistic and CG energies obtained for the residues GLY···ALA and ASN···ASN, which correspond to the best and worst fitting, respectively. As can be seen, even in the worst case, the CG potentials derived from this procedure are able to provide acceptable results. Interestingly, the analytical expression obtained for GLY···ALA (eq 9) is significantly simpler than that derived for ASN···ASN (eq 10):

$$V^{i,i+1}_{\mathrm{GLY-ALA}} = -399.50 + \frac{6.40}{\alpha_{\mathrm{O1-H1-O2}}} + 2.31\varphi_{\mathrm{H1-O1-H2-O2}} + 123.70 d_{\mathrm{O1-H2}} + 22.12(\alpha_{\mathrm{H1-O1-H2}} - 1.64)^2 + 6.31(d_{\mathrm{C2-O1}} - 4.70)^2 - 0.15(d_{\mathrm{C2-O2}} - 3.31)^2 \quad (9)$$

$$V^{i,i+1}_{\mathrm{ASN-ASN}} = 4.67 - 1.70(\alpha_{\mathrm{X1-C1-H1}} - 1.32)^2 + 1.30(d_{\mathrm{H1-C2}} - 5.63)^2 - \frac{39.91}{d_{\mathrm{X1-H2}}{}^2} - 73.00(\alpha_{\mathrm{O1-H2-O2}} - 0.87)^2 + 5.01(\alpha_{\mathrm{C1-H1-O1}} - 0.91)^2 + \frac{35.51}{d_{\mathrm{X1-X2}}{}^2} + 1.00(\alpha_{\mathrm{H2-C2-X2}} - 1.23)^2 - 0.03(\varphi_{\mathrm{X1-C1-C2-X2}} - 6.28)^2 - 46.41(\alpha_{\mathrm{H1-O1-H2}} - 1.63)^2 - 14491(d_{\mathrm{O1-H2}} - 3.11)^2 \quad (10)$$

where the nomenclature is identical to that explained above for eq 8. The analytical expressions of all of the potentials derived for the 1krr building block are provided in the Supporting Information.

Figure 8 shows the averaged rms and rms[rel] values of the energies associated with the interactions between two consecutive residues for the 11 different amino acids contained in the 1krr building block. In general, the values of rms and rms[rel] are higher than those obtained for the electrostatic and van der Waals interactions between residues i and $i+n$ with $n \geq 2$ (Figure 6). This is an expected result, since, typically, the error associated with the derived CG potentials increases when the separation between the interacting residues decreases. In spite of this, the results displayed in Figure 8 indicate that the calculated CG potentials are able to describe very satisfactorily the $i, i+1$ inter-residue interactions for the 11 amino acids. On the other hand, the highest values of rms and rms[rel] were found for the ASN amino acid, which also provided the poorest results for the nonbonding interactions between residues i and $i+n$ with $n \geq 2$. Thus, the coarse-graining of the interactions associated with this specific polar amino acid is not an easy

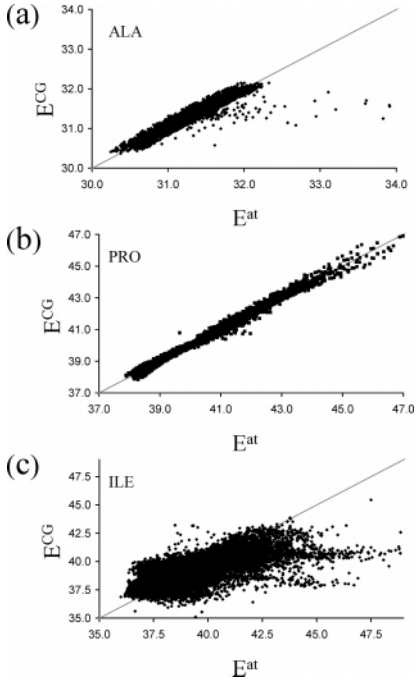

Figure 9. Graphical representation of the atomistic intra-residue energies (E^{at}; in kcal/mol) vs the CG intra-residue energies (E^{CG}; in kcal/mol) for the (a) ALA and (b) ILE residues. The 3×10^4 atomistic microstructures generated and relaxed for the 1krr building block have been considered. CG energies were calculated using the procedure and potentials explained in the text.

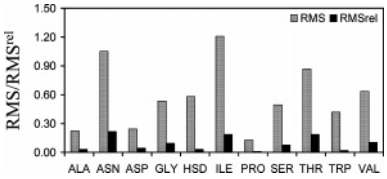

Figure 10. rms and rms[rel] of the energies associated with the intra-residue interactions for the different amino acids contained in the 1krr building block.

task. This should be attributed to the range of energy values that are captured by the CG potential, which is significantly wider than that for other amino acids (see Figure 7c).

Intra-residue Interactions. The CG potentials for intra-residue interactions were defined using a procedure identical to that described above for interactions between residues i and $i+1$. Thus, intra-residue energies were adjusted to a linear function that contains distances, angles, and dihedrals as variables. However, in this case, the number of parameters that were considered in the fitting of the atomistic and CG energies was significantly reduced. Thus, for amino acids different from GLY, ALA, and VAL, these geometric parameters consist of two dihedrals ($\varphi_{\mathrm{X-C-O-H}}$ and $\varphi_{\mathrm{X-C-H-O}}$), three angles ($\alpha_{\mathrm{X-C-H}}$, $\alpha_{\mathrm{X-C-O}}$, and $\alpha_{\mathrm{C-H-O}}$), and six distances ($d_{\mathrm{H-O}}$, $d_{\mathrm{H-C}}$, $d_{\mathrm{H-X}}$, $d_{\mathrm{O-C}}$, $d_{\mathrm{O-X}}$, and $d_{\mathrm{C-X}}$). For ALA and VAL, only one angle and three distances ($\alpha_{\mathrm{C-H-O}}$, $d_{\mathrm{H-O}}$, $d_{\mathrm{H-C}}$, and $d_{\mathrm{O-C}}$) were considered, while $d_{\mathrm{H-O}}$ was the only geometric parameter used for GLY. According to the strategy previously used for the interactions between residues i and $i+1$, the geometric parameters that improve the correlation coefficient were the only ones explicitly included in the CG potential; i.e., the contribution of each geometric parameter to the quality of the fitting was tested automatically using suitable routines.

CG potentials were specifically developed for the 11 amino acids contained in the 1krr building block. The simplest, the best, and the worst adjustments, which are displayed in Figure

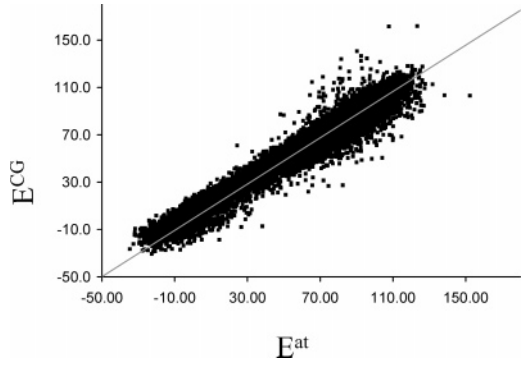

Figure 11. Graphical representation of the atomistic (E^{at}; in kcal/mol) vs the CG total energies (E^{CG}; in kcal/mol) for 5×10^4 independent configurations of the 1krr building block, which have not been used in the development of the CG model.

9, were obtained for ALA, PRO, and ILE, respectively, with the resulting CG potentials being

$$V_{ALA} = -45.27 + \frac{8.16}{\alpha_{C-H-O}^{2}} + \frac{473.46}{d_{C-H}^{2}} - 0.5 d_{H-O}^{2} + 9.34 \alpha_{C-H-O}^{2} \quad (11)$$

$$V_{PRO} = 34.19 - 6.18 d_{C-X} + \frac{108.67}{d_{O-X}^{2}} + 14.43 \cos^2 \varphi_{X-C-O-H} + \frac{49.39}{d_{C-H}^{2}} + \frac{0.40}{d_{C-X}^{2}} - \frac{17.55}{\alpha_{X-C-O}^{2}} + 12.76 \cos \varphi_{X-C-O-H} - \frac{19.11}{\alpha_{X-C-H}} - \frac{118.75}{d_{H-O}^{2}} - 1.45 \alpha_{X-C-O}^{2} \quad (12)$$

$$V_{ILE} = 29.65 - 0.76 d_{H-X}^{2} + 0.33 d_{H-O}^{2} + 1.02 \cos \varphi_{X-C-O-H} + 0.76 d_{O-C}^{2} - \frac{2.69}{\sqrt{\alpha_{X-C-H}}} + 0.84 \cos^2 \varphi_{X-C-H-O} + 0.19 \cos \varphi_{X-C-H-O} + \frac{0.01}{\alpha_{X-C-H}^{2}} + 0.48 \cos \varphi_{X-C-O-H}^{2} - \frac{0.71}{\sqrt{\alpha_{X-C-O}}} + \frac{0.01}{\alpha_{X-C-O}^{2}} + \frac{13.73}{\alpha_{C-H-O}^{2}} \quad (13)$$

The potentials for the remaining amino acids are provided in the Supporting Information. Figure 10 shows the averaged rms and rms^{rel}values of the intra-residue energies for the 11 amino acids contained in the building block of 1krr. As can be seen, the rms values found for ASN, ILE, and THR are significantly high, i.e., 1.05, 1.21, and 0.91 kcal/mol, respectively. However, the effect of the intra-residue energy contribution on the general stability of the β-helical motif is very small. Furthermore, the importance of this contribution to the general stability of the nanotubes formed by self-assembling of β-helical building blocks, which is our final objective, is expected to be almost negligible. This point will be specifically checked in the next studies devoted to coarse-graining the self-assembly of the 1krr building blocks.

Test Calculations

The reliability of the CG model proposed in this work has been tested by comparing the atomistic and CG total energies calculated for 5×10^4 independent configurations, which were not used in the development of the model, of the 1krr building block. Figure 11 compares the energies of such structures calculated using the atomistic and CG potentials. As can be seen, there is an excellent agreement between the two methodologies, with the rms and rms^{rel} values being 4.6 kcal/mol and 4.1%, respectively. It should be noted that these statistical parameters are remarkably low, especially if we consider that the atomistic energies range from -32 to 155 kcal/mol.

These results clearly demonstrate that the procedure presented in this work to develop CG potentials for protein building blocks is very satisfactory. Thus, atomistic energies are reproduced accurately using a simplified CG model. At the present time, we are developing a CG potential to calculate the interaction energies between different building blocks, which combined with that presented in this work is expected to allow mesoscopic simulations of large nanotubes formed by self-assembled β-helical building blocks.

Conclusions

This work reports a general procedure to develop CG models of protein building blocks, which has been applied to the 1krr β-helical fragment that forms stable tubular nanoconstructs. In this model, the explicit atoms of each residue have been replaced by a small set of blobs located at chemically intuitive positions that are able to retain the essential physics of the systems. Specifically, the 493 explicit particles contained in the detailed atomistic model of the 1krr building block have been reduced to 127 blobs in the developed CG model. This reduction involves a significant decrease in the number of interactions, especially between nonbonding pairs, that require energy evaluation.

On the other hand, the interactions between blobs have been represented considering three different contributions: nonbonding interactions between residues i and $i + n$ with $n \geq 2$, which involve both electrostatic and van der Waals terms; interactions between consecutive residues (i and $i + 1$); and intra-residue interactions. The potentials developed to represent these interactions are very satisfactory, as indicated by the low rms and rms^{rel} values, which were calculated by comparing the atomistic and CG energies. It is worth noting that the CG model is about 2 orders of magnitude faster than the atomistic model. In general, the intra-residue energies provided the largest values for these statistical parameters, while the smallest values correspond to the energies associated with the nonbonding interactions between residues i and $i + n$ with $n \geq 2$. This is a very important result, since the latter contribution is essential to retain the β-helix conformation of the building block.

Acknowledgment. The authors are indebted to the Universitat de Lleida and to the Centre de Supercomputació de Catalunya (CESCA) for computational facilities. Dr. David Zanuy is thanked for his kind assistance in the molecular dynamics simulations. This project has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, under Contract No. N01-CO-12400. The content of this publication does not necessarily reflect the view of the policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organization imply endorsement by the U.S. Government. This research was supported (in part) by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research.

Supporting Information Available: Potentials to describe the interactions between residues i and $i + 1$ and intra-residue

interactions. This material is available free of charge via the Internet at http://pubs.acs.org.

References and Notes

(1) Liao, S.; Seeman, N. C. *Science* **2004**, *306*, 2072.
(2) Yan, H.; Park, S. H.; Finkelstein, G.; Reif, J. H.; LaBean, T. H. *Science* **2003**, *301*, 1882.
(3) Chworos, A.; Severcan, I.; Koyfman, A. Y.; Weinkam, P.; Oroudjev, E.; Hansma, H. G.; Jaeger, L. *Science* **2004**, *306*, 2068.
(4) Percec, V.; Dulcey, A. E.; Balagurusamy, V. S. K.; Miura, Y.; Smidrkal, J.; Peterca, M.; Nummelin, S.; Edlund, U.; Hudson, S. D.; Heiney, P. A.; Hu, D. A.; Magonov, S. N.; Vinogradov, S. A. *Nature* **2004**, *430*, 764.
(5) Rajagopal, K.; Schneider, J. P. *Curr. Opin. Struct. Biol.* **2004**, *14*, 480.
(6) Valery, C.; Paternostre, M.; Robert, B.; Gulik-Krzywicki, T.; Narayanan, T.; Dedieu, J. C.; Keller, M. L.; Cherif-Cheikh, R.; Calvo, P.; Artzner, E. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 10258.
(7) Yokoi, H.; Kinoshita, T.; Zhang, S. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 8414.
(8) Main, E. R. G.; Lowe, A. R.; Mochrie, S. G. J.; Jackson, S. E.; Regan, L. *Curr. Opin. Struct. Biol.* **2005**, *15*, 464.
(9) Kajander, T.; Cortajarena, A. L.; Main, E. R. G.; Mochrie, S. G. J.; Regan, L. *J. Am. Chem. Soc.* **2005**, *127*, 10188.
(10) Tsai, H.-H.; Tsai, C.-J.; Ma, B.; Nussinov, R. *Protein Sci.* **2004**, *13*, 2753.
(11) Main, E. R. G.; Stott, K.; Jackson, S. E.; Regan, L. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 5721.
(12) Tovar, J. D.; Claussen, R. C.; Stupp, S. I. *J. Am. Chem. Soc.* **2005**, *127*, 7337.
(13) Reches, M.; Gazit, E. *Science* **2003**, *300*, 625.
(14) Zhang, S. *Nat. Biotechnol.* **2003**, *21*, 1171.
(15) Alemán, C.; Zanuy, D.; Jiménez, A. I.; Cativiela, C.; Haspel, N.; Zheng, J.; Wolfson, H.; Nussinov, R. *Phys. Biol.* **2006**, *3*, S54.
(16) Tsai, C.-J.; Zheng, J.; Alemán, C.; Nussinov, R. *Trends. Biotechnol.* **2006**, *24*, 449.
(17) Tsai, C.-J.; Zheng, J.; Zanuy, D.; Haspel, N.; Wolfson, H.; Alemán, C.; Nussinov, R. *Proteins*, in press.
(18) Tsai, C.-J.; Zheng, J.; Nussinov, R. *PLoS Comput. Biol.* **2006**, *2*, e42.
(19) Tsai, H.-H.; Tsai, C.-J.; Ma, B.; Nussinov, R. *Protein Sci.* **2004**, *13*, 2753.
(20) Haspel, N.; Zanuy, D.; Aleman, C.; Wolfson, H.; Nussinov, R. *Structure* **2006**, *14*, 1137.
(21) Tsai, C.-J.; Maizel, J. V., Jr.; Nussinov, R. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 12038.
(22) Zheng, J.; Zanuy, D.; Haspel, N.; Tsai, C.-J.; Alemán, C.; Nussinov, R. *Biochemistry* **2007**, *46*, 1205.
(23) Zanuy, D.; Jiménez, A. I.; Cativiela, C.; Nussinov, R.; Alemán, C. *J. Phys. Chem. B* **2007**, *111*, 3236.
(24) Tschöp, W.; Kremer, K.; Batoulis, O. J.; Bürger, T.; Hahn, O. *Acta Polym.* **1998**, *49*, 61.
(25) Theodorou, D. N. *Comput. Phys. Commun.* **2005**, *169*, 82.
(26) Reith, D.; Pütz, M.; Müller-Plathe, F. *J. Comput. Chem.* **2003**, *24*, 1624.
(27) Kremer, K. *Macromol. Chem. Phys.* **2003**, *204*, 257.
(28) Fukunaga, H.; Takimoto, J.-i.; Doi, M. *J. Chem. Phys.* **2002**, *116*, 8183.
(29) Padding, J. T.; Briels, W. J. *J. Chem. Phys.* **2002**, *117*, 925.
(30) Curcó, D.; Alemán, C. *Chem. Phys. Lett.* **2007**, *436*, 189.
(31) Saiz, L.; Klein, M. L. *Acc. Chem. Res.* **2002**, *35*, 482−489.
(32) Briels, W. J.; Mulder, P.; den Otter, W. K. *J. Phys.: Condens. Matter* **2004**, *16*, S3965.
(33) Goetz, R.; Lipowsky, R. *J. Chem. Phys.* **1998**, *108*, 7397.
(34) Goetz, R.; Gompper, G.; Lipowsky, R. *Phys. Rev. Lett.* **1999**, *82*, 221.
(35) Salmon, J.-B.; Colin, A.; Manneville, S. *Phys. Rev. Lett.* **2003**, *90*, Art. No. 228303.
(36) Shkulipa, S. A.; den Otter, W. K.; Briels, W. J. *Biophys. J.* **2005**, *89*, 823.
(37) Bagci, Z.; Jernigan, R. L.; Bahar, I. *Polymer* **2002**, *43*, 451.
(38) Micheletti, C.; Seno, F.; Maritan, A.; Banavar, J. R. *Comput. Mater. Sci.* **2001**, *20*, 305.
(39) Rojnuckarin, A.; Sangtae, K.; Subramanian, S. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 4288.
(40) Erman, B.; Bahar, I.; Jernigan, R. L. *J. Chem. Phys.* **1997**, *107*, 2046.
(41) Ulrich, P.; Scott, W.; van Gunsteren, W. F.; Torda, A. E. *Proteins* **1997**, *27*, 367.
(42) Gohlke, H.; Thorpe, M. F. *Biophys. J.* **2006**, *91*, 2115.
(43) Fernández, A. *Chem. Phys. Phys. Chem.* **1999**, *1*, 861.
(44) Doruker, P.; Jernigan, R. L.; Bahar, I. *J. Comput. Chem.* **2002**, *23*, 119.
(45) Erman, B. *Biophys. J.* **2001**, *81*, 3534.
(46) Haliloglu, T.; Bahar, I. *Proteins* **1998**, *31*, 271.
(47) Bruscoline, P. *J. Chem. Phys.* **1997**, *107*, 7512.
(48) Curcó, D.; Alemán, C. *J. Chem. Phys.* **2003**, *119*, 2915.
(49) Siepmann, J. I.; Frenkel, D. *Mol. Phys.* **1992**, *75*, 59.
(50) Main, E. R. G.; Lowe, A. R.; Mochrie, S. G. J.; Jackson, S. E.; Regan, L. *Curr. Opin. Struct. Biol.* **2005**, *15*, 464.
(51) Pye, V. E.; Tingey, A. P.; Robson, R. L.; Moody, P. C. E. *J. Biol. Chem.* **2004**, *279*, 40729.
(52) Govaerts, C.; Wille, H.; Prusiner, S. B.; Cohen, F. E. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 8342.
(53) Kreisberg, J. F.; Betts, S. D.; King, J. *Protein Sci.* **2000**, *9*, 2338.
(54) Jenkins, J.; Mayans, O.; Pickersgill, R. *J. Struct. Biol.* **1998**, *122*, 236.
(55) Curcó, D.; Alemán, C. *J. Comput. Chem.* **2004**, *25*, 790.
(56) Curcó, D.; Laso, M.; Alemán, C. *J. Phys. Chem. B* **2004**, *108*, 20331.
(57) Cornell, W. D.; 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.
(58) de Pablo, J. J.; Laso, M.; Suter, U. W. *J. Chem. Phys.* **1992**, *96*, 2395.
(59) Leontidis, E.; de Pablo, J. J.; Laso, M.; Suter, U. W. *Adv. Polym. Sci.* **1994**, *116*, 283.
(60) Curcó, D.; Alemán, C. *J. Chem. Phys.* **1994**, *121*, 9744.
(61) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926.
(62) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684.
(63) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327.
(64) 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.
(65) Press, W. H.; Teukolsy, S. A.; Vetterling, W. T.; Flannery, B. P. *Numerical Recipes in C*; Cambridge University Press: 1992.