Generic coarse-grained model for protein folding and aggregation

Tristan Bereau^{a)} and Markus Deserno^{b)}

Department of Physics, Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh, Pennsylvania 15213, USA

(Received 21 January 2009; accepted 19 May 2009; published online 18 June 2009)

A generic coarse-grained (CG) protein model is presented. The intermediate level of resolution (four beads per amino acid, implicit solvent) allows for accurate sampling of local conformations. It relies on simple interactions that emphasize structure, such as hydrogen bonds and hydrophobicity. Realistic α/β content is achieved by including an effective nearest-neighbor dipolar interaction. Parameters are tuned to reproduce both local conformations and tertiary structures. The thermodynamics and kinetics of a three-helix bundle are studied. We check that the CG model is able to fold proteins with tertiary structures and amino acid sequences different from the one used for parameter tuning. By studying both helical and extended conformations we make sure the force field is not biased toward any particular secondary structure. The accuracy involved in folding not only the test protein but also other ones show strong evidence for amino acid cooperativity embedded in the model. Without any further adjustments or bias a realistic oligopeptide aggregation scenario is observed. © 2009 American Institute of Physics. [DOI: [10.1063/1.3152842](https://doi.org/10.1063/1.3152842)]

I. INTRODUCTION

Proteins are the building blocks of biology. They are evolutionarily optimized heteropolymers, whose physical and material properties more often than not exceed what can be readily understood from conventional polymer physics reasoning, which derives much of its strength from uniformity, randomness, and the law of large numbers. In contrast, the complexity of proteins rests on the different physical and chemical properties of their monomers, the 20 physiological amino acids, and their intricate combination into what at cursory inspection only *seems* to be a random heteropolymer sequence. Moreover, the main interactions that drive their folding into intricate secondary, tertiary, and quaternary physiological structures are weak, comparable to thermal energy. The overall stability of a protein is perilously marginal,^{1,2} so proteins very often rely on cooperative effects to keep them in their native structure—one appealing reason for why they might be so much bigger than what their comparatively small active centers would make one suspect. This aspect also makes them very hard to model and to coarse grain, because it is extremely difficult to understand from first principles which interactions are essential and which can be approximated.

As for many other soft matter systems, the success of atomistic protein simulations is limited by available computer power. It is not so much the sheer number of atoms involved that poses the main challenge but rather the long equilibration time associated with a system that in a highly nontrivial phase space can so easily get stuck. Coarse-grained (CG) simulations intend to address this problem by lowering the level of resolution.³ A smaller number of quasiatoms, or “beads,” decreases the computational require-

ments and accelerates the speed of Monte Carlo (MC) or molecular dynamics (MD) simulations. It also smoothenes out the free energy landscape by reducing molecular friction, which artificially accelerates the dynamics and makes phase space both smaller and more navigable. However, compared to many other examples in soft matter, it is often precisely these small local interactions that contribute to the overall stability of the native state which makes the process of “throwing away detail for the benefit of the greater good” so much more daring. One might thus prefer to wait until computers have become even faster, and the progress in atomistic simulations looks indeed promising. Yet, the undeniable need to access really big systems of crucial relevance as well as the insatiable scientific interest to find what really matters in these systems both drive the development of new CG protein models.

The field of CG protein modeling is very diverse and has a rich history owing to a wide variety of problems to tackle, as well as length and time scales to look at.^{4–13} Various levels of resolution have been designed to study many different problems. On the coarser side of particle-based simulations, conformational effects of hydrophobic interactions were studied using lattice simulations.¹⁴ This is a very powerful tool that is still widely used when looking at large-scale cooperativity effects. Soon, off-lattice simulations were developed using one bead per amino acid with implicit solvent; famous examples are Gō models.¹⁵ This level of resolution allows for much more conformational freedom, which is key to structural studies. One underlying constraint in such models is that structure is biased toward the native configuration of the protein because the remaining degrees of freedom do not suffice to accurately represent the system’s phase space, including secondary structure motifs. Intermediate resolution models (more than one bead per amino acid) have been designed to investigate structural properties of proteins while emphasizing certain aspects. For instance, the recently intro-

^{a)}Electronic mail: bereau@cmu.edu.

^{b)}Electronic mail: deserno@andrew.cmu.edu.

duced MARTINI force field¹⁶ opts for a high resolution on the protein's side chains, while the backbone is represented by only one bead per amino acid. The force field was parametrized using partitioning coefficients between water and a (similarly CG) lipid membrane. By doing so, structural properties in transmembrane proteins can be accurately investigated (see, e.g., Ref. 17). Other models with a comparable overall resolution shift the emphasis (in terms of modeling detail) on the backbone instead of the side chain in order to look at structure and conformational properties without biasing the force field to the native configuration. Several force fields (see, e.g., Refs. 18–20) have been reported to fold *de novo* helical proteins. These models incorporate only a subset of amino acids, emphasizing their chemical effects (e.g., hydrophobic, polar, glycine residue).

Intermediate level resolution models have shown promising results in capturing local conformations and reproducing basic aspects of secondary structure recognition while gaining much computational efficiency compared to atomistic models. This is partly due to the removal of solvent, which allows for significant speedup, as water typically represents the bulk of a simulation in such systems. As a result, it is necessary to treat important solvent effects implicitly, as they are determining factors in a protein's conformation.

While α -helices are comparatively easy to obtain in such models, β -sheets and structures are more difficult to stabilize. There are several reasons for this. First, the enthalpic gain per amino acid is weaker compared to α -helices.²¹ Second, Yang and Honig²¹ showed that side chain–side chain interactions have a decisive role in sheet formation. And third, the stabilization energy contains a contribution from interactions between dipoles of successive peptide bonds that is usually neglected in simple models, yet it favors the β -over the α -structure.²² Apart from these local effects, the stability of extended conformations therefore also depends greatly on cooperativity. Other than stabilizing folds, this can also lead to peptide aggregation. Besides being an interesting physical problem, peptide aggregation is associated with countless biological processes. It also plays a crucial role in many diseases, ranging from sickle cell anemia²³ to Alzheimer.²⁴

In this work, we present a CG model of a four bead per amino acid model in implicit solvent. It differs from previously mentioned intermediate level force fields in several ways. First, by improving on full amino acid specificity it provides a more detailed free energy landscape. Second, protein folding is quantitatively probed by comparing our MD simulations with experimental data instead of the lowest energy structure that was sampled. Third, after tuning our force field with respect to one protein (in terms of tertiary structure reproduction), others are tested to understand how reliable this procedure is. Fourth, an important design criterion for our model is its ability to sample extended conformations. By requiring a balanced proportion between α -helical and β -extended conformations, our model aims to avoid a bias toward any particular secondary structure. Finally, we monitor the aggregation of small peptides (into β -sheets) to test whether a realistic aggregation scenario in the long-time and large length-scale regime can be achieved.

FIG. 1. Schematic figure of the local geometry of the protein chain. The solid beads represent one amino acid. Neighboring amino acid beads are represented in dashed lines.

In order to parametrize and test our force field as finely as possible, we systematically compare the performance of our CG model with experimental data. We hasten to add, though, that refining CG models is no attempt to compete with atomistic force fields. Such an endeavor strikes us as neither likely to succeed nor to be in line with the reasons one pursues coarse graining in the first place, namely, to gain a physical understanding of fundamental mechanisms and universals of complex molecular structures. However, in systems as delicate as marginally stable proteins a subtle local interaction can have a substantial global impact, and uncovering causations of this type is well within the scope of CG studies.

The paper is divided into several parts: the mapping scheme will explain how atomistic details were coarse-grained (CG) out, the different interactions as well as parameter tuning and simulation methods will be described, and finally several applications will show to what extent the model can reproduce structural properties.

II. MAPPING SCHEME

A. Overall geometry

An amino acid is modeled by three or four beads (Fig. 1). These beads represent the amide group N, central carbon C_α , carbonyl group C' , and (for nonglycine residues) a side chain C_β . The first three beads belong to the backbone of the protein chain, whereas the last one represents the side chain and is responsible for amino acid specificity. This high level of backbone resolution is necessary to account for the characteristic conformational properties underlying secondary protein structure. As far as reducing the number of degrees of freedom is concerned, this high resolution is regrettable, as the backbone is represented almost atomistically. Indeed, models that do not require the CG protein to represent local structure generally do away with most (if not all) backbone beads (e.g., Ref. 16). However, here we explicitly aim at a model that is capable of finding secondary structure by itself. This is, for instance, necessary in applications where this structure is known to change (e.g., misfolding, spontaneous aggregation) or not known at all.

B. Parameter values

Geometric parameters were taken from existing peptide models^{9,18,19} and are reported in Table I. Even though the spatial arrangement of the beads was fixed beforehand, the

TABLE I. Bonded interaction parameters used in the model. The dihedrals denoted with an asterisk were determined during parameter tuning (see Sec. IV). All parameters are expressed in terms of the intrinsic units of the system. k represents the interaction strength of Fourier mode n (see main text), with equilibrium value φ_0 . ω_{pro} refers to the ω dihedral around the peptide bond for a proline residue. The sign of the improper dihedral angle φ_0 is linked to the chirality of the isomer; the L form requires a negative sign. For each angular potential, only a single mode n was used.

	Bond lengths				
	NC_α		$\text{C}_\alpha\text{C}'$	$\text{C}'\text{N}$	$\text{C}_\alpha\text{C}_\beta$
r_0 (Å)	1.455		1.510	1.325	1.530
k_{bond} ($\mathcal{E}/\text{\AA}^2$)	300		300	300	300
	Bond angles				
	$\text{NC}_\alpha\text{C}_\beta$	$\text{C}_\beta\text{C}_\alpha\text{C}'$	NC_αC	$\text{C}_\alpha\text{C}'\text{N}$	$\text{C}'\text{NC}_\alpha$
θ_0 (deg)	108	113	111	116	122
k_{angle} ($\mathcal{E}/\text{deg}^2$)	300	300	300	300	300
	Dihedrals				
	ϕ^*	ψ^*	ω	ω_{pro}	Improper
k ($\mathcal{E}$)	−0.3	−0.3	67.0	3.0	17.0
n	1	1	1	2	1
φ_0 (deg)	0	0	180	0	∓ 120

van der Waals radii were left as free parameters. Following the abovementioned references, C_β was set at the location of the first carbon of the side chain (hence our nomenclature), directly connected to the backbone. Its location will generally not coincide with the center of mass of the atomistic side chain (which for larger and flexible side chains has no fixed position with respect to the backbone), but the concomitant substantial reduction in tuning parameters is necessary for our parametrization scheme, as we will see below.

All side chain beads have been given the same van der Waals radius, except for glycine, which is modeled without a side chain. This accounts for the biggest difference in the Ramachandran plot of amino acids, namely, the large flexibility of an achiral glycine residue, as opposed to the substantial chiral sterical clashes between all the others.²⁵ On the other hand, it does not represent the size differences between nonglycine residues and will thus likely cause problems if packing issues are important, e.g., inside globular proteins.

Both the location and the size of the side chain are thus modeled in an approximate and highly simplified way. Why not be more sophisticated? Since these degrees of freedom are accounted for, one might as well give them the best possible parameter values. Ideally this is indeed what one would like to do, but the catch is that the necessary tuning is very difficult. Having 20 different amino acids gives—in the worst case— $20^3=8000$ local Ramachandran plots for the (ϕ, ψ) angles between three consecutive amino acids. These would first need to be determined atomistically and then—via some suitable matching procedure—translated into CG side chain properties. Clearly, many obvious simplifications would be possible and the task is not nearly as daunting. The number of free parameters would nevertheless be substantially increased and their tuning would require both automated techniques and enormous computing resources. In contrast, in the present model we aim to keep the number of

free parameters as low as possible, such that judicious tuning by hand is still a viable option. We will see below that it is also successful. While optimization of side chain parameters will remain a long term goal, this is certainly not the point where to start.

Finally, amino acids that are in the middle of a protein chain form peptide bonds with their neighbors. This is not so at the ends of the chain, and the structure is slightly different. Nonetheless, we model the end beads identically.

C. Units

All lengths are measured in units of $\mathcal{L}$, which we choose to be 1 Å. For the energies we found it convenient to relate them to the thermal energy, since it is this balance which determines the overall protein conformation. We thus define the energy unit $\mathcal{E}=k_B T_r=1.38 \times 10^{-23} \text{ J K}^{-1} 300 \text{ K} \approx 4.1 \times 10^{-21} \text{ J} \approx 0.6 \text{ kcal mol}^{-1}$ as the thermal energy at room temperature.

Masses will be measured in the unit “ $\mathcal{M}$,” which is the mass of a single CG bead. We will assume all beads to have the same mass. An amino acid weighs on average 110 Da. By distributing mass equally among the four beads N, C_α , C_β , and C' , this gives an average mass of $\mathcal{M} \approx 4.6 \times 10^{-26} \text{ kg}$.

The natural time unit in our simulation is $\tau=\mathcal{L}\sqrt{\mathcal{M}/\mathcal{E}}$. Using the length, energy, and mass mappings from above, we find $\tau \sim 0.1 \text{ ps}$. This unit of time correctly describes the instantaneous dynamics of a fictitious CG bead-spring system (e.g., it leads to a value of the instantaneous velocity and associated kinetic energy that satisfies the equipartition theorem). However, it is crucial to understand that it does *not* measure the time which the real protein system requires to undergo the same conformational changes as observed in the simulation. The reason is that the reduction in degrees of freedom removes friction (smoothes the free energy land-

scape) and speeds up the motion through phase space. Translating τ into a reasonable measure for actual dynamics requires the determination of the associated speedup factor, which is typically accomplished by mapping an easily observable dynamic process between the experimental system and the CG simulation (such as diffusion). However, in the case of the conformational dynamics of proteins the identification of a suitable dynamic process is much less obvious. We defer this task to future work. It should be recalled that as far as equilibrium questions are concerned the precise time mapping is, of course, irrelevant.

III. INTERACTIONS

A. Bonded interactions

The local structure is constrained by bonded interactions. Bonds and angle potentials are chosen to be harmonic:

$$V_{\text{bond}}(r) = \frac{1}{2}k_{\text{bond}}(r - r_0)^2, \quad (1a)$$

$$V_{\text{angle}}(\theta) = \frac{1}{2}k_{\text{angle}}(\theta - \theta_0)^2. \quad (1b)$$

The spring constants k_{bond} and k_{angle} are set high enough to keep these coordinates close to their minimum (within $\sim 5\%$). Table I reports these parameters.

Up to thermal fluctuations bonds and angles are thus fixed. Flexibility of the overall structure enters through the dihedrals, the possibility to rotate around a chemical bond. In the case of proteins, two out of three backbone dihedrals are very flexible and are responsible for the diverse set of local conformations. These dihedrals are the ϕ and ψ coordinates, defined by the sets of beads $C'NC_\alpha C'$ and $NC_\alpha C'N$, respectively (see Fig. 1). They describe the angle between two planes (e.g., ϕ is the angle between the planes $C'NC_\alpha$ and $NC_\alpha C'$) and obey the following convention: taking any four beads 1, 2, 3, and 4 and looking along the vector from bead 2 to bead 3, the angle “0” will correspond to the conformation in which beads 1 and 4 point into the same direction (i.e., when they visually overlap). The rotation of plane 1, 2, 3 with respect to plane 2, 3, 4 away from this state defines the angle; the counterclockwise sense counts positive. Because the potential of rotation around the bond between sp^3 - and sp^2 -hybridized atoms has a rather low barrier compared to thermal energy at room temperature, we let the beads rotate freely. However, we will later include a contribution to the coordinates ϕ and ψ accounting for an effective nonbonded dipolar interaction (see below).

The third dihedral along the backbone chain, ω , defined by $C_\alpha C'NC_\alpha$, is located at the peptide bond (see Fig. 1). This bond corresponds to the rotation around two sp^2 -hybridized atoms, which involves a symmetric potential with two minima, separated by a rather high barrier. The two conformations, *cis* and *trans*, have angles of 0° and 180° , respectively. The *cis* conformation tends to be sterically unfavored for most amino acids, except for proline where there is no specific preference due to its special side chain linkage.

Generally, dihedrals can be written as a Fourier series in the rotation angle. Here we will restrict to a single mode and describe the interaction as

$$V_{\text{dih}}(\varphi) = k_n[1 - \cos(n\varphi - \varphi_{n,0})], \quad (2)$$

with coefficient k_n and phase $\varphi_{n,0}$. In our model we represent the peptide bond using only one minimum ($n=1$) centered around the *trans* conformation. In this case $\varphi_0 \equiv \varphi_{1,0}$ is the equilibrium orientation of the dihedral and $k \equiv k_1$ is the stiffness describing deviations from the equilibrium angle. For a peptide bond located right before a proline residue, we model the isomerization by a dihedral potential with two minima ($n=2$, $k \equiv k_2$), one at the *cis* conformation and the other one at *trans*. This allows for a more natural representation of the different conformations proline can take. Depending on the problem one is interested in, the energy barrier can be tuned to either freeze the isomerization or set to a low value to allow efficient sampling. We chose the latter in this work. This choice will of course affect the kinetics of the system.

The central carbon C_α not only links the backbone to the side chain; its sp^3 hybridization imposes a tilted orientation of the $C_\alpha C_\beta$ vector compared to the $NC_\alpha C'$ plane. Its four bonds are located at the vertices of a tetrahedron, linking the backbone atoms N and C' , as well as the C_β side chain and an extra hydrogen (not modeled by us). This has an important consequence, because a carbon atom with four different substituents is *chiral* and hence optically active. All amino acids except glycine exist as two different stereoisomers. The L form is realized in native amino acids: looking at the central carbon C_α , with the hydrogen atom pointing away, the isomer has L form if the three other substituents C' , C_β , and N are arranged in a counterclockwise fashion (“CORN rule”). This amino acid chirality is a central feature in proteins and their secondary structure, and we account for it by including an “improper dihedral” between the beads $NC_\alpha C'C_\beta$. This keeps a tilt between the backbone plane, $NC_\alpha C'$, and the plane intersecting the side chain with two backbone beads, $C_\alpha C'C_\beta$, such that all angles are correct and the CORN rule is satisfied. The interaction has the same form as other dihedrals, given by Eq. (2). The two stereoisomers only differ in the sign of the dihedral equilibrium angle φ_0 and can thus both be modeled.

B. Nonbonded interactions

Probably the biggest challenge in any coarse graining scheme is determining the nonbonded interactions. Unlike bonded interactions, their form is not intrinsically obvious and the system behavior depends very sensitively on them. In the following section every interaction introduced will require at least one free parameter that has to be determined by tuning. The key technical difficulty of this enterprise is that all parameters are typically highly correlated. Optimization is thus an intrinsically multidimensional problem and we therefore intend to limit the number of free parameters as much as possible. While one might envision “hands-off” tuning schemes in which optimization occurs in an automated fashion,²⁶ for the present problem we found this difficult to implement for two reasons: First, parameter variations often have a rather inconspicuous impact on target observables, and the determination of the right gradient in parameter space thus can require very substantial computer time. And

second, some optimization aims are hard to quantify in numbers and rather require judgment and choice—e.g., the question how one balances the quality of a local Ramachandran plot against global folding characteristics. While we are aware of several obvious extensions and improvements of our present model that would ultimately benefit from such automated fine-tuning, this is not the point where we wish to start.

1. Backbone

Steric interactions are closely linked to secondary and tertiary structures for two reasons: first, local interactions along the protein chain will shape the Ramachandran plot; second, contact between distant parts of the amino acid chain will determine protein packing on larger scales. In order to model a local excluded volume, we use a purely repulsive Weeks–Chandler–Andersen (WCA) potential

$$V_{\text{bb}}(r) = \begin{cases} 4\epsilon_{\text{bb}} \left[\left(\frac{\sigma_{ij}}{r} \right)^{12} - \left(\frac{\sigma_{ij}}{r} \right)^6 + \frac{1}{4} \right], & r \leq r_c, \\ 0, & r > r_c, \end{cases} \quad (3)$$

where $r_c = 2^{1/6} \sigma_{ij}$ and σ_{ij} is the arithmetic mean between the two bead sizes involved, following the Lorentz–Berthelot mixing rule. Just like the bead sizes, the energy ϵ_{bb} is a free parameter, though we use only one parameter for all backbone–backbone and backbone–side chain interactions, since for the WCA potential the energy scale is largely immaterial. Following the practice in atomistic simulations, we do not calculate excluded volume interaction between beads that are less than three bonds apart.

2. Side chain interactions

Amino acids differ in their water solubility. This can be quantified experimentally by measuring the partitioning of residues between water and a hydrophobic environment (e.g., Ref. 27). The ratio of densities of a residue in the two environments can be translated into a free energy of transfer from one medium to another. Hydrophobicity is one prominent cause for certain amino acids to attract. However, there are other reasons why residues interact (e.g., charges or hydrogen bonds between side chains) and this combination can be probed by statistical analyses of residue–residue contacts in proteins.^{28–32} One then arrives at a phenomenological interaction energy between any two residues A and B that depend on the number of close AB contacts that are found in a pool of protein structures. This mean-field approach (it averages over all neighboring contacts) not only contains information on the relative hydrophobicity of amino acids but also partially incorporates effects coming from additional interactions (e.g., salt bridges or side chain hydrogen bonds). In the absence of explicit solvent we represent this phenomenological cohesion by introducing an effective attraction (of standard Lennard-Jones 12-6 type) between C_β side chain beads, whose strength is mapped to such a statistical analysis of residue–residue contacts. Specifically, we used Miyazawa and Jernigan’s (MJ) statistical analyses²⁸ to extract a *relative* attraction strength between residues. To translate this into an *absolute* scale, one additional free parameter ϵ_{hp} is needed.

Miyazawa and Jernigan analyzed residue–residue contacts in crystallized proteins. By modeling interactions via square-well potentials, they obtained interaction strengths $\epsilon_{ij}^{\text{MJ}}$ for every i - j pair of residues. We reduced the resulting 20×20 interaction matrix further by deconvolving it into 20 interaction parameters ϵ_i (one for each amino acid), which approximately recreate all interactions as the geometric mean of the two amino acids involved, $\epsilon_{ij}^{\text{MJ}} \approx \epsilon_{ij} = \sqrt{\epsilon_i \epsilon_j}$, following the Lorentz–Berthelot mixing rule. Each term is then normalized,

$$\epsilon'_i = \frac{\epsilon_i - \min_k \epsilon_k}{\max_k \epsilon_k - \min_k \epsilon_k}, \quad (4)$$

such that the most hydrophilic residue has a weight of 0 and the most hydrophobic a weight of 1, and the normalized interaction contact is denoted $\epsilon'_{ij} = \sqrt{\epsilon'_i \epsilon'_j}$. Finally, we multiply this term by the overall interaction scale ϵ_{hp} . One limitation in varying the interaction strength of a Lennard-Jones potential is that a low ϵ'_{ij} will tend to flatten out the repulsive part of the interaction. This will, as a result, fade the excluded volume effect for certain side chain beads, which is likely to exacerbate packing problems in dense regions. To overcome this issue and keep the same excluded volume for all side chain beads, we model the overall interaction by using a Lennard-Jones potential for the attractive part linked to a purely repulsive WCA potential for smaller distances. We join the two potentials at the minimum value of the interaction in such a way that both the potential and its first derivative are continuous. Overall, the interaction will have the following form:

$$V_{\text{hp}}(r) = \begin{cases} 4\epsilon_{\text{hp}} \left[\left(\frac{\sigma_{C_\beta}}{r} \right)^{12} - \left(\frac{\sigma_{C_\beta}}{r} \right)^6 \right] + (\epsilon_{\text{hp}} - \epsilon'_{ij}), & r \leq r_c, \\ 4\epsilon_{\text{hp}} \epsilon'_{ij} \left[\left(\frac{\sigma_{C_\beta}}{r} \right)^{12} - \left(\frac{\sigma_{C_\beta}}{r} \right)^6 \right], & r_c \leq r \leq r_{\text{hp,cut}}, \\ 0, & r > r_{\text{hp,cut}}. \end{cases} \quad (5)$$

Relative (un-normalized) coefficients ϵ_i were calculated by minimizing the expression

$$\chi^2 = \frac{1}{N} \sum_{i,j \geq 1}^N \chi_{ij}^2, \quad (6)$$

where $\chi_{ij} = \epsilon_{ij}^{\text{MJ}} - \sqrt{\epsilon_i \epsilon_j}$, N is the number of matrix coefficients (210 independent elements in a 20×20 symmetric matrix), and the sum goes over all such elements. The normalized coefficients ϵ'_i that were obtained by simulated annealing followed by proper scaling [Eq. (4)] are reported in Table II.

Let us quantify the quality of our deconvolution and the suitability of our amino acid specific hydrophobic strength ϵ'_i . Recall that the correlation coefficient c between two data sets $\{X_i\}$ and $\{Y_i\}$ is defined as

TABLE II. Normalized scale of amino acid hydrophobicities ϵ_i using the Lorentz-Berthelot mixing rule for the cross terms, as well as relative and absolute error, $\Delta\epsilon_i$ and χ_{ij} , from the diagonal elements of the MJ matrix (see text for definition). Note that the side chain of glycine (marked with an asterisk in the table) is not modeled.

	Lys	Glu	Asp	Asn	Ser	Arg	Gln	Pro	Thr	Gly*	His	Ala	Tyr	Cys	Trp	Val	Met	Ile	Phe	Leu
	K	E	D	N	S	R	Q	P	T	G*	H	A	Y	C	W	V	M	I	F	L
ϵ_i ($\mathcal{E}$)	0.00	0.05	0.06	0.10	0.11	0.13	0.13	0.14	0.16	0.17	0.25	0.26	0.49	0.54	0.64	0.65	0.67	0.84	0.97	1.00
$\Delta\epsilon_{ii}$ ($\mathcal{E}$)	4.00	0.50	0.16	0.01	0.05	0.20	0.20	0.10	-0.01	-0.05	-0.11	0.00	0.03	-0.14	0.05	-0.02	0.01	0.02	0.04	0.05
χ_{ij} ($\mathcal{E}$)	-0.48	-0.45	-0.19	-0.02	-0.08	-0.31	-0.31	-0.17	0.02	0.11	0.35	0.01	-0.14	0.76	-0.24	0.12	-0.05	-0.12	-0.32	-0.38

$$c = \frac{1}{n} \sum_{i=1}^n \left(\frac{X_i - \bar{X}}{\sigma_X} \right) \left(\frac{Y_i - \bar{Y}}{\sigma_Y} \right), \quad (7)$$

where n is the number of data points in each set, $\bar{X}$ and $\bar{Y}$ are their averages, and σ_X and σ_Y are their standard deviations, respectively. Our inferred 210 ϵ_{ij} values and their original $\epsilon_{ij}^{\text{MJ}}$ counterparts have a correlation coefficient of 98%, which decreases by only three points when comparing the MJ matrix to the normalized interaction contacts ϵ'_{ij} . Moreover, the 20 individual values ϵ_i as well as the ϵ'_i have an 87% correlation with the experimental hydrophobic scale measured by Fauchere and Pliska.²⁷ Since the MJ matrix accounts for more than hydrophobicity, this further drop in the correlation coefficient is expected. However, its still relatively large value suggests that the hydrophobic effect is the dominant contribution to the MJ energies. This is the reason why we refer to the interactions (5) summarily as “hydrophobicity.” The fitting procedure gave a χ^2 value of 0.064, which translates into an average relative error $\Delta\epsilon = 0.25$ between coefficients along the diagonal of the MJ matrix, where this deviation is defined by $\Delta\epsilon_{ii} = \chi_{ii} / \epsilon_{ii}^{\text{MJ}}$. Even though most coefficients did not deviate more than 15% from the MJ matrix, lysine, the most hydrophilic residue, is off by a factor of 4. Various sets of parameters with a comparable χ^2 value showed equivalent correlation properties, even though deviations were located on different amino acids. This rules out the hypothesis of a systematic failure of our $\mathcal{N}^2 \rightarrow \mathcal{N}$ deconvolution.

It is possible to account for solvent effects in even further detail, for instance, by including the layering of water molecules around the solute into the effective potentials.³³ In our attempt to develop a simple force field and only keep a few important aspects of protein interactions and in view of the approximation already made, we decided against such local details.

3. Hydrogen bonds

Since our model does not contain any electrostatics, it is necessary to model hydrogen bonds implicitly as well. The interaction depends on the relative distance and orientation of an amide and a carbonyl group. A real amide group is composed of a nitrogen with a hydrogen, whereas the carbonyl group has a carbon double-bonded to an oxygen. The hydrogen bond is favored when the N, H, and O atoms are aligned. Several interaction potentials for hydrogen bonding have been proposed in the literature.^{11,18,19,34–36} For its simplicity and corresponding CG mapping, we follow Irbäck *et al.*¹⁸ by using a radial 12-10 Lennard-Jones potential combined with an angular term,

$$V_{\text{hb}}(r, \theta_N, \theta_C) = \epsilon_{\text{hb}} \left[5 \left(\frac{\sigma_{\text{hb}}}{r} \right)^{12} - 6 \left(\frac{\sigma_{\text{hb}}}{r} \right)^{10} \right] \times \begin{cases} \cos^2 \theta_N \cos^2 \theta_C, & |\theta_N|, |\theta_C| < 90^\circ, \\ 0, & \text{otherwise,} \end{cases} \quad (8)$$

where r is the distance between the two beads N and C', σ_{hb} is the equilibrium distance (Table III), and θ_N is the angle

TABLE III. Nonbonded interactions. The length σ represents the diameter of a bead. Most parameters were determined after parameter tuning, except the ones denoted by an asterisk. See Sec. IV.

Backbone excluded volume			
σ_N (Å)	σ_{C_α} (Å)	$\sigma_{C'}$ (Å)	ϵ_{bb} (ε)
2.9	3.7	3.5	0.02
Hydrophobicity			
σ_{C_β} (Å)	ϵ_{hp} (ε)	$r_{hp,cut}$ (Å)	
5.0	4.5	10*	
Hydrogen bonding			
σ_{hb} (Å)	ϵ_{hb} (ε)	$r_{hb,cut}$ (Å)	
4.11*	6	8*	

formed by the atoms HNC' and θ_C corresponds to the angle NC'O (Fig. 2). The main motivation for using a power of 10 instead of 6 in the Lennard-Jones potential is a narrower confinement of the hydrogen bond length. Since our model does not represent hydrogens and oxygens, these particle positions were calculated via the local geometry of the backbone. Any NC' pair can form a hydrogen bond, except if N belongs to proline, since its side chain connects to the preceding amide on the backbone. The hydrogen bond leads to one more free parameter, its interaction strength ϵ_{hb} .

4. Electrostatics

There is no explicit treatment of side chain charges in the force field. Specifically, we do not model the interaction between charged residues. However, this piece of information is partially included in the MJ matrix, as the method is based on statistical analysis of residue-residue distances. The electrostatic interaction involved between two charged residues will be implicitly sampled, and its effect reflected in the interaction coefficient. Nevertheless, an explicit treatment of charged residues would allow one to look into properties that depend on the environment's pH or ionic strength. For a solution that has a high salt concentration (e.g., under physiological conditions), ions are able to screen most of the electrostatics, such that a Debye-Hückel potential would be appropriate to model this interaction. By compensating for the difference in binding energy for all the coefficients involved, one could disentangle charge effects from the MJ matrix. This, however, has not been done in the present model.

FIG. 2. Schematic figure of the hydrogen bond interaction. The light beads (H and O) are not explicitly modeled in the simulation; their positions are inferred from their bonded neighbors.

5. Dipole interaction

The interactions described above were sufficient to fold and stabilize α -helices but not β -sheets. Chen *et al.*²² pointed out that there is an important contribution usually neglected in generic models: carbonyl and amide groups at the peptide bond form dipoles that interact with each other. Mu and Gao³⁴ showed that the nearest-neighbor interaction is enough to favor β over α content. Effectively, all dipoles along a helix are parallel compared to more favorable antiparallel neighboring dipoles on a β -sheet.

From a computational standpoint, a dipole-dipole interaction,

$$V_{dd}(\mathbf{p}_i, \mathbf{p}_j) = \frac{\epsilon_{dd}}{r^3} [\mathbf{p}_i \cdot \mathbf{p}_j - 3(\mathbf{p}_i \cdot \hat{\mathbf{r}})(\mathbf{p}_j \cdot \hat{\mathbf{r}})], \quad (9)$$

between two dipoles $\mathbf{p}_i$ and $\mathbf{p}_j$ at a distance $\mathbf{r}$ from each other is inconvenient because it is long ranged. However, nearest-neighbor dipoles are all separated roughly by the same distance, as all amino acids have the same backbone geometry. All dipoles also have the same magnitude, as they are formed from the same atoms. Therefore, the key component of the interaction lies in the relative orientation between dipoles and not on their magnitude or relative distance. Successive dipoles therefore capture the orientation of the local backbone geometry. Two neighboring dipoles will effectively measure the angle difference between the two planes $C'_{i-1}N_iC_{\alpha,i}$ and $C_{\alpha,i}C'_iN_{i+1}$, where the index keeps track of the amino acid involved (see Fig. 1). As the effect is completely localized and only affects the conformation of the amino acid backbone, we treat this interaction as a *bonded* one by effectively biasing the dihedral potentials of ϕ and ψ . To do so, we first calculated Eq. (9) for all combinations of dihedral angles with a 5° resolution. The result is plotted on the upper part of Fig. 3. The (sterically forbidden) central part of the plot was removed to emphasize local differences in allowed regions.

In order to be efficient, the potential should decouple along the two coordinates, i.e., it must be expressible as a sum $U(\phi, \psi) \approx U(\phi) + U(\psi)$. We use a single cosine function centered around $\phi, \psi = 0$ with identical amplitude along both coordinates to approximate the neighboring dipole potential (Fig. 3, bottom). Higher modes in the series have shown to be negligible:

$$V_{dip}(\phi, \psi) = k_{dip}[(1 - \cos \phi) + (1 - \cos \psi)]. \quad (10)$$

The value of the optimally tuned free parameter k_{dip} is reported in Table I. The discrepancy between the plots is due to the enforced decoupling of the two coordinates ϕ and ψ . Even though the final result looks rather inaccurate on the whole domain of the function, it nevertheless recreates the one important effect of the interaction: the β region is more favored than the α region (see labels in Fig. 3). Moreover, the quality of the fit should only be tested along the physically relevant domains of the Ramachandran plot, most notably the α and β regions. In this sense, Eq. (10) makes for a good approximation of the dipole interaction and is enough to recreate the physics that favors β regions.

FIG. 3. (Color) Map of the nearest-neighbor dipole-dipole interaction for all sets of dihedral angles ϕ and ψ (top) and the decoupled Fourier series approximation (bottom). The central part of the upper plot was not reproduced in order to emphasize local difference in other regions of the plot (as can be seen in Fig. 4, this anyway is a sterically hindered region). Sterically favored regions of the plot are circumscribed by a thick line, in addition to labels of α and β regions. The two graphs were shifted and scaled for comparison.

IV. PARAMETER TUNING

There are various ways CG force fields can be parametrized. For instance: only allowing nonbonded interactions between native contacts (Gō-type models),¹⁵ partitioning measurements of amino acids between water and a hydrophobic medium,¹⁶ structure-based coarse graining based on all-atom simulations,⁴ and knowledge-based potentials which intend to optimize parameters by using large pools of existing structures.¹²

Parameter tuning in top-down CG models aims at reproducing a selected subset of structural or energetic system properties. Since these parameters tend to be correlated, a given set needs to be tested at all scales. In our force field, *local* conformations are tuned to reproduce probability distribution functions of dihedral angles, which by a slight extension of standard terminology we also called Ramachandran plots (see Sec. IV A). Large-scale (*global*) properties are targeted by studying folding events of a helical peptide (see Sec. IV B). The final set of parameters was identified as the one we felt most capable in reproducing properties on both levels. The physical conditions (temperature, density, etc.) of the force field will be set by the systems we try to match.

Note that on the global level we tune our parameters using only one protein. Of course, adding more proteins into

TABLE IV. Table of free parameters in this CG model. The main test that was used to determine a given parameter is denoted in the second column.

Free parameters	Tuning method
$\sigma_N, \sigma_{C_\alpha}, \sigma_{C'}, \sigma_{C_\beta}, \epsilon_{bb}$	Ramachandran plot
$\epsilon_{hp}, \epsilon_{hb}, k_{dip}$	Folding characteristics

the “training set” would incorporate more information, presumably leading to a better founded force field. There exist various successful parametrization schemes that rest on large ensembles of data.^{37,12,38} This, however, needs to be balanced against the need to test how reliable a given force field handles proteins that were not part of its training set—a point we deemed more relevant.

Table IV lists the eight free parameters that need to be determined. Because of time constraints and to obtain some intuition and feeling of each interaction involved, we made a point of having our model tunable by hand, which is why we required the number of free parameters to be as low as possible. As explained above, this is the main reason why we decided against amino acid specific bead sizes. Adding even more free parameters, on top of being time consuming, would make it difficult to obtain a consistent set of parameters that would correctly describe both local and global conformations. Different bead sizes would involve different Ramachandran plots, and all backbone parameters would need to be consistent throughout.

The free parameters were tuned by trying to constrain parameter space as much as possible, for instance, by eliminating unphysical behavior (e.g., sterically hindered β region in the Ramachandran plot or too much helicity in secondary structures). Combining both local and global tests was enough to settle for a satisfying set of parameters using the constraint that the dipole interaction strength was maximized. Even though this may sound arbitrary when looking for a realistic α/β content ratio, it turned out to be very difficult to use β structures as tests because they are so weakly stabilized. Indeed, we have found that the final set point is still not strong enough to fully stabilize β -sheets during folding events (see below). This shows that maximizing the dipole interaction strength in this model does not lead to oversampling of β content but merely sampling as much extended conformations as possible before the force field cannot stabilize helical structures anymore. Other simple tests can be used to exclude regions of parameter space. For example, a hydrogen bond interaction that is too strong will lead to proteins that fold into one long helix. Too strong hydrophobic interactions will collapse proteins into globules, even native elongated helical structures. Bead size parameters were initially taken from other CG models (e.g., Refs. 9, 18, and 19) and tuned as little as possible to recreate enough sampling of α/β content while suppressing sterically hindered regions.

As for any physical system, the representative sampling of its phase space is prerequisite to obtaining accurate thermodynamic information. Different schemes have been developed to characterize and estimate the population of thermodynamic states.^{39–42} In the present case, thermodynamic

calculations were performed by combining parallel tempering⁴³ with the weighted histogram analysis method (WHAM).^{44–46} The main idea is to combine energy histograms from canonical simulations at various temperatures in order to reconstruct the density of states of the system. The information contained in these histograms is used to calculate a consistent set of free energy differences between simulations. Converging these free energies was done by using a recently developed highly efficient algorithm.⁴⁷ Once the density of states is reconstructed, one can obtain continuous approximations to all thermodynamic observables. By combining WHAM with parallel tempering, we effectively improve sampling by reducing correlations between data points.

A. Local conformations: Ramachandran plot

The Ramachandran plot⁴⁸ records the occurrence and frequency of successive (ϕ, ψ) angles in a protein. Since backbone flexibility is almost exclusively due to these two coordinates, the Ramachandran plot is an ideal reporter of local (secondary) structure: α -helices and β -sheets belong to peaks in different regions of the plot. And since proteins are highly constrained systems, low energy points on the Ramachandran plot are rather well localized. Their accurate sampling is therefore prerequisite to the formation and stabilization of reliable structures on larger scales. In the following we will be concerned with the (thermal) distribution of the (ϕ, ψ) angles surrounding some particular amino acid and, in a slight stretch of standard terminology, also refer to this probability density as a Ramachandran plot.

The free parameters that most directly constrain the Ramachandran plot are the different bead sizes $(\sigma_N, \sigma_{C_\alpha}, \sigma_{C'}, \sigma_{C_\beta})$ and, to a lesser extent, the excluded volume energy prefactor ϵ_{bb} . We disentangled hydrogen bond and hydrophobicity effects from the Ramachandran plot by studying systems made of only three amino acids. From a steric point of view we only distinguish between glycine and nonglycine amino acid, by either not having a side chain bead at all (Gly) or by using a generic bead representing the 19 other amino acids (Ala, for the sake of concreteness). It is then sufficient to study the two Ramachandran plots of Gly-Gly-Gly and Gly-Ala-Gly tripeptides, the smallest systems that contain relevant information on successive dihedral angles ϕ and ψ . The reason why we surround the amino acid of interest with two Gly is to avoid hydrophobic interactions between neighboring side chains. As a result, we solely probe steric effects. The Ramachandran plots derived from the final set of parameters are shown in Fig. 4 as free energy plots obtained from using parallel tempering at temperatures $k_B T / \epsilon \in \{0.5, 0.7, 1.0, 1.3, 1.6, 1.9, 2.2, 2.5\}$ and reconstructing the density of states with WHAM. The free energy plot is calculated at our reference temperature $k_B T / \epsilon = 1$. The shading represents the free energy difference with respect to the lowest conformation, in units of $k_B T$. Notice the inherent asymmetry in the Gly-Ala-Gly system, which reflects the chirality of the α -carbon. Both α -helix $(-60^\circ, -60^\circ)$ and β -sheet $(-60^\circ, 130^\circ)$ regions are well populated, in agreement with Ho *et al.*⁴⁹ Proper balance and connectivity be-

FIG. 4. (Color) Free energy plots of tripeptides Gly-Gly-Gly (top) and Gly-Ala-Gly (bottom) as a function of successive dihedrals ϕ and ψ , calculated at our reference temperature $T = 1\epsilon/k_B$. The coloring represents the free energy difference with the lowest conformation, in units of $k_B T$.

tween the two regions are crucial for protein folding. This is tuned by the bead sizes and excluded volume energy but also depends on the dipole interaction k_{dip} (see below). The achiral glycine, on the other hand, has no side chain and permits many more conformations. One therefore often finds glycine residues at the ends of helices.

A particular challenge was the fact that we model neither the amide-hydrogen nor the carbonyl-oxygen explicitly, yet their steric effects strongly shape the Ramachandran plot.⁴⁹ This required subtle adjustments of the bead sizes of the N and C' atoms compared to their conventional van der Waals radii.

A poor sampling of local conformations can thwart the formation of realistic secondary structure. Moreover, the *relative* weight of characteristic regions of the Ramachandran plot determines to a large extent the α/β content. Even though the analysis of abovementioned tripeptides accounts for steric effects and the dipole interaction, it does not consider hydrogen bonds and side chain interactions which are also important to stabilize secondary structure. For this reason it is difficult to ascertain the quality of conformational distributions without studying larger structures.

B. Folding of a three-helix bundle

In this section we study full size proteins to parametrize large-scale interactions. We used proteins found in the Protein Data Bank that were resolved experimentally in aqueous solvent.

TABLE V. Structure and amino acid sequence of all proteins studied in this paper.

PDB ID	Structure	Sequence
2A3D	Three-helix bundle	MGSWAEFKQLAAIKTRLQALGGSEAEAAFEKEIAA... FESELQAYKKGKNPEVEALRKEAAAIRDELQAYRHN
1LQ7	Three-helix bundle	GSRVKALEEKVKALEEKVKALGGGGRIEELKKKW... EELKKKIEELGGGGEVKKVEEEVKKLEEEIKKL
1P68	Four-helix bundle	MYGKLNLDLEDLQEV LKLNHKNWHGGKDNLHDVDNHL... QNVIEDIHDFMQGGGSGGKLQEMMKEFQQVLDEL... NNHLQGGKHTVHHIEQNIKEIFHHLEELVHR
2JUA	Four-helix bundle	MYGKLNLDLEDLQEV LKHVNQHWQGGQKNMNKVDHHL... QNVIEDIHDFMQGGGSGGKLQEMMKEFQQVLDEL... KQQLQGGDNLHNHVENIKEIFHHLEELVHR
1R69	Five short helices	SISSRVKSKRIQLGLNQAELAQKVGTTQQSIEQLENG... KTKRPRFLPELASALGVSDWLLNGTSDSNVR
1K8B	Two helices and a four stranded β -sheet	EILIEGNRTIIRNFRELAKAVNRDEEFFAKYLLKETG... SAGNLEGGRLILQRR
1K43	β -hairpin	RGKWTYNGITYEGR
	β -hairpin	VVVVV ^D PGVVVVV

Our choice of reference protein is constrained by the limitations of our model. For instance, salt or disulfide bridges cannot yet be represented and should thus play no role in the reference protein either. Also, it was important to start with a simple structure rather than a globular protein for which packing and cooperativity are more important. Following Irbäck *et al.*¹⁸ and Takada *et al.*,¹⁹ we also tuned our force field on a *three-helix bundle*. Direct comparisons with their models are difficult, though. First, these authors do not incorporate specificity on every amino acid and only represent a few amino acid types (e.g., hydrophobic, polar, glycine residue). Second, they only compared their simulations to the lowest energy structure found during the simulation rather than experimental data. In contrast, we use the *de novo* protein 2A3D (73 residues) and systematically compare our results with the real structure resolved experimentally (using NMR).⁵⁰ The amino acid sequence is given in Table V. A similar protocol was followed by Favrin *et al.*²⁰ in order to study a different three-helix bundle (1BDD).

A first attempt in tuning parameters consisted of simulating proteins starting from their native structure. Testing for stability is a rapid means to constrain parameter space but not sufficiently so as to actually determine their values. This is consistent with the picture of a deep funnel-like free energy landscape:⁶ the free energy minimum of a native state is sufficiently deep compared to unfolded states that a folded protein is very stable against force field parameter variations. Further tuning was therefore mainly achieved by studying folding events using a set of trial runs with different parameters. Observation of three-dimensional structures with VMD (Ref. 51) was well suited to characterize simulations. The software was also used to render protein images in this paper.

Folding was studied in the following way: The only input into our simulations was the sequence of amino acids and the temperature. The initial conformation (determined by the collection of dihedral angles ϕ and ψ) was chosen randomly,

and the integration started by warming up nonbonded interactions to relax high energy steric clashes. We used parallel tempering for all simulations to avoid kinetic traps. Structural observables were measured at $k_B T = 1\mathcal{E}$, the temperature at which the force field was tuned. Simulations were set at eight different temperatures: $k_B T/\mathcal{E} \in \{1.0, 1.1, 1.2, 1.3, 1.4, 1.6, 1.9, 2.2\}$. MC swaps between different temperatures were attempted every 10τ , the average acceptance rate was around 10%. We tested convergence to a global minimum by checking that different initial conditions consistently equilibrate to the same structure. A combination of thermodynamic and kinetic studies (see below) will allow us to show two important features. First, the temperature used for parameter tuning, $k_B T = 1\mathcal{E}$, is below the folding temperature T_f of 2A3D, above which the unfolded conformation becomes the most stable state. Second, $k_B T = 1\mathcal{E}$ is above the glass transition temperature T_g , below which the energy landscape becomes very rugged and creates severe kinetic traps. It was indeed possible to observe folding events in conventional (i.e., not using parallel tempering) simulations within this range of temperature.

Quantitative comparison between the CG and the experimental structures can be made by calculating the root-mean-square deviation (RMSD) between corresponding α -carbons on the two chains (after optimal mutual alignment). Figure 5 reports the RMSD of a protein in the lowest ($k_B T = 1\mathcal{E}$) replica of a parallel tempering MD run as a function of time, using the RMSD trajectory tool within the VMD package.⁵¹ These results were obtained with the parameters reported in Tables I–III. The average error between the equilibrated simulation and the NMR structure is around 4 Å after about 100 000 τ and at $k_B T/\mathcal{E} = 1$, temperature at which the native conformation represents the free energy minimum. A superposition of the simulated structure with the experimental one is shown in Fig. 6. The STRIDE algorithm⁵² was used to as-

FIG. 5. (Color online) RMSD of the CG proteins 2A3D (full line) and 1P68 (dashed line) compared with experimentally resolved structures. Both simulations were run at $T=1\mathcal{E}/k_B$.

sign secondary structure. Overall the conformation is very well reproduced considering that we have a resolution of only four beads per amino acid and that no *a priori* knowledge of secondary/tertiary structure was provided to the force field. Helix regions had formed at the right place, and amino acids were arranged in order to bury hydrophobic beads between the three helices, away from the implicit solvent.

To characterize the stability of this protein, we also performed thermodynamic calculations using WHAM and parallel tempering at the temperatures $k_B T/\mathcal{E} \in \{0.8, 0.9, 1.0, 1.1, 1.2, 1.3, 1.4, 1.6, 1.9, 2.2\}$. By reconstructing the density of states, we can estimate the folding temperature $k_B T_f \approx 1.2\mathcal{E}$, the point at which the folded and unfolded states are equally populated. This gives a measure of the stability of the system: below T_f the native state is the most likely conformation. In Fig. 7, we plot the free energy below, at, and above the folding temperature as a function of the nativeness order parameter Q as introduced by Takada *et al.*¹⁹ It measures the distance r_{ij} between pairs i and j of C_α beads between the NMR data and CG simulations:

$$Q = \left\langle \exp \left[-\frac{1}{9\sigma^2} (r_{ij}^{\text{NMR}} - r_{ij}^{\text{CG}})^2 \right] \right\rangle_{ij}, \quad (11)$$

where the average goes over all pairs ij . The folded conformation lies in the basin $Q \gtrsim 0.6$ whereas all unfolded conformations

FIG. 6. (Color) Equilibrated structures of (a) 2A3D and (b) 1P68 sampled at $T=1\mathcal{E}/k_B$. Superposition of simulated structure (opaque) with experimental data (transparent) is displayed. The STRIDE algorithm (Ref. 52) was used for secondary structure assignment (thick ribbons represent α -helices on the figure).

FIG. 7. (Color online) Free energy profile as a function of a nativeness order parameter Q below ($T=1.1\mathcal{E}/k_B$), at, and above ($T=1.3\mathcal{E}/k_B$) the folding temperature $T_f=1.2\mathcal{E}/k_B$.

occur for $Q \lesssim 0.5$. It should be noted that all three curves in the graph have been calculated by using the same reference point, meaning that the vertical shift between curves accounts for the free energy difference in going from one temperature to another. The folding temperature is close to $1.2\mathcal{E}/k_B$. To make sure the model is also able to sample this important part of phase space in conventional simulations, we provide a stability run at the folding temperature starting from a random conformation. It can be seen that the system repeatedly switches between folded and unfolded states and roughly spends as much time in either one (Fig. 8).

In 13 out of 15 independent parallel tempering simulations the protein folded to the native state at a temperature $T=1\mathcal{E}/k_B$. However, the folding time varied substantially between different simulations. The kinetics of folding of this protein was studied by running conventional simulations at various temperatures. For each temperature $k_B T/\mathcal{E} \in \{0.7, 0.8, 0.9, 1.0, 1.1, 1.2\}$ we ran ten simulations and measured the average time it took to fold the protein to its native conformation, if it ever did in the time scale of the simulation ($2 \times 10^6 \tau$). The results are reported in Fig. 9. Temperatures $0.7\mathcal{E}/k_B$ and $0.8\mathcal{E}/k_B$ did not yield a single folding event, suggesting the onset of glassy behavior.^{6,53} The glass transition temperature T_g can be estimated following a simple pragmatic scheme suggested by Socci and Onuchic:⁵³ it is the temperature where the mean folding time is the average of the minimum folding time $\tau_{\min}$ (lowest point in the graph) and the largest time scale one is willing to

FIG. 8. (Color online) Conventional (i.e., not using parallel tempering) simulation of 2A3D at $T=1.2\mathcal{E}/k_B$. The nativity parameter Q is plotted against time. The protein alternates between folded ($Q \gtrsim 0.6$) and unfolded conformations ($Q \lesssim 0.5$).

FIG. 9. (Color online) Kinetic studies of the 2A3D three-helix bundle CG protein. The average folding time t_f is plotted against temperature. For temperatures ranging from $T=0.7\epsilon/k_B$ to $T=1.2\epsilon/k_B$, about ten simulations were run and we measured the first passage time to the native state. The line represents the average between the minimum folding time and the time scale of the simulation. This can be used to estimate the glass transition temperature (see text).

invest in the simulation, $\tau_{\max}$ (highest boundary in the graph): $\tau_g = (\tau_{\min} + \tau_{\max})/2$. This average time is plotted as a horizontal line in the graph. One can then estimate what temperature this folding time corresponds to (Fig. 9). In our case, we can safely assume that $T_g < 0.9\epsilon/k_B$, meaning that the protein does not experience glassy behavior when simulating at our reference temperature $T=1\epsilon/k_B$. Moreover, combining results from thermodynamic calculations and kinetic studies shows that there is a range of temperatures $T_g < T < T_f$ in which the system is not experiencing glassy behavior but is still “cold” enough such that the native state is the most stable conformation.

Irbäck *et al.*¹⁸ as well as Takada *et al.*¹⁹ reported a degeneracy in the CG structures of their helix bundles: there are two ways three helices can pack (see Fig. 10), and their models were not able to discriminate the two different tertiary structures. NMR experiments on 2A3D found a ratio between clockwise and counterclockwise topologies of several percentage, leading to a free energy difference of a few $k_B T$ at room temperature.⁵⁴ From 15 independent simulations we ran, one of them did not fold within 300 000 τ , and 13 converged to the NMR structure—a counterclockwise topology [Fig. 10(a)]; only one had the other topology [(illustrated in Fig. 10(b)). While it is encouraging to see that our model is able to distinguish these topologies, it is not guaranteed that this will work equally well for other proteins.

V. SIMULATION DETAILS

MD simulations were performed with the ESPRESSO package.⁵⁵ Simulations in the canonical ensemble (NVT)

FIG. 10. Schematic figure of the two possible topologies in forming a three-helix bundle. (a) The native fold of protein 2A3D corresponds to a counterclockwise topology and (b) that of 1LQ7 is clockwise.

were achieved by using a Langevin thermostat with friction constant $\Gamma = \tau^{-1}$. The temperature was expressed in terms of the intrinsic unit of energy, ϵ . The force field is parametrized in order to reproduce a temperature of $T=300$ K. The integration time step used for all simulations is $\delta t=0.01\tau$.

VI. APPLICATIONS AND TESTS

A. Folding

All simulations mentioned from this point onward have not been part of the parameter tuning training set. They come out as independent checks and features of the force field. Thermodynamic and kinetic studies were not performed for the different proteins of this section. Here, we study the equilibrium conformations of various sequences at a temperature of $k_B T/\epsilon=1$, which lies between T_g and T_f for our reference protein, 2A3D. In this respect, we expect to avoid glassy behavior for similarly complex proteins whose native state is folded.

In order to test the folding features of the model, we first studied another *de novo* three-helix bundle, 1LQ7. Even though the fold is very similar to 2A3D, it has 67 amino acids and a completely different primary sequence. Also, the native structure, obtained from NMR,⁵⁶ has the *opposite* topology (clockwise) compared to 2A3D. From ten independent parallel tempering runs, 300 000 τ long each, one of them did not fold within this amount of time (helices formed but did not arrange properly). Out of the nine remaining structures, five folded consistently to the native clockwise topology [Fig. 10(b)] and four to the other one [Fig. 10(a)]. It should be noted that this sequence had been designed such that its native structure leads to favorable salt-bridge interactions.⁵⁶ As we do not incorporate electrostatics (and thus salt bridges) explicitly, we expect the CG model to have difficulties in discriminating between the two tertiary structures.

In order to further probe the folding features of different α -helical rich folds, we studied a four-helix bundle, 1P68, consisting of 102 amino acids.⁵⁷ Even though the secondary structure is overall rather similar to the abovementioned three-helix bundle, the tertiary structure and amino acid sequence are completely different. Again, the reference structure is taken from experimental data.⁵⁷ From six independent parallel tempering runs, each 600 000 τ long, our force field successfully folded the protein into a four-helix bundle for every simulation except one, which did not have time to properly align its fourth helix. The RMSD is shown in Fig. 5 for a simulation which converged to the right topology. As can be seen, the RMSD went below 4 Å, which, again, is very satisfactory considering the level of resolution and the complete absence of structure bias in the force field. It should be noted that what appears as large fluctuations on the graph are actually frequent MC swaps between replicas of the parallel tempering ladder. Even though their potential energies are comparable (which is the reason why they swap temperatures), their structures are fairly different, as can be seen on the RMSD plot. Just as in the three-helix bundle case, this protein can fold into several different topologies. Out of the five simulations which converged to a four-helix

bundle tertiary structure, two of them represented the NMR topology. RMSD values for other topologies ranged between 5 and 8 Å. A snapshot of the equilibrated structure is shown in Fig. 6(b).

Also a second *de novo* four-helix bundle was used to test the force field. Even though the tertiary structure resembles the abovementioned 1P68, the amino acid sequence of 2JUA is completely independent (though it also has 102 amino acids) and the topology is different. Out of three independent runs, all of them successfully folded in a four-helix bundle structure within 600 000 τ by comparing qualitatively the CG protein with the NMR structure.⁵⁸ However, none of them converged to the right topology.

Our model has proven very efficient in finding the equilibrium conformation of various helical structures, up to small deviations, and independent of their tertiary structure (i.e., number of helices) or sequence of amino acids. The fact that none of these proteins was part of the parameter tuning strongly indicates that our CG model captures important aspects of protein physics.

The limits of the model were reached when simulating globular proteins, such as 1R69 (Ref. 59) and 1K8B.⁶⁰ The chain collapsed into a molten globule, but the arrangement of secondary structures (collections of α -helices and β -sheets) was not accurately reproduced, leading to an incorrect tertiary structure. This suggests a missing sufficiently deep free energy minimum, most likely due to the limitations of the CG model in terms of cooperativity and realistic packing (recall that all side chains have the same bead size). The RMSD values did not drop below 10 Å.

Stabilizing a single β -hairpin in small proteins is difficult because this relies on very weak interactions. We simulated the *de novo* 1K43 peptide for 300 000 τ . It consists of 14 residues and forms a β -hairpin in water.⁶¹ Our model is not able to stabilize it. The simulation shows a high tendency to form an α -helix, where 40% of all conformations are helical, whereas only 2% are extended (β -sheet-like). However, the CG model can successfully fold a designed β -hairpin, sequence V₅^DPGV₅, which contains a D-proline in order to sterically favor hairpin formation.⁶² This peptide has been recently characterized using atomistic⁶³ and structure-based CG simulations.⁶⁴

B. Aggregation

Gsponer *et al.*⁶⁵ recently reported atomistic simulations of small aggregation events in water. Heptapeptides GNNQQNY from the yeast prion protein Sup35 were shown to form β -sheet aggregates. These authors did a quantitative analysis of the number of 2 and 3 aggregates in the system at room temperature.

We studied the abovementioned scenario by simulating 15 identical peptides in a box of size of 40 Å, without matching density with the atomistic run. Indeed, while Gsponer *et al.* simulated their system in a restricted sphere of 150 Å diameter and applying forces to constrain the system in the center, we set periodic boundary conditions in a cubic box. Even though this represents a rather dense system in order to drive aggregation, we checked that similar structures were

FIG. 11. (Color online) Specific heat of 15 GNNQQNY peptides in a cubic box of size of 41 Å. The peak around $T=0.95\epsilon/k_B$ separates a low temperature phase, rich in high- β content aggregates, from a high temperature phase where no aggregates form.

sampled when simulating more dilute systems. Initial configurations were chosen randomly, and we ran parallel tempering simulations at temperatures $k_B T/\epsilon \in \{0.7, 0.8, 0.85, 0.9, 0.95, 1.0, 1.1, 1.2\}$ for 500 000 τ each. We used WHAM to calculate the specific heat of the system (Fig. 11). A clear peak occurs between lower temperatures, with formation of long-range fibrillar structures (Fig. 12), and higher temperatures where the system mostly samples random coil monomers.

At lower temperatures, where aggregation occurs, we mostly observe parallel sheets over antiparallel. Interestingly, this is in agreement with the study of Gsponer *et al.* and could be due to the hydrophobic interactions of the C-terminal tyrosine. To test this, we performed single point mutations in order to create a symmetric sequence. In this case parallel β -sheets also turned out to be more stable than antiparallel ones, which is unexpected since antiparallel β -sheets are generally believed to be lower in free energy.²⁵ One possible explanation is that the model is lacking electrostatic interactions at the N and C termini of the chains, which will favor antiparallel sheets, as the two ends have opposite charges.

These parallel GNNQQNY β -sheets also have the tendency to align within a plane, with the C termini facing each other. This evidently results from the attraction between the C-terminal tyrosines, the most hydrophobic amino acid in this peptide.

To show that their force field was not biased toward

FIG. 12. (Color) Snapshot of a typical cluster that forms by peptide aggregation in the low temperature phase ($T=0.8\epsilon/k_B$) of Fig. 11.

aggregation, the authors also simulated a water-soluble control peptide SQNGNQQRG and found a difference in the amount of β -sheets formed. We compared the phase behavior of GNNQQNY and SQNGNQQRG by using WHAM on both sequences but did not find statistically significant differences over the studied temperature range. This suggests that some of the details necessary to distinguish the thermodynamics of these two peptides are too subtle for our force field to represent. Since the simulation temperature of Gsponer *et al.* in our case maps to $T = 1\mathcal{E}/k_B$, which is where we essentially find the phase transition (Fig. 11), effects only captured by the atomistic force field can indeed be expected to lead to substantial differences. Previous studies have shown how differences in CG force field parameters affect structure, β -sheet propensity, and aggregation behavior of different sequences.⁶⁶

All of these aggregation results were obtained using the same force field with no additional parameter adjustment. Other CG models have previously demonstrated aggregation events on a larger scale.^{67,68} Here our goal was to show that we can study aggregation events using a force field that is tuned to reproduce simple folding features without biasing secondary or tertiary structure. This is important when looking at spontaneous aggregation or misfolding pathways, where one aims to reproduce general behavior without constraining the protein's structure toward a certain state that might not even be known or well defined.

VII. CONCLUSION

We have presented a new CG implicit solvent peptide model. Its intermediate resolution of four beads per amino acid permits accurate sampling of local conformations and thus secondary structure. Following cautious parameter tuning, the CG model is able to fold simple proteins such as helix bundles. Folding of a three-helix bundle was used to incorporate large-scale aspects of the force field, whereas the successful folding event of other helical bundles provided independent checks of reliability. Thermodynamic and kinetic studies of the three-helix bundle were carried out to make sure the folding temperature T_f was above the glass transition temperature T_g for this protein. The model was systematically compared to NMR data in order to optimize parameter tuning and precisely determine how much fine-scale information this CG model still contains. Of course, our model is not intended to compete with atomistic simulations, which is not the point of CG models; yet, carefully balancing several key contributions to the force field is a prerequisite to perform meaningful studies involving secondary and tertiary structure formation. Globular shaped proteins have proven more difficult to stabilize, presumably because accurate packing and strong cooperativity are not well enough captured. We also observe aggregation events of small β -sheets without retuning the force field. A realistic α/β balance, coupled with basic folding features, make the CG model very suitable for the large-scale and long-term regime that many biological processes require. Indeed, a force field that is not biased toward the protein's native conformation will likely give rise to insightful thermodynamic

and kinetic studies when the structure is not known, not well defined, strongly perturbed from the native state, or adjusts during aggregation events.

ACKNOWLEDGMENTS

We thank Zunjing Wang, Christine Peter, Cem Yolcu, and Bill DeGrado for valuable discussions and useful comments.

- ¹C. N. Pace, *Trends Biochem. Sci.* **15**, 14 (1990).
- ²C. N. Pace, B. A. Shirley, M. McNutt, and K. Gajiwala, *FASEB J.* **10**, 75 (1996).
- ³*Coarse-Graining of Condensed Phase and Biomolecular Systems*, edited by G. A. Voth (Taylor & Francis, New York, 2008).
- ⁴G. S. Ayton, W. G. Noid, and G. A. Voth, *Curr. Opin. Struct. Biol.* **17**, 192 (2007).
- ⁵V. Tozzini, *Curr. Opin. Struct. Biol.* **15**, 144 (2005).
- ⁶J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes, *Proteins: Struct., Funct., Genet.* **21**, 167 (1995).
- ⁷A. Arkhipov, P. L. Freddolino, and K. Schulten, *Structure (London)* **14**, 1767 (2006).
- ⁸A. Arkhipov, Y. Yin, and K. Schulten, *Biophys. J.* **95**, 2806 (2008).
- ⁹F. Ding, J. M. Borreguero, S. V. Buldyrey, H. E. Stanley, and N. V. Dokholyan, *Proteins: Struct., Funct., Genet.* **53**, 220 (2003).
- ¹⁰J. M. Sorenson and T. Head-Gordon, *J. Comput. Biol.* **7**, 469 (2000).
- ¹¹A. Voegler Smith and C. K. Hall, *Proteins: Struct., Funct., Genet.* **44**, 344 (2001).
- ¹²Y. Fujitsuka, S. Takada, Z. A. Luthey-Schulten, and P. G. Wolynes, *Proteins: Struct., Funct., Genet.* **54**, 88 (2004).
- ¹³T. Head-Gordon and S. Brown, *Curr. Opin. Struct. Biol.* **13**, 160 (2003).
- ¹⁴K. F. Lau and K. A. Dill, *Macromolecules* **22**, 3986 (1989).
- ¹⁵N. Go, *Annu. Rev. Biophys. Bioeng.* **12**, 183 (1983).
- ¹⁶L. Monticelli, S. K. Kandasamy, X. Periole, R. G. Larson, D. P. Tieleman, and S.-J. Marrink, *J. Chem. Theory Comput.* **4**, 819 (2008).
- ¹⁷L. Thøgersen, B. Schiøtt, T. Vosegaard, N. C. Nielsen, and E. Tajkhorshid, *Biophys. J.* **95**, 4337 (2008).
- ¹⁸A. Irback, F. Sjunnesson, and S. Wallin, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 13614 (2000).
- ¹⁹S. Takada, Z. Luthey-Schulten, and P. G. Wolynes, *J. Chem. Phys.* **110**, 11616 (1999).
- ²⁰G. Favrin, A. Irback, and S. Wallin, *Proteins: Struct., Funct., Genet.* **47**, 99 (2002).
- ²¹A. S. Yang and B. Honig, *J. Mol. Biol.* **252**, 366 (1995).
- ²²N. Y. Chen, Z. Y. Su, and C. Y. Mou, *Phys. Rev. Lett.* **96**, 078103 (2006).
- ²³H. Lodish, A. Berk, S. L. Zipursky, P. Matsudaira, D. Baltimore, and J. Darnell, *Molecular Cell Biology* (Freeman, New York, 2000).
- ²⁴P. T. Lansbury and H. A. Lashuel, *Nature (London)* **443**, 774 (2006).
- ²⁵A. V. Finkelstein and O. B. Pittsyn, *Protein Physics* (Academic, New York, 2002).
- ²⁶H. Meyer, O. Biermann, R. Faller, D. Reith, and F. Müller-Plathe, *J. Chem. Phys.* **113**, 6264 (2000).
- ²⁷J. L. Fauchere and V. Pliska, *Eur. J. Med. Chem.* **18**, 369 (1983).
- ²⁸S. Miyazawa and R. L. Jernigan, *J. Mol. Biol.* **256**, 623 (1996).
- ²⁹J. Skolnick, L. Jaroszewski, A. Kolinski, and A. Godzik, *Protein Sci.* **6**, 676 (1997).
- ³⁰S. Miyazawa and R. L. Jernigan, *Proteins: Struct., Funct., Genet.* **36**, 357 (1999).
- ³¹M. R. Betancourt and D. Thirumalai, *Protein Sci.* **8**, 361 (1999).
- ³²Z. H. Wang and H. C. Lee, *Phys. Rev. Lett.* **84**, 574 (2000).
- ³³M. S. Cheung, A. E. Garcia, and J. N. Onuchic, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 685 (2002).
- ³⁴Y. Mu and Y. Q. Gao, *J. Chem. Phys.* **127**, 105102 (2007).
- ³⁵E. H. Yap, N. L. Fawzi, and T. Head-Gordon, *Proteins: Struct., Funct., Bioinf.* **70**, 626 (2008).
- ³⁶C. L. Guo, M. S. Cheung, H. Levine, and D. A. Kessler, *J. Chem. Phys.* **116**, 4353 (2002).
- ³⁷L. A. Mirny and E. I. Shakhnovich, *J. Mol. Biol.* **264**, 1164 (1996).
- ³⁸S. Matysiak and C. Clementi, *J. Mol. Biol.* **363**, 297 (2006).
- ³⁹P. Das, M. Moll, H. Stamati, L. E. Kavraki, and C. Clementi, *Proc. Natl. Acad. Sci. U.S.A.* **103**, 9885 (2006).
- ⁴⁰C. Micheletti, A. Laio, and M. Parrinello, *Phys. Rev. Lett.* **92**, 170601 (2004).

- ⁴¹G. M. Torrie and J. P. Valleau, *J. Comput. Phys.* **23**, 187 (1977).
- ⁴²S. Park and K. Schulten, *J. Chem. Phys.* **120**, 5946 (2004).
- ⁴³R. H. Swendsen and J. S. Wang, *Phys. Rev. Lett.* **57**, 2607 (1986).
- ⁴⁴A. M. Ferrenberg and R. H. Swendsen, *Phys. Rev. Lett.* **61**, 2635 (1988).
- ⁴⁵S. Kumar, D. Bouzida, R. H. Swendsen, P. A. Kollman, and J. M. Rosenberg, *J. Comput. Chem.* **13**, 1011 (1992).
- ⁴⁶S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman, *J. Comput. Chem.* **16**, 1339 (1995).
- ⁴⁷T. Bereau and R. H. Swendsen, "Optimized convergence for multiple histogram analysis," *J. Comput. Phys.* DOI:10.106/j.jcp.2009.05.011 (to be published).
- ⁴⁸G. N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan, *J. Mol. Biol.* **7**, 95 (1963).
- ⁴⁹B. K. Ho, A. Thomas, and R. Brasseur, *Protein Sci.* **12**, 2508 (2003).
- ⁵⁰S. T. R. Walsh, H. Cheng, J. W. Bryson, H. Roder, and W. F. DeGrado, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 5486 (1999).
- ⁵¹W. Humphrey, A. Dalke, and K. Schulten, *J. Mol. Graphics* **14**, 33 (1996).
- ⁵²D. Frishman and P. Argos, *Proteins: Struct., Funct., Genet.* **23**, 566 (1995).
- ⁵³N. D. Socci and J. N. Onuchic, *J. Chem. Phys.* **101**, 1519 (1994).
- ⁵⁴F. W. DeGrado (personal communication).
- ⁵⁵H. J. Limbach, A. Arnold, B. A. Mann, and C. Holm, *Comput. Phys. Commun.* **174**, 704 (2006).
- ⁵⁶Q. H. Dai, C. Tommos, E. J. Fuentes, M. R. A. Blomberg, P. L. Dutton, and A. J. Wand, *J. Am. Chem. Soc.* **124**, 10952 (2002).
- ⁵⁷Y. N. Wei, S. Kim, D. Fela, J. Baum, and M. H. Hecht, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 13270 (2003).
- ⁵⁸A. Go, S. Kim, J. Baum, and M. H. Hecht, *Protein Sci.* **17**, 821 (2008).
- ⁵⁹A. Mondragon, S. Subbiah, S. C. Almo, M. Drott, and S. C. Harrison, *J. Mol. Biol.* **205**, 189 (1989).
- ⁶⁰S. Cho and D. W. Hoffman, *Biochemistry* **41**, 5730 (2002).
- ⁶¹M. T. Pastor, M. L. de la Paz, E. Lacroix, L. Serrano, and E. Perez-Paya, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 614 (2002).
- ⁶²S. H. Gellman, *Curr. Opin. Chem. Biol.* **2**, 717 (1998).
- ⁶³P. Ferrara, J. Apostolakis, and A. Caflisch, *J. Phys. Chem. B* **104**, 5000 (2000).
- ⁶⁴I. F. Thorpe, J. Zhou, and G. A. Voth, *J. Phys. Chem. B* **112**, 13079 (2008).
- ⁶⁵J. Gsponer, U. Haberthur, and A. Caflisch, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 5154 (2003).
- ⁶⁶G. Bellesia and J.-E. Shea, *J. Chem. Phys.* **126**, 245104 (2007).
- ⁶⁷S. Peng, F. Ding, B. Urbanc, S. V. Buldyrev, L. Cruz, H. E. Stanley, and N. V. Dokholyan, *Phys. Rev. E* **69**, 041908 (2004).
- ⁶⁸N. L. Fawzi, Y. Okabe, E. H. Yap, and T. Head-Gordon, *J. Mol. Biol.* **365**, 535 (2007).