A refined hydrogen bond potential for flexible protein models

Marta Enciso[a] and Antonio Rey[b]
Departamento de Química Física I, Facultad de Ciencias Químicas, Universidad Complutense, E-28040 Madrid, Spain

(Received 20 April 2010; accepted 6 May 2010; published online 15 June 2010)

One of the major disadvantages of coarse-grained hydrogen bond potentials, for their use in protein folding simulations, is the appearance of abnormal structures when these potentials are used in flexible chain models, and no other geometrical restrictions or energetic contributions are defined into the system. We have efficiently overcome this problem, for chains of adequate size in a relevant temperature range, with a refined coarse-grained hydrogen bond potential. With it, we have been able to obtain nativelike α-helices and β-sheets in peptidic systems, and successfully reproduced the competition between the populations of these secondary structure elements by the effect of temperature and concentration changes. In this manuscript we detail the design of the interaction potential and thoroughly examine its applicability in energetic and structural terms, considering factors such as chain length, concentration, and temperature. © *2010 American Institute of Physics*. [doi:10.1063/1.3436723]

I. INTRODUCTION

The backbone hydrogen bond is one of the most common and relevant interactions in proteins. It plays an essential role in their structure and folding, as it is the main driving force in the formation of α-helices and β-sheets.[1–3] Apart from its ubiquity in native proteins, it is also important for aggregation: aggregates share a β-type structure,[4] regardless of the native state of the original protein,[5] that can be linked to the formation of non-native intermolecular hydrogen bonds among protein backbones.

Understanding the internal basis of these interactions has become, therefore, an active research field in the past few years.[6,7] In this context, the computational approach must be specially highlighted,[8–11] as it can analyze this interaction individually, without considering the rest of energetic contributions (hydrophobicity, electrostatics, etc.).

One interesting spotlight within this field is the obtention of a complete folding landscape of peptidic systems in terms of the formation of backbone hydrogen bonds. As for any other kind of interaction, this task generally requires a huge computational effort, as proteins are complex systems with multiple degrees of freedom. One profusely accepted strategy to overcome this drawback is the use of coarse-grained models.[12–15] Lowering the level of detail of the protein description allows a faster exploration of the conformational space, losing the specific information of each atom, yet trying to provide a realistic description of the behavior of the protein as a whole.

The use of coarse-grained models implies a careful design of the energy functions, as important simplifications are often made, sometimes blurring the biochemical information of the polypeptidic chain. In the case of backbone hydrogen bonds, a good potential must fulfill three requirements. First of all, it should be able to generate the secondary structure elements by itself. Additionally, both helices and sheets should be equally favored so that changes in the system conditions (e.g., concentration) may lead to the different stable regular structures experimentally observed. Finally, these structures should be realistic: They should follow the structural and geometrical patterns of the α-helices and β-sheets found in native proteins.

The design of hydrogen bond potentials has raised a considerable attention among coarse-grained modelers. Consequently, a wide range of intermediate resolution models has been proposed recently.[10,16–19] Combined with other potentials for the rest of energetic interactions or applied to rigid fragments,[20] they give good results in most cases, leading to recognizable structures that are sensitive to the system conditions.

However, our aim is not to extract the role of backbone hydrogen bonds from a potential where other interactions are involved, but to analyze hydrogen bonding in flexible systems where it is the only energetic contribution taken into account. Under this situation, the obtention of the preferred structures for a given system fully relies on the careful design of the backbone hydrogen bond potential, as no other interaction is present and the system description lacks a complete biochemical information. The directionality of the hydrogen bonds and the absence in many coarse-grained models of an explicit consideration of the amino and carbonyl groups—which form the hydrogen bond between the backbones of two real amino acids—constitute the main hindrances.

Therefore, some of the models previously published create, in some conditions of interest, distorted structures that are not fully compatible with the natural topology of proteins. For instance, some authors report the presence of alternative helical structures that coexist with nativelike ones.[21] In our laboratory, we have also acknowledged this fact for other backbone hydrogen bond models, as the one

[a]Electronic mail: martaenciso@quim.ucm.es.
[b]Electronic mail: jsbach@quim.ucm.es.

introduced in Ref. 17. Using only the hydrogen bond contribution of this potential in our simulations, helices and sheets present local geometries that are overly compressed or expanded, respectively, in relation to those found in native elements. In addition, nativelike helices are only dominant at very low temperatures, whereas alternative helical structures—also incompatible to native helices—are the stable species near the folding/unfolding transition. Although other contributions in the full potential of the model in Ref. 17 may overcome these problems, the hydrogen bond model alone, as we try to use it here, is not completely satisfactory.

For these reasons, we present in this work a modified hydrogen bond potential based on an α-carbon (C_α) representation of the polypeptidic chain. Inspired by knowledge-based potentials,[22] we have analyzed the geometrical patterns that underlie the native secondary structure elements. To do this, we have statistically studied the geometry of backbone hydrogen bonds in native proteins deposited in the Protein Data Bank (PDB) database,[23] and extracted several geometrical requirements for angles and distances, in the same spirit as other authors presented in recent literature.[17,18] It is important to state, then, that the definition of our potential does not rely in a complex mathematical expression depending on several parameters, as it is usually the case. Instead, our potential defines a hydrogen bond interaction between two residues in the model if several geometrical restrictions are fulfilled. Thus, the definition of our interaction potential consists in the choice of an adequate set of restrictions and their limits.

To explore the scope of the resulting simple model, and the possible inherent limitations it may still present, we have performed a broad set of simulation tests to validate our potential. We have studied the influence of the chain length, as well as the effect of concentration and temperature changes in peptidic systems. We have analyzed the helix-coil transition and the sheet-helix-coil one, obtaining a range of conditions where our model shows to be fully valid. This way, we finally present a complete energy landscape for these systems.

II. MATERIALS AND METHODS

A. The backbone hydrogen bond potential

In this work we use a coarse-grained off-lattice representation of the polypeptidic chain. Each amino acid i is represented by a single center of interaction, a rigid sphere centered at the C_α position, denoted by a position vector $\mathbf{r}_i$ (see Fig. 1). Neighbor beads along the sequence are linked by a virtual bond vector ($\mathbf{v}_i$); the norm of this vector is fixed at 3.8 Å, corresponding to the length of a *trans* peptide bond. Our model reflects the chain flexibility within the limits imposed by the chemistry of the real bonds: The virtual bond angle associated with three consecutive C_α beads is allowed to range from 65° to 150°.

We have also built an auxiliary unit vector for each bead i ($\mathbf{h}_i$) that is perpendicular to the plane defined by the preceding ($\mathbf{v}_i$) and following ($\mathbf{v}_{i+1}$) virtual bond vectors of each bead. This vector $\mathbf{h}_i$, colored in red in Fig. 1, approximately indicates the direction of the hydrogen bond in real proteins.

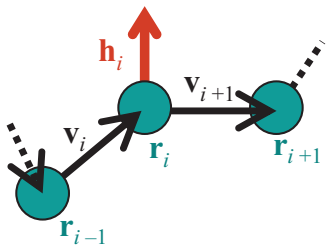

FIG. 1. Model description of a polypeptidic chain. The interaction centers (in blue) are placed at the C_α positions, given by their position vectors $\mathbf{r}$. Black arrows represent the virtual bonds, $\mathbf{v}$, and the red one is the auxiliary vector, $\mathbf{h}$.

Due to the reduced representation of our model, in our simulations, backbone hydrogen bonds are formally established among α-carbons, and therefore along the direction of the vector $\mathbf{r}_{ij}$ between the considered units. This approach requires a "renumbering" of the interacting residues, as it is done in Ref. 17, as the real backbone hydrogen bond between the ith and the jth residues is replaced by the interaction between the ith and the $(j-1)$th beads in our model.

The total energy of a given conformation is calculated using a pair-additive potential. In it, the $(i,i+4)$ interaction between beads within the same chain is skipped to favor a better geometry of helices and tight loops.[17] Interactions among residues near in sequence ($|j-i|<3$) are not considered if both beads belong to the same chain.

The core of the interaction definition for our hydrogen bond model lies in the fact that each individual energy contribution, u_{ij}, must be carefully designed, as the use of a simple description for the protein chain implies that the real geometry of the hydrogen bond interaction must be reflected elsewhere. This geometrical information has been extracted from a large amount of hydrogen bonded pairs that belong to real secondary structure elements (obtained from a representative set of over 1600 proteins via the DSSP files[24] available in the PDB database[23]). Although this is a common procedure in knowledge-based potentials, it presents some distinct features in the case of hydrogen bonds—compared to other interactions such as hydrophobic or electrostatic potentials—due to the covalent nature of hydrogen bonds. This covalency implies that the backbone hydrogen bond potential must reflect both a spatial and an orientational character.

Some previously published potentials use a single expression where both aspects are closely interrelated,[16,19] while others split them into different geometrical restrictions.[17,18] We have used this latter approach, identifying three representative quantities (named R1, R2, and R3) that can be computed using only the α-carbons of the i and j amino acids.

- R1 is a spatial restriction that designates the distance between the two α-carbons of the hydrogen bonded residues,

$$\mathrm{R1} = |\mathbf{r}_{ij}| = |\mathbf{r}_j - \mathbf{r}_i|. \tag{1}$$

This is a standard restriction for any pair of interacting residues, no matter what the type of interaction is.

- R2 is an orientational restrain which computes the cosine of the angle associated to the relative orientation between the auxiliary vectors of both residues,

$$R2 = |\cos(\mathbf{h}_i, \mathbf{h}_j)|. \quad (2)$$

In principle, these two vectors would be close to parallel (or antiparallel) for residues forming a hydrogen bond.

- R3 is also an orientational quantity that computes the cosine of the angle between the direction of the tentative hydrogen bond in the model and each of the auxiliary vectors; thus, R3 is independently calculated for both i and j beads (R3i and R3j),

$$R3 \begin{cases} R3i = |\cos(\mathbf{h}_i, \mathbf{r}_{ij})| \\ R3j = |\cos(\mathbf{h}_j, \mathbf{r}_{ij})|. \end{cases} \quad (3)$$

Again, the vectors considered should be almost parallel (or antiparallel) for the restrictions in R3. The simultaneous consideration of the orientational restrictions R2 and R3 helps to frame the hydrogen bonds found in real proteins in a better way than when only one of them is considered.

We have calculated these restrictions for each pair of hydrogen bonded amino acids within our 1600 protein database. We have separated this analysis in two different sets, depending on the formation of what we call a local interaction, present in α-helices and formed between the ith and the $(i+3)$th residues, according to our model, or a nonlocal one, that mainly leads to the formation of β-sheets and is defined between residues separated by more than three residues. The corresponding histograms for every one of our geometrical restrictions (separated for local and nonlocal pairs) are represented in Fig. 2. As we can see, most native hydrogen bonds clearly accumulate in relatively narrow ranges for the three geometrical quantities described above: R1 gets the most probable distances among residues below 6 Å, and the most probable values for R2 and R3 are close to unity, as expected for a nearly parallel (or antiparallel) orientation of the vectors involved in each restriction.

From the shape of the histograms, we have selected an optimal range of values for each restriction and type of interaction. The choice of these intervals is a critical aspect of our potential. On one hand, a very large proportion of native hydrogen bonds should be identified by our model (with our chosen range, it is nearly 80%). On the other hand, the selected intervals should be narrow enough to discriminate between nativelike and abnormal backbone hydrogen bonds, since this is precisely the type of result we want to enforce in our interaction model.

This selection has been a difficult task and we have had to perform many proofs using different possibilities, finally resulting in the optimum ranges summarized in Table I, which are also stripped and marked with arrows in Fig. 2. The choice of these intervals has been thoroughly optimized to exhibit the best properties under our numerical experiments. The interval limits, however, may vary within approximately 5% of the data in Table I with minor consequences in the model performance. The less flexible interval is the one corresponding to restriction R3 due to its wide distribution in native hydrogen bonds [especially for nonlocal pairs, see Fig. 2(f)]. Narrowing this interval considerably reduces the proportion of identified hydrogen bonds, while its extension favors the presence along the simulations of abnormal structures, such as those reported with other models.[21]

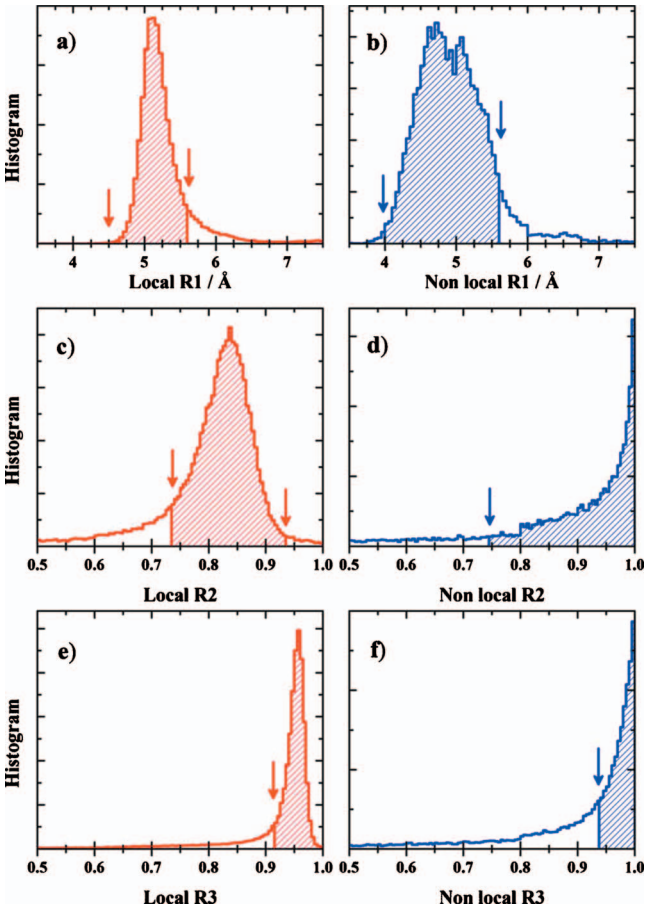

FIG. 2. Statistics over the PDB of native hydrogen bonds for the geometrical restrictions R1, R2, and R3. Histograms obtained from local pairs are colored in red (left column graphs) and nonlocal pairs are colored in blue (right column graphs). The stripped regions and arrows indicate the selected range of values in our model.

Once the choice of the proper intervals for the different restrictions is done, the definition of the hydrogen bond potential is really simple. Our interaction is built in two steps. First, the three geometrical restrictions are checked for each tentative pair of beads at a given conformation of the model. Then, we calculate its energy using the following expression:

TABLE I. Optimal ranges for the three geometrical restrictions chosen in our model for backbone hydrogen bonds.

Restriction	Local range	Nonlocal range
R1	4.7 Å ≤ R1 ≤ 5.6 Å	4.0 Å ≤ R1 ≤ 5.6 Å
R2	0.74 ≤ R2 ≤ 0.93	0.75 ≤ R2 ≤ 1.00
R3	0.92 ≤ R3 ≤ 1.00	0.94 ≤ R3 ≤ 1.00

$$\begin{cases} u_{ij} = u_{hb} & \text{if R1, R2, R3}i\text{, and R3}j \text{ are fulfilled} \\ u_{ij} = 0.25u_{hb} & \text{if R1, R2, and only one of R3}i \text{ or R3}j \text{ are fulfilled} \\ u_{ij} = 0 & \text{otherwise.} \end{cases} \quad (4)$$

This "steplike" potential is suitable for the hydrogen bond interaction, as it reflects the all-or-none nature of these bonds, related to its covalent character. Furthermore, it reduces the number of parameters of the model and accelerates the energy calculus.

The full energy of a single hydrogen bond, u_{hb}, is different depending on whether it is a local or a nonlocal bond, as experiments have proven that local bonds provide more stability than nonlocal ones.[25] After several proofs, we have found that the energy of a nonlocal bond should be 90%–95% of a local one. Therefore, we have chosen the following optimal value, in arbitrary units:

$$\begin{cases} u_{hb} = -10.0 & \text{for local bonds} \\ u_{hb} = -9.3 & \text{for nonlocal bonds.} \end{cases} \quad (5)$$

Finally, as an important technical detail, our potential considers an additional aspect of backbone hydrogen bonds: the right-hand chirality of native α-helices. Thus, we have assigned a certain chirality to each local bond (given by the sign of the triple product of the involved virtual bond vectors). Then, local interactions are only stabilized if their chirality is compatible with that found in real α-helices.

In our model, the first and last residues of a chain need a special treatment, as the auxiliary vector **h** cannot be built in these cases (see Fig. 1). Some models simply ignore these hydrogen bonds[16,17] but we have found that when doing so, in a model which includes a hydrogen bond potential alone, the conformation of terminal residues becomes almost completely free, in an unnatural way. For this reason, we have preferred the approach of Hoang *et al.*[18] We have used the same philosophy as above, keeping R1 unchanged, and merging R2 and R3 into a new restriction, which we call R2x, for the hydrogen bonds involving terminal residues. It computes the cosine of the angle formed by the tentative hydrogen bond direction in the model and the virtual bond vector of the terminal residue with its only neighbor. For example, if both i and j are end residues,

$$\text{R2x} \begin{cases} \text{R2x}i = |\cos(\mathbf{v}_i, \mathbf{r}_{ij})| \\ \text{R2x}j = |\cos(\mathbf{v}_j, \mathbf{r}_{ij})|. \end{cases} \quad (6)$$

Histograms and optimal values for R2x are displayed in Fig. 3 and Table II, respectively. Note that the selected values for our restriction R2x are close to zero, indicating that the involved vectors should be nearly perpendicular in this case.

Terminal hydrogen bonds are worse defined in native structures due to the usually higher intrinsic mobility of this protein region. The impossibility of the construction of an auxiliary vector in these cases constitutes an additional inconvenient for our model. Therefore, a terminal hydrogen bond in our model is formed only if every restriction is fulfilled. In addition, we have energetically penalized them with respect to hydrogen bonds involving only residues at the chain interior,

$$\begin{cases} u_{ij,\text{term}} = 0.75u_{hb} & \text{if R1, R2x}i\text{, and R2x}j \text{ are fulfilled} \\ u_{ij,\text{term}} = 0 & \text{otherwise,} \end{cases} \quad (7)$$

with u_{hb} given in Eq. (5).

It should be finally stressed that real amino acids can only form a maximum of two hydrogen bonds (one as donor and one as acceptor), but our model does not explicitly limit the number of hydrogen bonds per residue, falling the observance of this rule into the accurate election of the geometrical restrictions and their allowed intervals. At this point, our model differs from others, as some of them include specifically this limitation and, together with that, an explicit cooperativity effect.[18] The simplification we are assuming considerably reduces the computational cost of our model, without any relevant effect on the good properties of our simulation data, as we shall show in Sec. III.

B. Simulation method

In order to study the characteristics of the whole energetic and structural landscape for our model, we have used a parallel tempering[26] Monte Carlo simulation algorithm, as previously described.[27] We have carried out single-chain and multichain numerical experiments, using periodic boundary conditions in a cubic simulation box when the concentration needs to be defined.

We have included a great amount of temperatures in each of our simulations (ranging from 40 to 56, depending on the complexity of the system), as our aim is to explore our systems at both intermediate temperatures (characterizing folding transitions and the competition among the populations of different structures) and low ones, where parallel

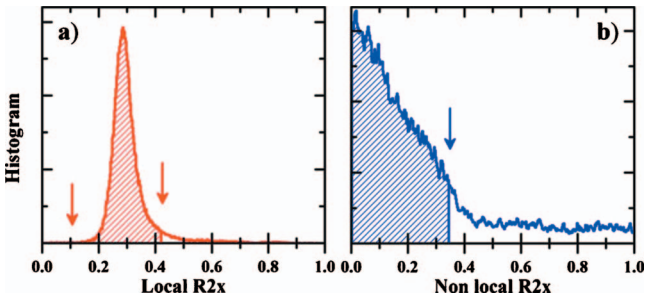

FIG. 3. Statistics over the PDB of native hydrogen bonds for the geometrical restriction R2x (terminal residues). Histograms obtained from local pairs are colored in red (left graph) and nonlocal pairs are colored in blue (right graph). The stripped regions and arrows indicate the selected range of values in our model.

TABLE II. Optimal range for the restriction R2x in hydrogen bonds involving terminal residues.

Restriction	Local range	Nonlocal range
R2x	$0.10 \leq R2x \leq 0.44$	$0.00 \leq R2x \leq 0.34$

tempering is used as a minimization technique. Each full simulation starts from a completely extended conformation for each chain and consists on 5×10^6 Monte Carlo cycles (at every temperature) after 2×10^6 equilibration cycles. In each cycle, every bead of the system is subjected to a trial Monte Carlo move.

The simplicity of the system description and the energy calculation reduces considerably the computational cost of our model. The times employed for the simulations in this work in single-processor machines range from approximately 3 h for single-chain systems of the smallest system studied (one chain of ten residues) to 55 h for simulating five chains of 12 residues and 56 temperatures. The algorithm is easily parallelized, which permits rather fast calculations in multiprocessor computers. The results presented here correspond to statistical averages over the sampling at every temperature and over different independent runs. For each system, four or five independent runs have been carried out.

III. RESULTS AND DISCUSSION

Apart from being carefully designed, a suitable potential must exhibit its good properties in simulated systems. For this reason, we have analyzed the behavior of a variety of flexible peptidic systems under our potential, considering factors such as the chain length and the system concentration. For each case, we have registered representative configurations from the equilibrium ensemble over a wide range of temperatures, from very low ones, where the system is nearly frozen, to intermediate and high ones, where we have explored the model behavior in the folding/unfolding transition. Therefore, we have tried to cover the complete energy landscape for our simulated systems.

A. Helix-coil transition

One of the most common tests for a backbone hydrogen bond potential is the study of the helix-coil transition,[28–32] as it is known that some homopeptidic chains naturally fold into a helix in diluted systems.[33,34] For us, this numerical experiment pursues two aims: to check whether our potential favors native helical states in diluted systems and to determine if other alternative structures are present in our simulations, allowed by the geometrical requirements of the defined interaction but lacking any physical meaning for a polypeptide chain.

We have simulated single-chain systems, i.e., infinite-dilution conditions, varying the peptide length (L) from 10 to 25 residues. In Fig. 4, we show the heat capacity curves of each system (computed from the energy fluctuations) in terms of temperature in reduced units, reflecting the energetic and structural transitions that take place. As we can see, short chains show a common behavior with only one peak in this curve, which corresponds to the transition from a folded state to a denatured (unfolded) one. On the contrary, the longest chain exhibits an unexpected behavior, with a double peak in the heat capacity curve (see Fig. 4 for $L=25$).

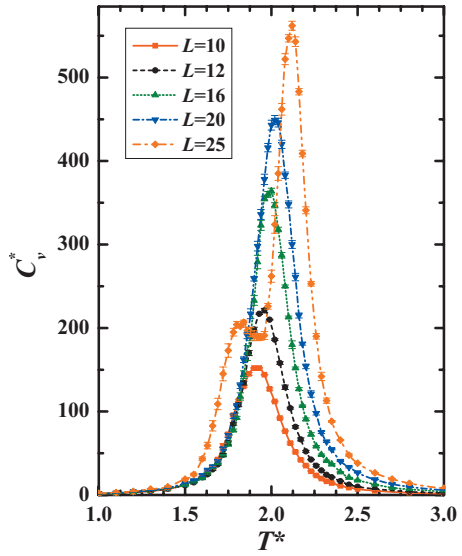

FIG. 4. Heat capacity curves vs temperature for infinite-dilution systems of different chain lengths (L). Note the use of reduced units in both axes.

We also show structural information for these systems, obtained from the registered configurations of our simulations. These structures correspond to α-helices, unfolded chains, and distorted wide helices with nonlocal bonds ($|j-i|>4$). Depending on the total number of hydrogen bonds and their kind (either local or nonlocal), we have assigned a characteristic type of structure to each configuration. This allows us to perform a population analysis, whose results are shown in Fig. 5. In this figure, a schematic plot of each kind of structure has also been drawn using visual molecular dynamics (VMD).[35] Again, the modification of the transition characteristics with the chain length is observed. At low temperatures, every chain folds into an α-helix, stabilized by the model interactions (i, $i+3$). The shortest chains ($L=10$ or 12) unfold at the transition temperature without any intermediates, as seen in Figs. 5(a) and 5(b). At intermediate temperatures, the increase in the chain length [Figs. 5(c)–5(e)] leads to the growth of the population of an alternative type of helix. This distorted structure, built thanks to nonlocal hydrogen bonds, corresponding in the model to (i, $i+5$) interactions, becomes dominant at intermediate temperatures for the 25-residue chain, explaining therefore the double peak in the heat capacity curve of Fig. 4.

According to our results, the population of distorted helices is null or almost negligible for chains shorter than 20 residues (less than 5% of the registered configurations). As the average length of a native α-helix in globular proteins is about 12 residues,[36] we can conclude that our model succeeds in the obtention of nativelike helices in infinite-dilution conditions for realistic chain lengths.

We have also performed several proofs including the (i, $i+4$) interaction in the energy calculations (data not shown). They highlight the importance of this detail: The forbiddance of (i, $i+4$) interactions widens the thermal sta-

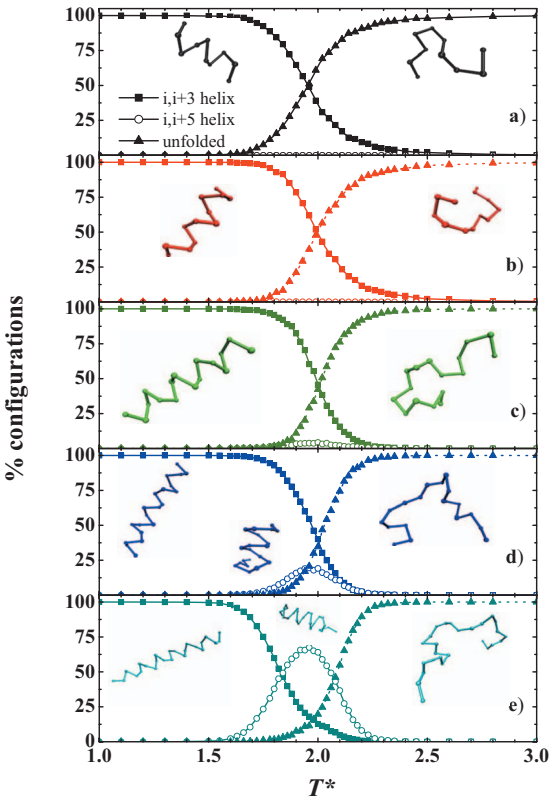

FIG. 5. Variation in the fraction of each type of structure, i.e., nativelike helices stabilized by $(i,\ i+3)$ hydrogen bonds (solid squares), distorted helices stabilized by nonlocal hydrogen bonds (open circles) or unfolded structures (solid triangles), with temperature for different chain lengths. Structures represented using VMD (Ref. 35). (a) $L=10$. (b) $L=12$. (c) $L=16$. (d) $L=20$. (e) $L=25$.

bility range for the $(i,\ i+3)$ ones, obtaining unique nativelike structures in the significant chain length range commented on above.

B. Sheet-helix-coil transition for 12-residue-long chains

As helical structures naturally appear in infinite-dilution conditions, increasing the system concentration leads to the formation of β-sheets.[37] Therefore, the change in concentration explores the competition between the population of α-helices and β-sheets, also known as the sheet-helix-coil transition.[38,39] This strategy is also related to aggregation, as it evaluates the propensity of interchain bonds (association of chains) against intrachain hydrogen bonds (single-chain folding).

To carry out this study, we have simulated several multichain systems consisting on five chains of twelve residues each. This chain length corresponds to the average length of a native α-helix and it is also long enough to participate in β-sheets.[36] In addition, its good helical behavior has been shown in Sec. III A. We have computed different concentrations for the system within the same order of magnitude. In this manuscript we shall show only the four most representative ones, ranging from 0.01 to 0.06 chain moles/L. Note that the numerical values of the simulated concentration are just representative of the variation analyzed in this work, but they do not try to reflect a real experimental concentration.

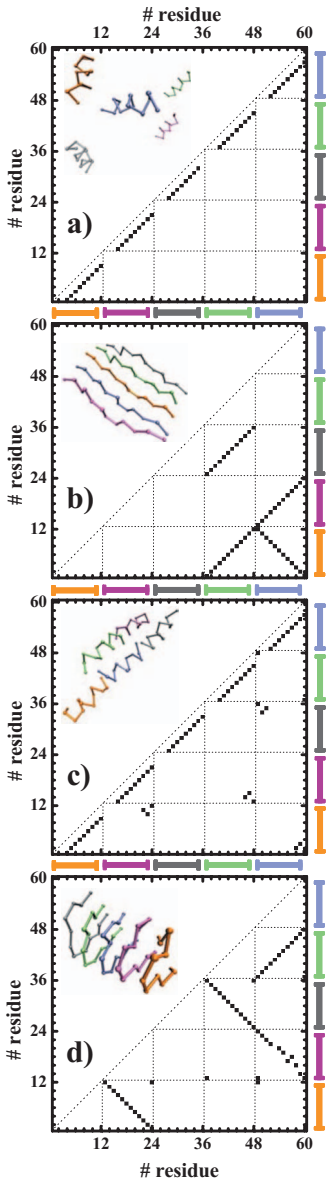

FIG. 6. Bidimensional energy maps for the different structures obtained in the multichain simulations. Structures represented using VMD (Ref. 35). The thin dotted lines indicate the end of a chain and the beginning of another one. (a) Structure A (free α-helices). (b) Structure B (nativelike β-sheet). (c) Structure C (oligomeric helices). (d) Structure D (overstabilized β- sheet).

In this section, we first describe the different types of structures we have observed in our simulations. Then, we evaluate their impact in the general landscape of multichain systems in terms of their stability range in concentration and temperature, ending with a schematic phase diagram for 12-residue peptides.

In our simulated trajectories with multichain systems, we have detected four types of regular structures, named in this manuscript as A, B, C, and D. A detailed representation of each of them is shown in Fig. 6. We present a cartoon image of the structures [drawn with VMD (Ref. 35)] and also a bidimensional energy map for each one, where a black spot indicates the presence of a hydrogen bond according to our model. Of course, we have also detected essentially random structures that have been classified as "unstructured."

Structure A is presented in Fig. 6(a). It corresponds to the single-chain folding of each peptide into a helix, stabilized only by local hydrogen bonds, i.e., $(i,\ i+3)$ interactions.

Structure B shows long range interchain hydrogen bond interactions, easily detected by the black spots on the block diagonals of the energy map in Fig. 6(b). It corresponds to a five-chain β-sheet, where both parallel and antiparallel arrangements appear. We have found different arrangements of the strands on the β-sheets in our simulations.

As well as these two nativelike structures, we have detected two other types of regular structures, undesirable in terms of the features found in native proteins. Structure C, similarly to structure A, is mainly helical. However, apart from the local intrachain hydrogen bonds, this structure also presents interchain associations between the terminal residues of a pair of chains. In this way, the simulated system finds more hydrogen bonds than those initially expected, creating a sort of oligomeric helical superstructure, where hydrogen bonds propagate as if the system were a large single chain, instead of multiple independent ones. This structure drastically minimizes the system energy through a strong reduction of the system mobility.

Structure D, displayed in Fig. 6(d), is based on interchain contacts. These nonlocal hydrogen bonds form a sort of distorted β-type structure, wrapped into itself by the formation of extra hydrogen bonds among terminal residues. This additional energetic stabilization illustrates, as in structure C, a violation of the "protein chemistry," at a high entropic cost.

As we have already stated, structures C and D do not match any relevant natural one, constituting artifacts of our model. They highlight an important feature of our potential: The number of hydrogen bonds for a given residue is not restricted in the potential definition. For inner residues, it does not imply any significant drawback, as the careful design of the restrictions and parameters of the model naturally limits the number of hydrogen bonds per bead. However the laxer definition of the interactions involving terminal residues fosters the formation of these abnormal structures. However, since we have penalized the energy of the hydrogen bonds involving terminal residues in our model, as shown in Eq. (7), the significant population of these structures can be restricted to very low temperatures, as we show in the next paragraphs.

To determine the significance of each structure over the whole energy landscape, we have followed the same strategy as in infinite-dilution systems. We have analyzed both the energetic and structural evolution with temperature for each of our simulated systems. First, we present in Fig. 7 the heat capacity curves versus temperature for each multichain system, from the most diluted one [Fig. 7(a)] to the most concentrated one [Fig. 7(d)]. In Fig. 8, we plot the evolution with the simulation temperature of the population of the four distinct structures introduced above. Each registered configuration can be easily classified into one of those four structures (or as unstructured), as the proportion of local/nonlocal and inner/terminal hydrogen bonds follows a clear pattern in each case.

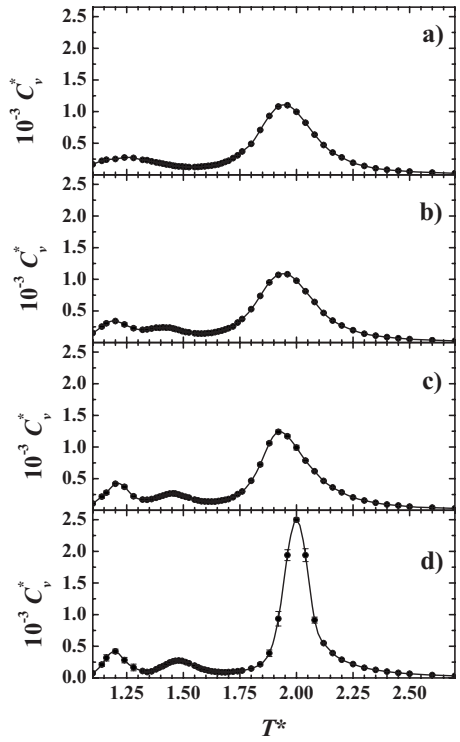

FIG. 7. Heat capacity curves vs temperature for each multichain simulated system. (a) 0.01 chain moles/L. (b) 0.02 chain moles/L. (c) 0.04 chain moles/L. (d) 0.06 chain moles/L.

Starting with the most diluted system (0.01 chain moles/L), it exhibits two well separated maxima in the heat capacity curve [see Fig. 7(a)]. These two energetic transitions define temperature ranges where a certain structure population is dominant. In Fig. 8(a) we observe that structure C, the helical oligomeric structure, is preponderant at very low temperatures. Its interchain interactions imply a large entropic cost, so a slight increment in the temperature of the system breaks these associations, resulting in a large stability region (from $T^*=1.30$ to its unfolding temperature, $T_m^*=1.95$) for structure A (isolated α-helices) as the unique stable structure within this range. In this sense, we have recovered the helix-coil transition studied in Sec. III A for infinite-dilution conditions—the black curve in Fig. 4. The presence of multiple chains in the current system is reflected by the existence of an extra transition at very low temperatures between the artifactual and natural helical structures, C and A, but without any unwanted effects in the not-frozen temperature range.

Following our discussion with a higher concentration system (0.02 chain moles/L), the corresponding heat capacity curve of Fig. 7(b) shows an additional peak in the low temperature region. At the lowest temperatures, structure C is again the most populated [see Fig. 8(b)] but a small temperature increase within this almost-frozen range leads to an energetic and structural transition to structure D, as interchain interactions are not so infrequent as before. It is also stable only at very low temperatures, as the entropic cost of blocking the terminal residues is still very high. For this reason, the stability region of these structures is small, essentially disappearing above $T^*=1.50$.

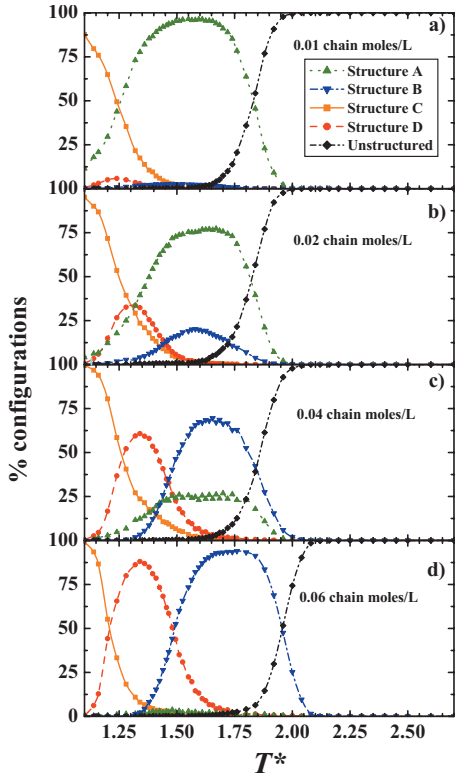

FIG. 8. Temperature evolution of the population of the different types of structures (A, B, C, or D, according to Fig. 6) observed in multichain systems. (a) 0.01 chain moles/L. (b) 0.02 chain moles/L. (c) 0.04 chain moles/L. (d) 0.06 chain moles/L.

In the subsequent temperature interval, structure A (isolated α-helices) is the predominant feature, although a smaller population of structure B (β-sheet) has also been detected. Thus, a concentration increment has revealed the emerging competition between the populations of the two nativelike secondary structure elements, mediated by this factor. It is also remarkable that they are the only stable structures within this relevant temperature range, so our results in the vicinity of the unfolding transition are not marred by the undesired presence of alternative structures.

The main results for the simulated system of 0.04 chain moles/L are shown in Figs. 7(c) and 8(c). This concentration displays a very similar behavior to the 0.02 chain moles/l system, with the same three energetic and structural transitions we have previously discussed. Importantly, the concentration increase has modified the relative population of helices and sheets (structures A and B) in the intermediate temperature region, predominating β-sheets in this case. This shows that concentration really modulates the competition between these two structures in our simulations.

Finally, Figs. 7(d) and 8(d) illustrate our results for the 0.06 chain moles/L system. In this case, the low temperature region exhibits the behavior previously described. However, the high temperature transition [$T_m^*=2.05$ in Fig. 7(d)] presents different characteristics, as it is higher and narrower than the transitions observed at lower concentrations. This can be linked to the absence of a significant population of structure A [see Fig. 8(d)], as interactions among chains (due to the high concentration of the system) are so common that finding isolated helices is very rare. In this concentration conditions, we have lost the competition between structures, obtaining the sheet-coil transition.

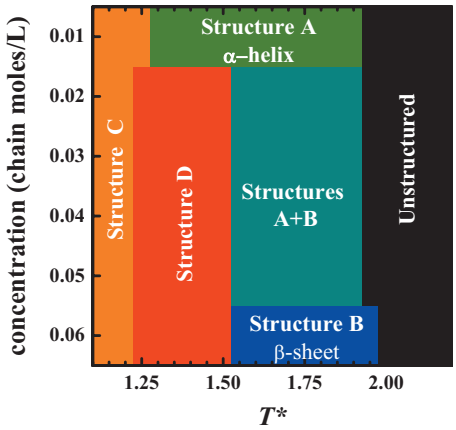

FIG. 9. Schematic phase for multichain systems according to our simulation model.

This fact also gives the clue for the similarity of this high temperature peak among the rest of the simulated systems. Sheets in intermediate concentration systems unfold via a helical intermediate, i.e., the chains that separate from the β-sheet form a helix before becoming completely unstructured at a higher temperature. This has also been acknowledged by direct observation of the simulation trajectories and might be probably linked to the experimental findings of a helical intermediate in aggregation processes.[40]

The computed simulations with our model have allowed us to explore the most relevant situations we would expect in systems driven by the formation of hydrogen bonds, from the structural helix-coil transition in diluted conditions to the sheet-coil one in concentrated systems. This information is condensed in Fig. 9, a sketched phase diagram where we have roughly matched the simulated structures to their stability regions as a function of concentration and temperature.

At very low temperatures, every system finds structure C, as it has the lowest energy due to its high helicity and additional hydrogen bonds. It is entropically disfavored, so a small temperature raise drastically destabilizes this structure. In very diluted systems, interchain interactions are scarce so the disappearance of structure C releases the chains. If the system concentration is moderate or high, our model produces a high population of structure D, thanks to the formation of interchain bonds. Its entropic cost, although lower than structure C's, is still high. For this reason, structure D also becomes unstable at relatively low temperatures. This way, the abnormal structures' interval ends at temperatures that, being too low, could be easily ignored in a numerical study of the simulation model around the folding/unfolding temperature.

Within the relevant temperature range, the structural situation depends on the system concentration. If concentration is low, isolated α-helices (structure A) are the stable feature until their complete denaturalization. In highly concentrated systems, β-sheets (structure B) are stable until un-

folding, as crowded systems promote interchain hydrogen bonds. For moderate concentrations, structures A and B coexist within the same temperature range.

Apart from the results shown here, we have also performed many additional simulations on multichain systems varying the chain length, the size of the simulation box, and the number of chains of the system for a comparable concentration range. We have found a similar behavior in all cases, confirming the general validity of the conclusions reported in this work.

IV. SUMMARY AND CONCLUSIONS

In this work, we have designed and evaluated a hydrogen bond potential based on a coarse-grained C_α representation of a protein, suitable for simulating flexible polypeptide chains. Our main goal is the obtention of secondary structure elements that fully resemble the native ones for relevant temperature and concentration conditions. Modulating these parameters, we have successfully reproduced the competition between α-helices, β-sheets, and unfolded chains.

Inspired by knowledge-based potentials, we have studied a representative set of native hydrogen bonded pairs, analyzing three geometrical features. They reflect the native tendencies of hydrogen bonds in relation to the distance between α-carbons of bonded residues (R1) and the relative orientation between auxiliary vectors in the model (R2) and between the auxiliary vector of each residue and the hydrogen bond direction in the model (R3). We have calculated these values for each pair of hydrogen bonded residues in a large set of protein structures, extracting recognizable tendencies for both local and nonlocal bonds. In the case of terminal residues, as auxiliary vectors cannot be built, the hydrogen bond definition is laxer. Selecting the best interval for each restriction and case, we have chosen a range of values where each condition has to be fulfilled for the hydrogen bond of our model to be considered.

To analyze the consequences of the model definition, we have used a parallel tempering Monte Carlo technique applied to a flexible chain model, simulating for each system a complete set of temperatures that ranges from a nearly frozen situation to the system unfolding and beyond. Remarkably, the potential calculation is very fast, obtaining full simulations of the complete energy landscape of the system in short CPU times (as much as tens of hours in single-processor machines). The main reasons for this efficiency are the election of a simple description of the system (only one interaction center per bead), the lack of complex expressions for the geometrical restrictions and energy calculations, and the absence of an explicit limitation in the number of hydrogen bonds per bead or cooperativity effects.

We have thoroughly examined the applicability of our model without any extra energetic contributions, presenting in this manuscript a complete set of tests to validate our hydrogen bond potential. First, we have simulated the whole folding transition of a polypeptidic system in infinite-dilution conditions. In this way, we have obtained the helix-coil transition for different chain lengths, reproducing nativelike energetic and structural properties for chains shorter than 20 residues, clearly above the average α-helix length in globular proteins (twelve residues).[36]

We have also studied the effect of concentration. For that purpose, in this work we report the results of a system composed by five polypeptidic chains of 12 residues each, whose concentration has been modified by changes in the size of the simulation box. In this case, the structural and energetic scenarios become more complex, as the number of interacting possibilities increases. Apart from detailed considerations for each individual system, a complete phase diagram has been obtained.

Four different types of structures have been observed. In the very low temperature region, where our system is nearly frozen, we observe two abnormal structures with more hydrogen bonds than those allowed by nature, being the result of a too intense energy minimization and the inherent (and always commented on) model limitations. The population of these structures, anyhow, is negligible at relevant temperatures, and therefore does not imply any serious pitfall of the model used in this work.

Thus, above this extremely low temperature region, we only observe nativelike structures and their unfolding processes. The system behavior depends on its concentration. Modifying this parameter within sensitive values, we have observed every expected relevant situation. Highly concentrated systems show the typical sheet-coil transition, whereas more diluted ones reflect a tough competition between the population of helices and sheets, the sheet-helix-coil transition. It is also remarkable that our model predicts a helical thermodynamic intermediate in the denaturalization of β-sheets when the sheet-helix-coil transition is observed, which may indicate the applicability of this model for numerical experiments in aggregation-prone conditions.

Altogether, we can say that with a very simple representation of the polypeptide chain, based only on the α-carbon positions of the residues, it seems to be essentially impossible to design a hydrogen bond potential which only produces the natural α-helices and β-sheets found in real proteins in any simulation condition. However, we have shown that our model is able indeed to clearly penalize other abnormal structures in the relevant temperature regime, and therefore they are only significantly populated at extremely low temperatures. This we have got with a definition of the interaction potential which is based on a careful choice of a proper set of geometrical restrictions, and their corresponding ranges.

In conclusion, our hydrogen bond model succeeds in the reproduction of the main hydrogen bond features in flexible chains without renouncing to a simple (and, thus, computationally fast) description. We have proven that this potential is completely suitable for the study of both secondary structure elements and aggregating conditions, showing excellent possibilities for present and future work.

ACKNOWLEDGMENTS

This work was partially supported by the Spanish Ministerio de Ciencia e Innovación (Grant Nos. FIS2009-

13364-C02-02 and CSD2007-00010), by the Comunidad Autónoma de Madrid (Grant No. S2009/PPQ-1551), and by the Universidad Complutense de Madrid/Banco Santander Central Hispano (Grant No. GR58/08-910068). M.E. acknowledges a Scholarship from Spanish Ministerio de Educación.

[1] K. A. Dill, Biochemistry **29**, 7133 (1990).
[2] D. F. Sticke, L. G. Presta, K. A. Dill, and G. D. Rose, J. Mol. Biol. **226**, 1143 (1992).
[3] D. Whitford, *Proteins: Structure and Function* (Wiley, New York, 2005).
[4] M. Sunde, L. C. Serpell, M. Bartlam, P. E. Fraser, M. B. Pepys, and C. C. F. Blake, J. Mol. Biol. **273**, 729 (1997).
[5] J. I. Guijarro, M. Sunde, J. A. Jones, I. D. Campbell, and C. M. Dobson, Proc. Natl. Acad. Sci. U.S.A. **95**, 4224 (1998).
[6] M. Wang, T. E. Wales, and M. C. Fitzgerald, Proc. Natl. Acad. Sci. U.S.A. **103**, 2600 (2006).
[7] R. L. Baldwin, J. Mol. Biol. **371**, 283 (2007).
[8] A. R. Fersht and V. Daggett, Cell **108**, 573 (2002).
[9] C. D. Snow, E. J. Sorin, Y. Min Rhee, and V. S. Pande, Annu. Rev. Biophys. Biomol. Struct. **34**, 43 (2005).
[10] A. V. Morozov and T. Kortemme, Adv. Protein Chem. **72**, 1 (2005).
[11] G. D. Rose, P. J. Fleming, J. R. Banavar, and A. Maritan, Proc. Natl. Acad. Sci. U.S.A. **103**, 16623 (2006).
[12] N. Gō, J. Stat. Phys. **30**, 413 (1983).
[13] F. Ding, S. V. Buldyrev, and N. V. Dokholyan, Biophys. J. **88**, 147 (2005).
[14] H. Imamura and J. Z. Y. Chen, Proteins **63**, 555 (2006).
[15] R. V. Pappu and R. Nussinov, Phys. Biol. **6**, 010301 (2009).
[16] J. Z. Y. Chen and H. Imamura, Physica A **321**, 181 (2003).
[17] A. Kolinski, Acta Biochim. Pol. **51**, 349 (2004).
[18] T. X. Hoang, A. Trovato, F. Seno, J. R. Banavar, and A. Maritan, Proc. Natl. Acad. Sci. U.S.A. **101**, 7960 (2004).
[19] D. K. Klimov, M. R. Betancourt, and D. Thirumalai, Folding Des. **3**, 481 (1998).
[20] D. De Sancho and A. Rey, J. Comput. Chem. **28**, 1187 (2007).
[21] J. R. Banavar and A. Maritan, Annu. Rev. Biophys. Biomol. Struct. **36**, 261 (2007).
[22] C. Zhang, S. Liu, H. Zhou, and Y. Zhou, Protein Sci. **13**, 400 (2004).
[23] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, Nucleic Acids Res. **28**, 235 (2000).
[24] W. Kabsch and C. Sander, Biopolymers **22**, 2577 (1983).
[25] Z. Shi, B. A. Krantz, N. Kallenchach, and T. R. Sosnick, Biochemistry **41**, 2120 (2002).
[26] U. H. E. Hansmann, Chem. Phys. Lett. **281**, 140 (1997).
[27] L. Prieto, D. de Sancho, and A. Rey, J. Chem. Phys. **123**, 154903 (2005).
[28] V. Varshney, T. E. Dirama, T. Z. Sen, and G. A. Carri, Macromolecules **37**, 8794 (2004).
[29] R. B. Best and G. Hummer, J. Phys. Chem. B **113**, 9004 (2009).
[30] U. H. E. Hansmann and Y. Okamoto, J. Chem. Phys. **110**, 1267 (1999).
[31] V. Daggett and M. Levitt, J. Mol. Biol. **223**, 1121 (1992).
[32] Y. Chen, Y. Zhou, and J. Ding, Proteins **69**, 58 (2007).
[33] T. Head-Gordon, F. H. Stillinger, M. H. Wright, and D. M. Gay, Proc. Natl. Acad. Sci. U.S.A. **89**, 11513 (1992).
[34] V. A. Bloomfield, Am. J. Phys. **67**, 1212 (1999).
[35] W. Humphrey, A. Dalke, and K. Schulten, J. Mol. Graphics **14**, 33 (1996).
[36] T. E. Creighton, *Proteins: Structures and Molecular Properties* (Freeman, New York, 1993).
[37] M. Morillas, D. L. Vanik, and W. K. Surewicz, Biochemistry **40**, 6982 (2001).
[38] B. Ilkowski, J. Skolnick, and A. Kolinski, Macromol. Theory Simul. **9**, 523 (2000).
[39] Y. Peng and U. H. E. Hansmann, Phys. Rev. E **68**, 041911 (2003).
[40] C. M. Dobson, Philos. Trans. R. Soc. London, Ser. B **356**, 133 (2001).