JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

Minimalist Model for the Dynamics of Helical Polypeptides: A Statistic-Based Parametrization

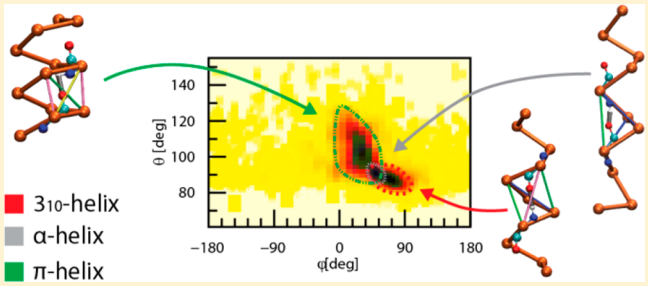
Giulia Lia Beatrice Spampinato,[†] Giuseppe Maccari,*[,‡] and Valentina Tozzini[†]

[†]NEST, Istituto Nanoscienze—CNR and Scuola Normale Superiore, Piazza San Silvestro 12-56127 Pisa, Italy

[‡]Center for Nanotechnology and Innovation @NEST, Istituto Italiano di Tecnologia, Piazza San Silvestro 12-56127 Pisa, Italy

Supporting Information

ABSTRACT: Low-resolution models are often used to address macroscopic time and size scales in molecular dynamics simulations of biomolecular systems. Coarse graining is often coupled to knowledge-based parametrization to obtain empirical potentials able to reproduce the system thermodynamic behavior. Here, a minimalist coarse grained (GC) model for the helical structures of proteins is reported. A knowledge-based parametrization strategy is coupled to the explicit inclusion of hydrogen-bonding-related terms, resulting in an accurate reproduction of the structure and dynamics of each single helical type, as well as the internal conformational variables correlation. The proposed strategy of basing the force field terms on real physicochemical interactions is transferable to different secondary structures. Thus, this work, though conclusive for helices, is to be considered the first of a series devoted to the application of the knowledge-based, physicochemical model to extended secondary structures and unstructured proteins.



INTRODUCTION

Atomistic molecular dynamics (MD) computer simulations based on empirical force fields (FF) are considered invaluable tools in the study of biological matter, capable of an insight hardly accessible to experiments.[1] Within this modeling framework, atoms interact by means of empirical potentials forming altogether the FF. In spite of the undoubted impact these approaches had in modern biochemistry and biophysics, they also present some weaknesses, basically related to the computational cost. Simulations of single proteins can currently reach the microsecond run length on single processors; however addressing larger macromolecular systems and/or for longer time-scales requires massively parallel computational resources.[2] Though computer power is constantly increasing with time, it was also observed that as the time-scale of simulations increases, new and subtler problems start to appear: traditional FFs show inaccuracies, especially in the evaluation of the relative energies of different secondary structures,[3,4] which specifically emerge when the thermodynamic limit is approached. A great effort is currently in the works to produce a new generation of FFs to fix these problems.[5,6] These new FFs, however, include a higher level of complexity and a number of additional parameters and will presumably require some time to be tested.

Apparently paradoxically, the reductionist approach can be considered as an alternative strategy to get around these problems. Simplifying the system by coarse-graining it,[7−10] i.e., representing a group of atoms by means of a single interacting center (bead), results in a reduction of computational cost. The direct consequence is that macroscopic time-space scales are immediately reachable. A second—less obvious—gain, resides in the reduced number of the model parameters. On one hand, this implies that accuracy, transferability, and predictive power might be reduced,[11] as well. On the other hand, taking advantage of the small number of parameters, more efficient optimization strategies can be adopted, and information from different sources (e.g., measured values of thermodynamics, or statistical observed) can be included, which is generally called "knowledge-based" parametrization. This allows creating classes of "minimalist models"[12] for specific tasks. At the same time, the coarse grained (CG) model must not disregard the atomistic level. Conversely, it must be coherent with it in order to have a complete and accurate representation of the system. This is, in fact, the core philosophy of the multiscale approach, for which Martin Karplus, Arieh Warshel, and Michael Levitt were recently awarded the Nobel Prize in Chemistry (2013).[13]

All of this considered, a large amount of work was recently devoted to the parametrization of "knowledge-based" potentials,[14] in particular to one specific aspect of knowledge-based derivation of parameters, namely their mining from statistical data sets of structures. In practice, this corresponds to finding the set of parameters that better reproduces in simulations the distributions of the internal variables derived from the data set,[15−17] returning the "statistically derived potentials."

In this work, we focus on one-bead CG models for proteins, representing each amino acid with a single bead placed on the α-carbon (Cα), although some of the results can be extended

Received: May 9, 2014
Published: July 22, 2014

   dx.doi.org/10.1021/ct5004059 | J. Chem. Theory Comput. 2014, 10, 3885−3895

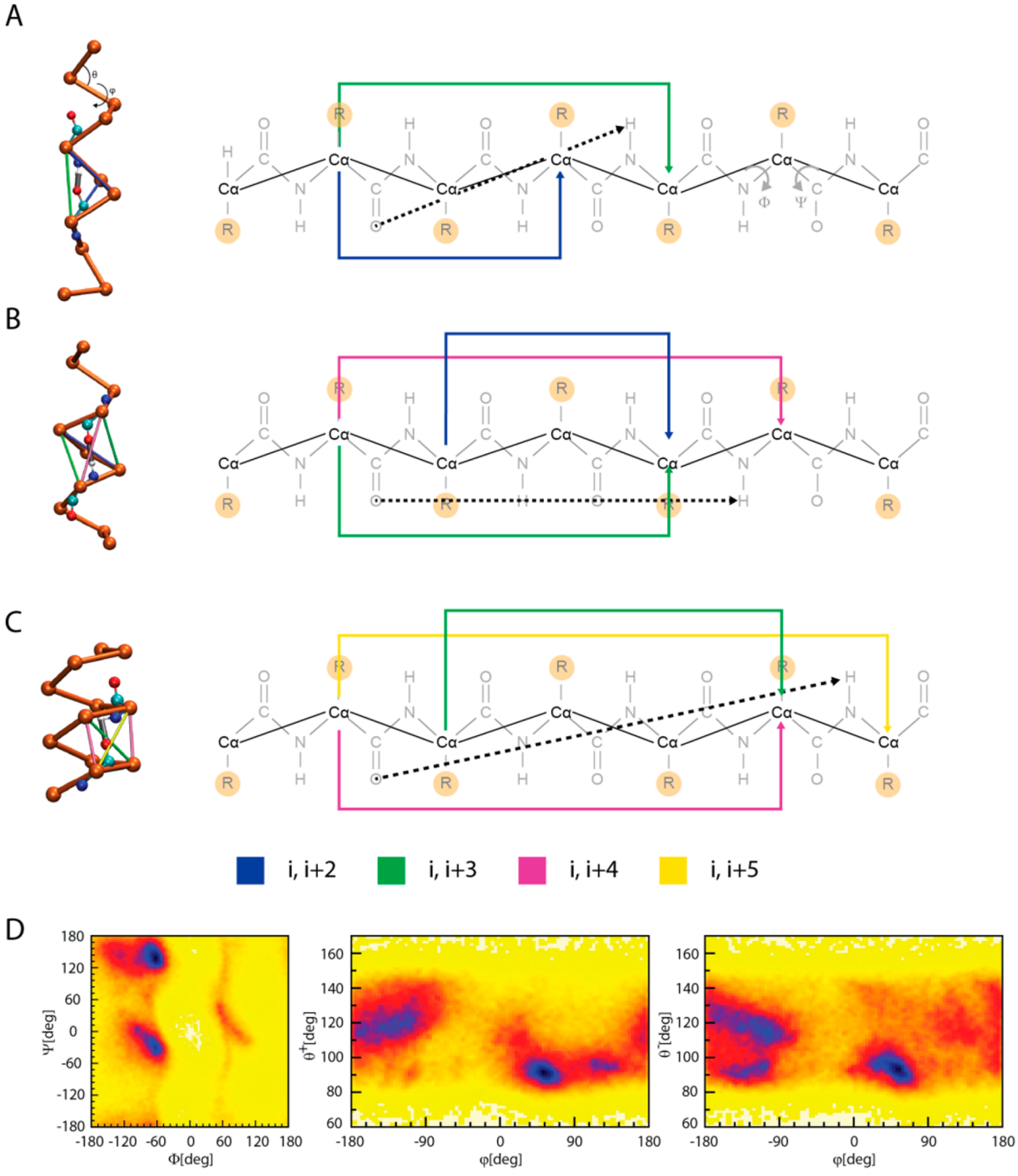


Figure 1. Helix structures and their internal variable correlations. (A–C): 3D structures and 2D schemes of the three helices types, 3_{10} (A), α (B), and π (C), respectively. On the left, the 3D CG structures are reported in beads and sticks representation in orange. Single hydrogen bonds are reported also in atomistic ball and sticks representation (standard coloring), surrounded by the CG bonding network (colored sticks). The conformational angles θ and φ are reported in A. On the right, the 2D CG schemes are superimposed to the atomistic ones (in dark gray). The dashed black arrow indicates the hydrogen bond connection between donor and acceptor in atomistic representation; the colored arrows indicate the corresponding hydrogen bonding network in CG representation. The color code is as in the 3D representations and is reported at the bottom of C. The conformational variables Φ,Ψ are reported in A. (D) Conformational maps for the unstructured proteins data set (see Table S3 for its composition). From left to right, the atomistic version, i.e. the Ramachandran map, and the CG versions, i.e. the $(\theta+,\varphi)$ and $(\theta-,\varphi)$ correlation maps, are shown.

to different CG models).[10] Even restricting to these, the question on how to choose at best the internal coordinates does not have a definitive answer. In fact, ranging from the network models[18] passing through the Go-like[19] and the partially biased[20,21] toward the unbiased models,[22] different authors have chosen different sets of internal variables. The "conformational variables," namely the set $\{\theta_i,\varphi_i\}$ of angles and dihedrals between subsequent Cα's evaluated along the chain (see Figure 1A) are recognized to have a special role. It was shown that under given conditions,[23] there is a one-to-one correspondence between the couples (θ,φ) and the dihedrals (Φ,Ψ) of the atomistic representation. This has two consequences: first, it

dx.doi.org/10.1021/ct5004059 | J. Chem. Theory Comput. 2014, 10, 3885–3895

allows for back-mapping from the CG to the atomistic representation, and second, different secondary structures can be distinguished, as they occupy separated regions in the (θ,φ) correlation plot. This also implies that the latter can be considered the one-bead CG equivalent of the Ramachandran plot (see Figure 1D). For these reasons, FF terms depending on θ_i and φ_i (hereafter "conformational terms") are included in most of the Cα based models as a set of internal variables. Furthermore, among the CG models, the Cα based models are the coarsest that allow the definition of a set of conformational variables with those properties; in this sense, they are called "minimalist."

In addition to conformational terms, specific and nonspecific bead−bead distance dependent terms are usually added to the FF.[24] A variety of possibilities were considered to account for the complex network of local interactions stabilizing proteins structures, which is a particularly difficult task considering the extreme simplification of the model. Anisotropic three-to-four body potentials were introduced to account for the secondary structure stabilizing backbone hydrogen bonds.[25,26] Alternatively, the hydrogen bond anisotropy can be reproduced by a network of isotropic local interactions.[27] It appears, however, that there is a relatively high freedom in the choice of the FF terms.

A general strategy for the definition of the terms and functional forms of the FF and for their parametrization is here proposed. The key concept of this model is simplicity and manageability, as the smallest number of terms and the simplest possible analytical functional form are chosen. Furthermore, each term is clearly related to a physical interaction, giving a straightforward interpretation of the model. This minimalist combined knowledge-based, physics-based approach has the advantage of reducing the freedom in the choice of the terms, providing a possible route to a "standard" form of this class of model, at the same time helping the parametrization, because physical knowledge can be more easily included in each term, at different stages.

The choice of the FF terms is illustrated in the first part of the Results section. This work focuses on the representation of the helical secondary structure class, which is considered as a paradigmatic case, also particularly useful for illustrating the procedure. Besides the most common α-helix, two other kinds of helices are reported: the more elongated 3_{10}-helix, usually found at the beginning or end of α-helices,[28] and the very elusive π-helix,[29] mostly present in the middle of α-helices.[30]

Parameterization was performed with a knowledge-based approach, by applying the Boltzmann inversion,[17] returning statistics based potentials. This requires a preliminary statistical analysis of the proteins' helical structures, which is reported in the second part of the Results section. In the last part of this section, a complete parametrization for the helical structures, which can be used in general dynamics simulations, is provided. A discussion of the results, conclusions, and illustration of future perspectives follow.

METHODS

Database. For each analyzed structure, a data set was built through the permutation of the combination of experimental determination techniques (X-ray or NMR) and secondary structure definition (DSSP or PDB). A total of 14 data sets were built, among which one for each helical structure and one for unstructured fragments was chosen, giving preference to NMR data when possible. Search results were filtered by the RCSB server imposing a structure similarity of less than 30%. A detailed description of the queries, together with a table summarizing the results, is reported in the SI (Table S3). Data were saved at the "minimalist" resolution level.

Simulation Protocol. The simulations were performed using a modified version of the DL_POLY software package,[31] where additional functional forms were introduced. In-house developed software was used to build the input and to analyze the output files. Distributions and correlations of internal variables from simulations were evaluated with SecStAnT. The leapfrog Verlet integrator was used[32] with a time step of 0.01 ps. The temperature was maintained by coupling the system to a Nosè−Hoover thermostat, with a relaxation constant of 0.5 ps. Since at this level no sequence information is included, at each bead was assigned an average amino acid mass (115 au).

Two kinds of simulations were performed for each system: folding and equilibrium structure simulations. In order to reproduce the near-equilibrium dynamic conformation, fragments representing the three helical structures were selected from the PDB. The typical length of α-helix structures ranges between 10 and 20 amino acids, while the 3_{10}-helix has an average length of seven amino acids. The chosen structures are then respectively 20 (extracted from the 1B87 PDB protein) and 11 (2L6U) amino acids long. The NMR and X-ray data sets of π-helices contain fragments up to five and seven amino acids, respectively. In order to perform a simulation with a helix long enough to contain the hydrogen bonding network, an "ideal" structure of 12 amino acids was built with the software Avogadro.[33] The aforementioned structures were considered as representative for each type of helix and used as starting and reference structures for equilibrium simulations. The system was initially relaxed in the local structural minima for 1 ns. An equilibration was then performed for 5 ns, in which the system is gradually heated up to 300 K. Finally, a production run at 300 K for 44 ns was performed.

The folding simulations started from completely unstructured fragments extracted from the PDB (1MEA). A 28mer was used in the case of the α-helix, while 10mers were used for 3_{10} and the π-helix, since these helices are intrinsically shorter. The folding simulation consisted of a local structure optimization of 1 ns, followed by a heating up to 300 K (5 ns) and a production run of 44 ns.

For the simulation without hydrogen bonds, the 28mer 1MEA PDB structure was used. The system was initially relaxed for 4 ns, and then an equilibration was performed for 5 ns in which the system was gradually heated to 300 K. The production run was performed at 300 K for 11 ns.

RESULTS AND DISCUSSION

Model and Force Field Definition. In our minimalist model for proteins, each residue is represented by a single bead centered on the Cα (Figure 1A−C). Due to the rigidity of the peptide bond, the distance between two subsequent Cα's has less than 1% variation; thus it can be safely kept fixed with a restraint at its equilibrium value (3.8 Å). The specific FF form proposed is the following:

$$U = U^{\theta}(\{\theta_i\}) + U^{\phi}(\{\phi_i\}) + U^{\mathrm{hb}}(\{r_{i,i+n}\}) + U^{\mathrm{nb}}(\{r_{i,j>i+n}\}) \quad (1)$$

where the first two are the already mentioned conformational terms describing the intrinsic conformational flexibility of the

dx.doi.org/10.1021/ct5004059 | J. Chem. Theory Comput. 2014, 10, 3885−3895

polypeptide, always present even in unstructured polypeptides. These two terms are thought to include only those effects present for all amino acids and secondary structures, namely the internal constraints due to the rigidity of the peptide bonds, the specific conformational geometry of the sp^3 hybridization of the Cα's, and the steric hindrance effect of the backbone atoms.

The medium-long-range interactions are included in the last term U^{nb}, depending on the distance between all $i-j$ couples of beads not involved in other interactions. U^{nb} includes the excluded volume and shape effects due to side chains, hydrophobicity, electrostatics, hydration−dehydration, and any other effect due to the implicit treatment of the solvent (such as, for instance, approximate treatment of hydrodynamics[34]).

Conversely, the U^{hb} term specifically describing backbone hydrogen bonds is kept separated because of its different formal and parametrization characteristics. This term is strongly dependent on the secondary structure: in unstructured peptides, hydrogen bonds are few and random, while in helical and sheet structures they follow a specific pattern. This gives the possibility to parametrize the U^{hb} term including a priori knowledge of the secondary structure when available, according to the "knowledge-based" force field philosophy. Considering the accuracy of current primary-to-secondary structure algorithms,[35] this approach can have a high predictive power even starting from the primary sequence.

Since most FFs of the Cα based models, though different and various, generally include all of the terms in eq 1 (and sometimes more), this can be considered a "minimal force field." We observe, however, that the proposed approach is not intended to rigorously determine the optimal set of internal variables for this model: other methodologies (e.g., the essential dynamics[36]) are more appropriate for this aim. However, these return unphysical internal coordinate sets and not easily interpretable FF terms, with little possibility of knowledge-based parametrization.

Coherently with the minimalist approach, the simplest possible functional forms were chosen for each term. Since this work is focused on the representation of the helical secondary structure, we chose the conformational terms as single welled:

$$U^{\theta} = \sum_i u^{\theta}(\theta_i) = \sum_i k_{\theta}\frac{1}{2}(\cos\theta_i - \cos\theta_0)^2$$

$$U^{\phi} = \sum_i u^{\phi}(\phi_i) = \sum_i A_{\phi}[1 - \cos(\phi_i - \phi_0)] \quad (2)$$

The optimized set of parameters in eq 2 for the helices (given in the next section) is independent from the specific kind of helix. The hydrogen bonding term is also expressed as a sum of potentials in the Morse form:

$$U^{hb} = \sum_{i,j\in S_{hb}} u_{ij}^{hb}(r_{ij}) = \sum_{i,j\in S_{hb}} \varepsilon_{ij}\{[1 - \exp(-\alpha_{ij}(r_{ij} - r_0^{ij}))]^2 - 1\} \quad (3)$$

The hydrogen bond handling differentiates this model from previous ones:[26] the directionality and multibody nature of the CG hydrogen bonding is obtained using a few (at most three) properly chosen two-body isotropic interactions for each hydrogen bond, which are included into the S_{hb} set, rather than with anisotropic multibody potentials. U^{hb} is the term that distinguishes the three different kinds of helices, namely the elongated 3_{10}-helix, the α-helix, and the shorter π-helix (see Figure 1A−C). These are stabilized by hydrogen bonds between the NH of amino acid i and the C═O of amino acid $i + n$, with $n = 3$ in the 3_{10}-helix, $n = 4$ in the α-helix, and $n = 5$ in the π-helix, respectively (see Figure 1A−C and Table S1 for structural parameters). The different hydrogen bonding topologies are translated into different S_{hb} sets in the minimalist model: specifically, S_{hb} includes $\{(i, i + 2), (i, i + 3)\}$ for the 3_{10}-helix, $\{(i, i + 2), (i, i + 3), (i, i + 4)\}$ for the α-helix, and $\{(i, i + 3), (i, i + 4), (i, i + 5)\}$ for the π-helix (see Table 1). The

Table 1. Helical Statistical Parameters[a]

helix type	3_{10} (NMR DSSP)	α (NMR PDB)	π (X-ray DSSP)
θ (deg)	88.1	90.8	100.0
φ (deg)	64.0	50.4	28.8
$r_{i,i+2}$ (Å)	5.33	5.42	5.78
$r_{i,i+3}$ (Å)	5.61	5.15	5.96
$r_{i,i+4}$ (Å)	8.12	6.05	4.88
$r_{i,i+5}$ (Å)	9.92	8.66	6.32

[a]Structural CG parameters for the three types of helices obtained as the peak values of the corresponding distributions and evaluated with the software Octave.[42]

rationale behind this choice is that these sets of bonds, together with the $(i, i + 1)$ bonds, form tetrahedrons whose vertices are Cα's, including a "real" hydrogen bond in each specific case (see Figure 1A−C, left side). Thus, in practice, restraining these distances is equivalent to controlling the relative orientation of donors and acceptors groups, which allows proper modeling of the helix conformation in the minimalist model. The same procedure was previously adopted to reproduce the helicity of DNA double helices[37] and RNA structures.[38] The parameters of the FF terms corresponding to these distances, optimized for the three helices, are reported in the next section. In a future perspective, the U^{hb} term parametrization can be extended to other secondary structures with the same algorithm, starting from the knowledge of the hydrogen binding topology.

U^{nb} is represented as a sum of two body interactions extended over all of the couples not involved in other interactions. In the specific case of helical structures here considered, U^{hb} already includes up to five nearest neighbor interactions, which are also the shortest and strongest. Consequently, the contribution of U^{nb} is likely to be minor, implying also that this particular case is not the best to optimize the parameters of this FF term. This task is, in fact, beyond the scope of this work and it is addressed in another work recently published by us,[39] which reports a set of statistics based potentials, including medium-long-range interactions and represented by double well potentials. Overall, the functional form of U^{nb} is

$$U^{nb} = \sum_{i,j\notin S_{hb}} u^{nb}(r_{ij})$$

$$u^{nb}(r) = \frac{1}{2}[u_1(r) + (u_2(r) - \Delta) - \sqrt{[u_1(r) - (u_2(r) - \Delta)]^2 + \lambda^2}]$$

$$u_i(r) = \varepsilon_i\{[1 - \exp(-\alpha_i(r - r_i^0))]^2 - 1\} \quad (4)$$

The two combined Morse-shaped wells (u_1 and u_2) include both local interactions (excluded volume, hydrophobicity, side

dx.doi.org/10.1021/ct5004059 | J. Chem. Theory Comput. 2014, 10, 3885−3895

chain hydrogen bonding) and average-long-range interactions (electrostatics). This allows for the possibility to represent side chain conformational changes and hydration−dehydration mediated interactions. Parameters are available both for an "average" amino acid and for an amino-acid-specific model.[39] We adopt here the first one, not to include a bias toward some given sequence. The model parameters are reported in Table 2, and the numerical values for eq 4 are taken from ref 39.

Table 2. Model Parameters[a]

FF	3_{10}-helix	α-helix	π-helix
u^{θ}		k_θ = 10	
		θ_0 = 92	
u^{φ}		A_φ = 1.0	
		φ_0 = 50	
$u^{r(i,i+2)}$	ε = 6.5	ε = 6.5	
	α = 1.4	α = 1.3	
	r_0 = 5.33	r_0 = 5.42	
$u^{r(i,i+3)}$	ε = 3.5	ε = 3.5	ε = 3.5
	α = 0.75	α = 0.8	α = 0.8
	r_0 = 5.61	r_0 = 5.15	r_0 = 5.96
$u^{r(i,i+4)}$		ε = 2.6	ε = 2.6
		α = 0.66	α = 0.66
		r_0 = 6.05	r_0 = 4.88
$u^{r(i,i+5)}$			ε = 2.0
			α = 0.6
			r_0 = 6.32

[a]Optimized parameters of each helical type are listed. Force constant k_θ, A, and ε are in kcal/mol; α in Å^{-1}; equilibrium angles and dihedrals in deg; and equilibrium distances in Å.

Statistical Analysis of Structural Data. The parametrization strategy here proposed consists of reproducing in simulations the statistical distribution of internal variables. For this reason, we first produced accurate distributions for the three different helices. Data sets representing each helical structure were compiled by selecting protein fragments from the RCSB database.[40] For this purpose, SecStAnT, a tool for the automatic extraction of fragments and analysis of internal variables, was employed.[41] In addition to helical structures, a database of unstructured fragments was also created, for the sake of comparison. For each structure, different queries were compiled through the permutation of experimental determination techniques (X-ray or NMR) and secondary structure definition (DSSP or PDB), resulting in 14 different data sets (Table S2 in the SI). For each secondary structure (and for unstructured proteins), a single data set was chosen on the basis of its statistical (number of hits) and physical relevance. For the latter point, when possible, preference was given to NMR data. In fact, they represent the protein conformation in a more physiological environment. The final data set configuration is reported in Table 1.

The internal variables distributions (Figure 2) and the two-variable correlations (Figure 3) were also evaluated with SecStAnT for each data set. The distributions of the unstructured chains data set (Figure 2, solid lines) are dispersed on the allowed range of values for each internal variable, though shoulders and large peaks are visible, reminiscent of a secondary structural tendency. A larger dispersion for unstructured proteins is clearly visible also in the (θ,φ) correlation maps (compare Figure 1D with Figure 3). For unstructured proteins, in addition, the (θ^+,φ) and (θ^-,φ) maps are different, $\theta^{+,-}$ being the angle following or preceding a given dihedral φ. This intrinsic directionality of the structure is however absent in the helices, which show indistinguishable (θ^+,φ) and (θ^-,φ) plots (see Figure S1). Thus, as the (θ,φ) correlation plot for the helices, we considered hereafter the average of (θ^+,φ) and (θ^-,φ).

The α-helix data set shows well-defined and peaked distributions (Figure 2). This is in fact the most abundant and stable secondary structure in proteins, allowing atoms to pack close together with few unfavorable contacts.[28] In contrast, 3_{10}-helices and π-helices data sets display more dispersed distributions. This is particularly evident for the $r_{i,i+4}$ and $r_{i,i+5}$ distance distributions, which in addition, suffer from very low statistics of representative fragments longer than six amino acids. The data set noise is exacerbated by the metastable nature of these helices, allowing more conformational states. In fact, in Table S1 different ideal values for (Φ,Ψ) found in the literature are reported, indicating the controversial issue of these secondary structures. In particular, the π-helix has been postulated to be the less stable for unfavorable steric contacts and unfavorable relative orientations of donor−acceptor in hydrogen bonding. We remark that, due to the very low statistical weight, the data for the π-helices are not used for the determination of θ_0 and φ_0 values, while they are used for the energetic terms as well as a benchmark to test the predictive power of the model.

Nevertheless, the conformational variables (especially θ) have similar values for the three helices, while the peaks in the $r_{i,i+n}$, especially with n = 4,5, are farther apart. This enforces our choice of using common values of θ_0 and φ_0 for the three structures, assigning to $u_{i,i+n}^{\mathrm{hb}}$ the task of producing different distributions/correlations for the different helices. To this aim, we observe that the FF term $u_{i,i+2}^{\mathrm{hb}}$ are somehow redundant with u^{θ}, and $u_{i,i+3}^{\mathrm{hb}}$ with u^{φ}, since the following relationships stand between these variables:

$$r_{i,i+2} = 2l\,\sin(\theta/2)$$
$$r_{i,i+3}^2 = l^2[4\,\sin^2(\varphi/2)\,\sin^2\theta + (1 - 2\cos\theta)^2] \quad (5)$$

(with l = 3.8 Å) in the first equation being always valid for trans peptide bonds, the second requiring additionally uniformity of the secondary structure along the chain. This condition, also related to the identity $(\theta^{\pm},\varphi)$ plots, is however well satisfied in helical structures. This also implies that no term to account for directionality is necessary in the FF for helices. Problems related to the nonuniformity and directionality are beyond the scope of this work and will be discussed in a forthcoming paper.[43]

Having chosen common u^{θ} and u^{φ} to all helices, the U^{hb} terms must account for all the differences in the distributions and in the correlation plots. These are (i) the different centers and width of all the $r_{i,i+n}$ distributions and (ii) the (θ,φ) plots for different helices. The (θ,φ) plots from experimental data sets in Figure 3 show that the three helices occupy partially superimposed areas, with different shapes, more or less disperse, roughly ellipsoidal, and with different "slopes" of the main ellipsoid axes. These differences are here reproduced by different networks of H-bonding. Again, to support the choice of the hydrogen bonding network, distance-constrained (θ,φ) values are superimposed to each helical (θ,φ) correlation plot. Namely, the blue lines are obtained constraining the $r_{i,i+2}$ distance (by imposing constant values on the left-hand side of the top eqs 5 and solving for θ); the green lines are obtained

dx.doi.org/10.1021/ct5004059 | J. Chem. Theory Comput. 2014, 10, 3885−3895

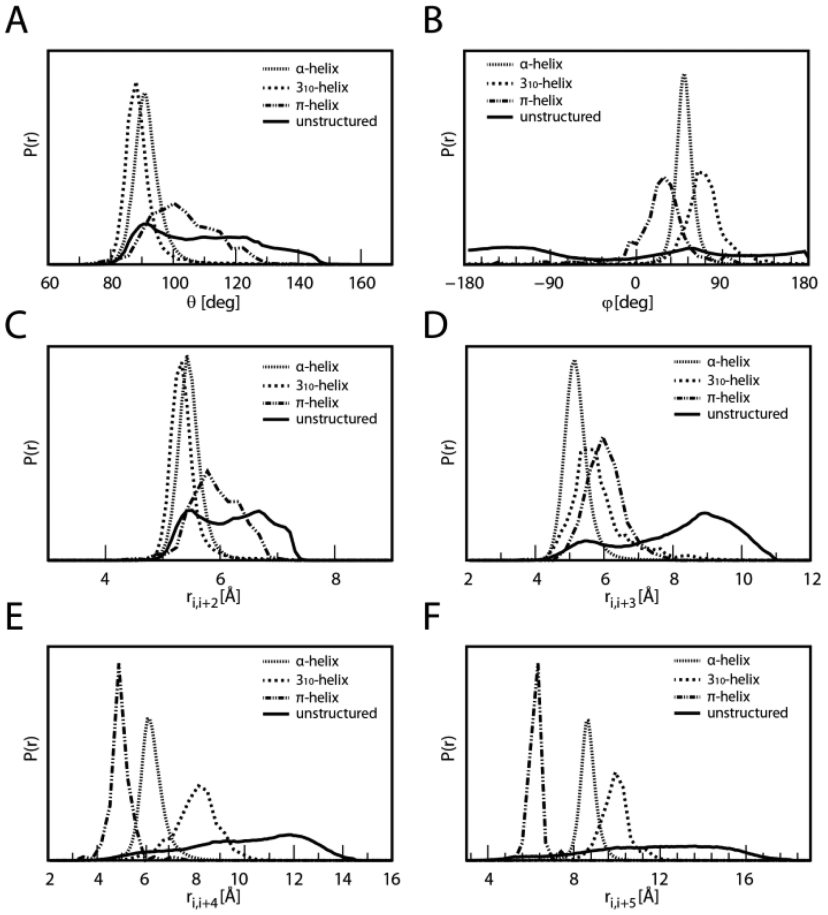


Figure 2. Statistical distributions of internal variables. Probability distribution of θ (A), φ (B), $r_{i,i+2}$ (C), $r_{i,i+3}$ (D), $r_{i,i+4}$ (E), $r_{i,i+5}$ (F). Different lines correspond to different secondary structure data sets, as reported within the plots (see Table 1 for the data sets properties). Distributions are normalized according to their integrals.

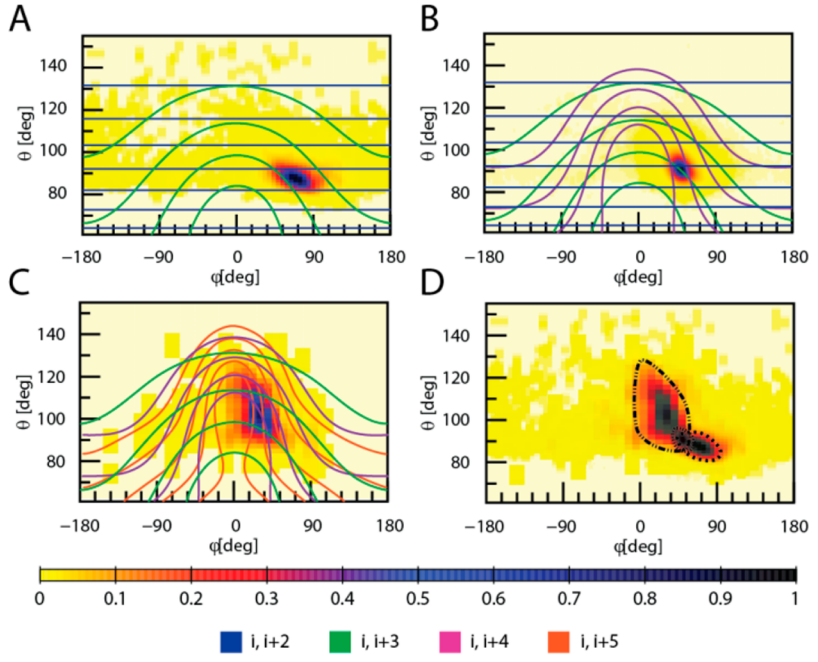


Figure 3. (θ,φ) correlation plots from experimental data sets. (A) Correlation plot for the 3_{10}-helix, average of (θ^{+},φ) and (θ^{-},φ). $r_{i,i+2}$ = const and $r_{i,i+3}$ = const lines are reported in blue and green, respectively. (B) Same for the α-helix. The $r_{i,i+4}$ = const line is additionally reported in purple. (C) Same for the π-helix. In addition, the $r_{i,i+5}$ = const line is reported in orange. (D) Superimposed correlation plots. The three helices correlation plots are contoured with lines of different dashing, as previously reported (Figure 2).

dx.doi.org/10.1021/ct5004059 | J. Chem. Theory Comput. 2014, 10, 3885−3895

constraining $r_{i,i+3}$ (by imposing constant values on the left-hand side of the bottom eqs 5 and then solving $\theta(\varphi)$); the purple and orange lines correspond respectively to constrained values of $r_{i,i+4}$ and $r_{i,i+5}$ and do not have a unique solution in terms of (θ,φ). These were obtained numerically by restraining the corresponding distances in the model, searching for the local minima of the corresponding structures.

For each helix type, only the lines corresponding to the involved hydrogen bonding network are reported. As can be seen from Figure 3A−C, passing from the 3_{10}- to the α- and the π-helix, the slope of the correlation plot increases. The same happens passing from the constrain lines $r_{i,i+n}$ as n increases. Consequently, the $r_{i,i+2}$ and $r_{i,i+3}$ lines superimpose to the (θ,φ) plot of the 3_{10}-helix, while the α-helix slope is more steep, indicating that also the $r_{i,i+4}$ line must be added. Coherently, the π-helix plot slope is even steeper, indicating that also the $r_{i,i+5}$ interactions are necessary, while in this case the "horizontal" $r_{i,i+2}$ interactions are probably not involved. In summary, the hydrogen bonding networks chosen for each helical type are also coherent with (θ,φ) correlations.

Parameter Optimization. In consideration of what was reported in the previous section, the parameters are optimized using the following criteria: (i) a unique parametrization is used for u^{θ} and u^{φ} terms, targeting the average centers and widths of the θ and φ distributions of the three helices; (ii) the parameters for the U^{hb} term are helix-specific and optimized by targeting the distributions of the $r_{i,i+2}$... $r_{i,i+5}$ distances and the (θ,φ) correlations. The unstructured distribution (restricted to the helical region) is used as a reference case with negligible hydrogen bonding.

The initial guess for the parameters of each force field terms was obtained fitting the direct Boltzmann inverted of the corresponding distribution (produced with SecStAnT) with the functional forms of eqs 2−4. The values of the distribution centers of θ and φ were used to fit a common value for the equilibrium parameters of u^{θ} and u^{φ}. A common value for the three helices force constants k_{θ} was chosen such as to include the width of the three distributions, allowing the three helical conformations. This also leads to a narrow range of acceptable values, within which k_{θ} was finally optimized upon the addition of U^{hb} terms (see Table 2). The distributions for the π-helix were marginally considered in the H-bonding term definition (i.e., given less weight in the averages), due to their low statistics and reliability. Knowledge-based and statistical-derived data were used jointly in the parametrization. The first guesses for u^{hb} were obtained with the same procedure, however separately for the three helices, including only the terms pertaining to each helix as explained previously. Also for the force constant of these terms, an acceptable range was defined; for example, each hydrogen bond energy lies in the interval ∼7−10 kcal/mol, in line with the experimentally known values.[44]

The force constants are then optimized following an iterative BI procedure (separately for each kind of helix), comparing at each step all the internal variable distributions and correlations obtained by the simulations performed with the DL_POLY software package[31] with the reference ones of Figures 2 and 3. The parameters were restrained in the acceptable ranges by a trial and error approach, driven by the knowledge of physics and chemistry. For example, the energetic contribution of the U^{hb} terms was chosen in order to be coherent with the experimental value. The optimized FF parameters are shown in Table 2, while the comparison between the reference and simulation distributions are reported in Figures 4 and 5 and discussed in the next section. For a detailed description of the parametrization process, please refer to Figure S5 in the SI.

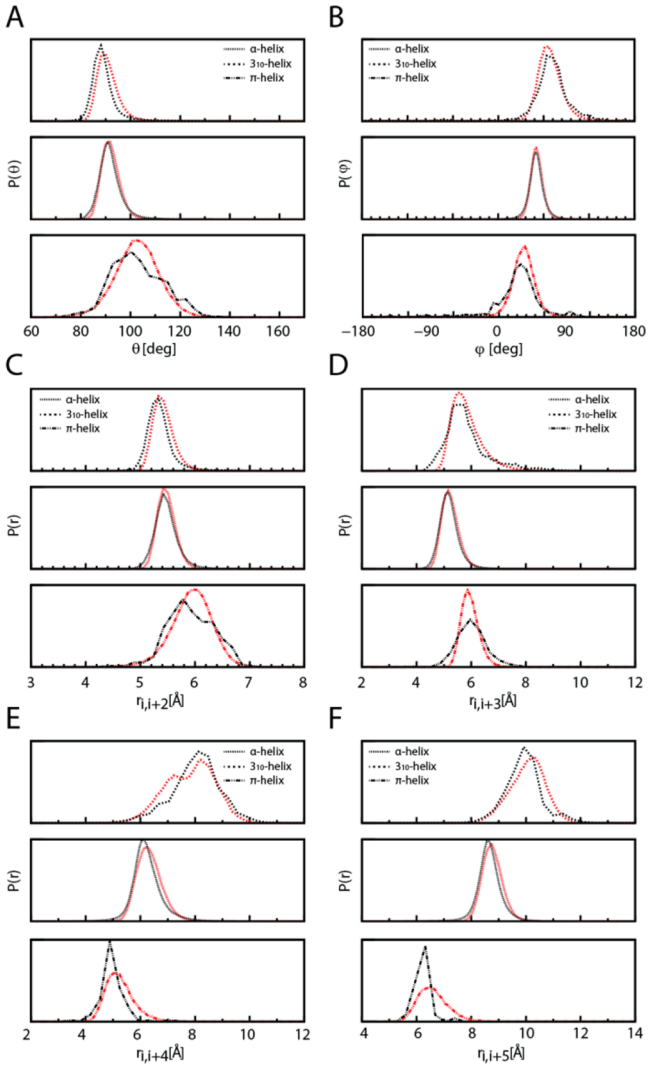


Figure 4. Comparison of experimental and simulation internal variable distributions. The probability distribution of θ (A), φ (B), $r_{i,i+2}$ (C), $r_{i,i+3}$ (D), $r_{i,i+4}$ (E), and $r_{i,i+5}$ (F) of experimental (black line) and simulation (red line) internal variables is compared. Different lines correspond to different secondary structure data sets, as reported within the plots.

Equilibrium Dynamics. To assess the quality of the FFs, two sets of simulations were performed for each type of helix, namely equilibrium dynamics and folding dynamics. Starting from the equilibrium dynamics, Figures 4 and 5 report the comparison between experimental and simulations statistical distributions and correlations, respectively. The distributions/correlations from simulations were evaluated in the production run part (see Figure S3 in the SI for the equilibrium dynamic energies plots). The agreement between simulation and experimental distribution (Figure 4) is excellent for the α-helix and good for the 3_{10}-helix. In the latter case, though the maximum of the distribution is often well reproduced, some discrepancies are observed in the dispersion, as the experimental distributions show more irregularities than the simulation ones. We note that the 3_{10}-helix data sets often

dx.doi.org/10.1021/ct5004059 | J. Chem. Theory Comput. 2014, 10, 3885−3895

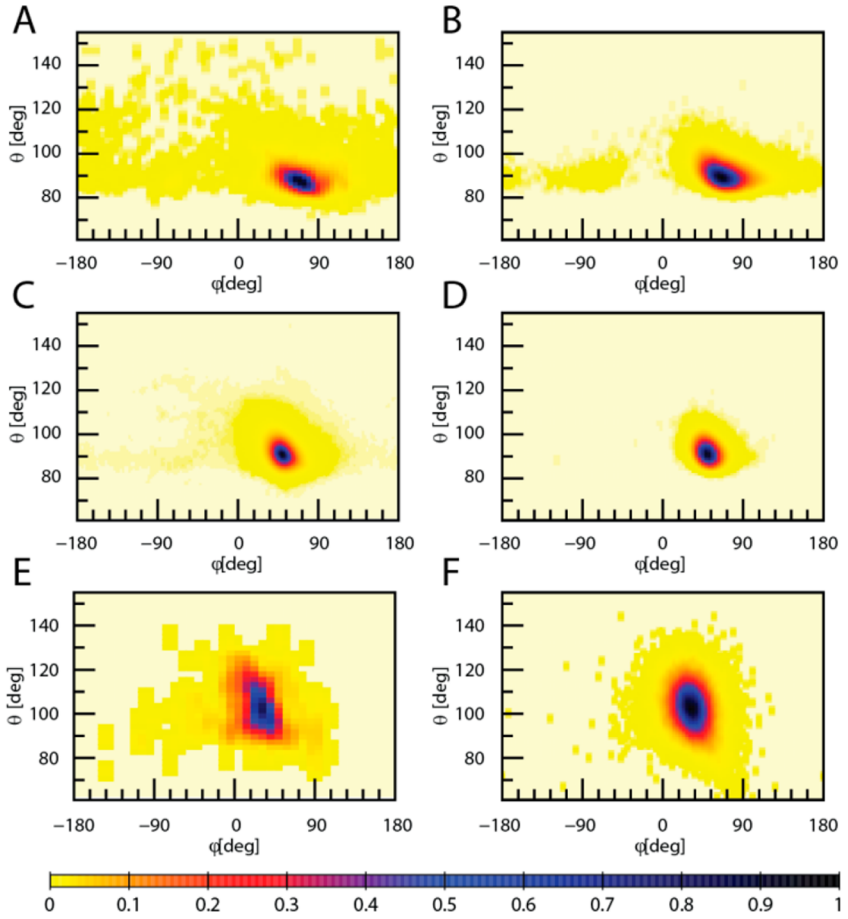


Figure 5. Comparison of (θ,φ) experimental and simulation correlations plots. (A, B) 3_{10}-helix; (C, D) α-helix; (D,E) π-helix. (A, C, E) Experimentally obtained data. (B, D, F) Simulated data from the equilibrium dynamics run.

contain wrongly identified α-helices, resulting in spurious data, which are very difficult to eliminate. In the $r_{i,i+4}$ distance distribution, for instance, the presence of α-helix residual manifests as an enlargement of the left-hand side. Noticeably, the simulations also reproduce a small α-like secondary peak, indicating that metastable α-like structures are explored by the model. The agreement between simulation and experiment is less good for π-helices. Again, the center of the distribution is reproduced, but the dispersion is not completely quantitative, especially for the distances. However, as stated, the statistics for π-helices are very low; thus the reference data on the distribution width cannot be considered very accurate. Interestingly, although the proposed model does not contain specific bistable terms able to induce transitions between different secondary structures, some spontaneous unfolding events can be observed in the three helical models (Figure S4).

The comparison between simulation and experiment is particularly interesting for what concerns the correlation plots (Figure 5). The location and average width of the occupied regions are reproduced coherently with the agreement of simulation−experiment in the corresponding single variable θ and φ distribution. In addition, the elongated shape and the slopes of the plots are also quantitatively reproduced. This is an effect of the proper choice of the hydrogen bonding networks for each helix and of their fine-tuned parametrization, since the u^{θ} and u^{φ} potentials are the same for the three helices. Again, the less quantitative agreement is for the π-helix. Very interestingly, the model reproduces the relatively larger tendency of the 3_{10}-helix (and to a lesser extent, of the π-helix) to form also left handed structures, which can be observed as populated areas at negative φ values. These are observed both in the simulated and in the experimental plots, which is a further validation of the model.

Folding Dynamics. Figure 6 reports the analysis of the folding dynamics for the three helices (black = α, red = 3_{10}, green = π). The temperature is raised with the same protocol (Figure 6 A), and in each case, the folding transition is seen as a stepwise decreasing of the internal energy and of the RSMD with respect to the folded state, which is reached after ∼7 ns in the three cases. The internal energy per residue in the three cases reflects the order of stability of the three helices, the α-helices being the most stable, followed by the 3_{10} and the π. By analyzing the energy variation of each energy term (panel B), it appears clear that in each case the main energy gain is due to the hydrogen bonding terms, which brings more than 90% of the total amount of energy gain upon folding (Table 3). This reflects our choice of imputing most of the responsibility of secondary structure formation to the hydrogen bonding terms. To further underline their importance, a simulation of the 28mer without hydrogen bonds was performed, keeping the other terms unchanged. The system forms a random coil, which however should not be compared with natural random coils or unstructured proteins. In fact, the residual helical bias in the bonding terms force it to occupy preferentially the right-handed helical regions. However, the occupied area is wider and less defined than in the helix simulations and does not assume the

dx.doi.org/10.1021/ct5004059 | J. Chem. Theory Comput. 2014, 10, 3885−3895

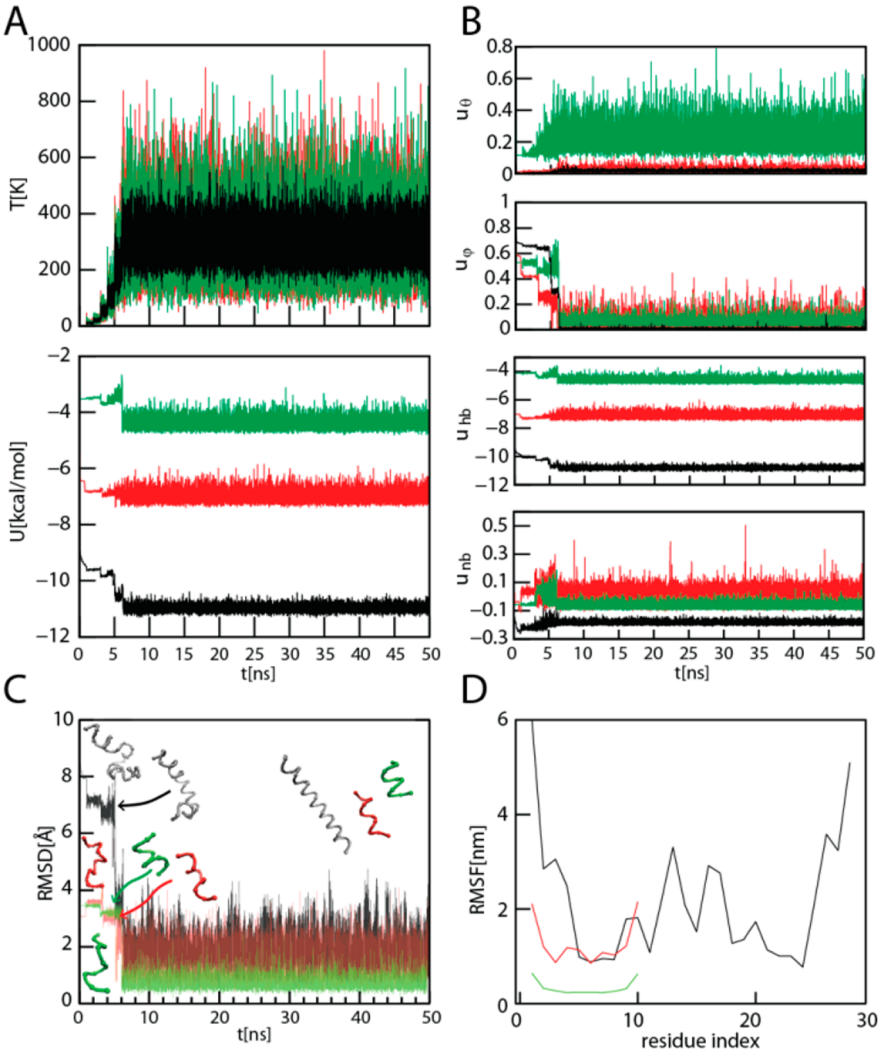


Figure 6. Folding dynamics. Folding simulation analysis of the three helices. Color code: black = α-helix, red = 3_{10}-helix, green = π-helix. (A) Simulation temperature (top) and total energy (bottom). (B) Force field energy contribution. From top to bottom: u^{θ}, u^{φ}, u^{hb}, u^{nb}. (C) Root mean squared deviation with respect to the configuration reached at 10 ns, after folding. Snapshots from the folding simulations are reported in the RSMD plot. The arrows indicate their corresponding time frame. (D) RMSF evaluated in the interval 10−50 ns. The energies are normalized on the number of beads.

correct shape and slope, as it is unable to reproduce the correct (θ,φ) correlations (see Figure S2 in SI). The conformational terms bring a small contribution, which is, however, more destabilizing for the π- and 3_{10}-helices with respect to α, as expected since the α-helix is the one whose structural parameters are nearer to the average ones, used for the parametrization (see also Table 3). As anticipated, the U^{nb} contribution is rather small, being however stabilizing for the α-helix, little stabilizing for π, and almost null for 3_{10}. Panel C reports the RSMD along the simulation with respect to the unfolded starting structure. The transition kinetics is clearly visible: each of the three helices passes through intermediate states, as reported in the figure. Finally, panel D reports the RSMF evaluated in the folded region. A W-shaped trend, typical of tubular-shaped structures, can be observed. This is due to the bending motion with two nodal points near the extremities and is normally observed in helical structures.[45]

■ CONCLUSIONS AND PERSPECTIVES

From the analysis of the simulations, the following conclusions can be drawn: (i) the structural characteristics of the three helical types are quantitatively and qualitatively reproduced, by the optimized parametrization both considering the single variable distributions and the correlations; (ii) this effect is obtained using common intrinsic conformational potentials and specific hydrogen bonding terms, mimicking the real interactions with a network of two or three pseudobonds. The different hydrogen bonding network identifies the helical type and, in turn, is defined once the secondary structure is known. This is a peculiarity of the proposed model: in previous similar ones,[20,22] the secondary structure is encoded both in the conformational and in the non-bonded interactions (including also hydrogen bonding), leading to a very complex primary and secondary structure-dependent parametrization. Here, we limited the secondary structure encoding to the hydrogen bonding term, which is both more adherent to reality and simpler and, consequently, more easily generalizable. In addition, at variance with other similar models that use complex anisotropic potentials,[25,26] the structure is accurately reproduced with simple isotropic potentials, efficiently handled in common molecular dynamics codes. The local directionality of the hydrogen bonding is obtained by the combination of two

Table 3. Folding Dynamics Total Energies and Terms Contribution[a]

		3_{10}-helix	α-helix	π-helix
average energies (kcal/mol)	U_{tot}	−69.53	−303	−42.64
	U_{θ}	0.21	0.45	2.19
	U_{φ}	1.05	2.3	1.13
	U_{hb}	−70.9	−300.5	−45.34
	U_{nb}	0.11	−5.24	−0.62
normalized energies (kcal/mol)	u_{tot}	−6.95	−10.82	−4.26
	u_{θ}	0.02	0.02	0.22
	u_{φ}	0.10	0.08	0.11
	u_{hb}	−7.09	−10.73	−4.53
	u_{nb}	0.01	−0.19	−0.06
Δ energies (kcal/mol)	ΔU_{tot}	18.59	97.11	20.06
	ΔU_{θ}	0.44	1.20	0.15
	ΔU_{φ}	5.63	24.22	5.36
	ΔU_{hb}	12.25	70.29	14.52
	ΔU_{nb}	0.27	1.40	0.04
normalized Δ energies (kcal/mol)	Δu_{tot}	1.85	3.47	2.0
	Δu_{θ}	0.04	0.04	0.01
	Δu_{φ}	0.56	0.86	0.53
	Δu_{hb}	1.22	2.51	1.45
	Δu_{nb}	0.027	0.05	0.004

[a]For each helical type, the average, normalized, Δ, and Δ-normalized energies are shown. Energies are normalized on the number of residues. Δ energies are considered between the average folded state energy and the first simulation frame.

or three isotropic bonds with specifically tuned equilibrium distances. As a result, the experimental distributions and correlations of the internal variables of the three helices (including the π-helix, within the error due to the low statistic) are accurately reproduced. Even if there is no guarantee that the relative population in structural data sets accurately reflects the relative stability, the BI-based analysis allowed accurate locatation of the "structural parameters" (i.e., equilibrium distances and well width) of each single helical structure. The BI-derived well depths are instead used as starting values and then adjusted based also on physicochemical considerations. In general, the parameters can be tuned to allow structural transitions on the basis of physical and statistical factors.

Though this work focuses on the helices, the parametrization strategy here outlined is straightforwardly extendable to any secondary structure. In perspective, the u^{θ} and u^{φ} terms shall account for the intrinsic tendency of the polypeptide for any secondary structural state and even for unstructured proteins. To this aim, it should be extended to be able to occupy also other regions of the (θ,φ) plane. Then these potentials should remain energetically weaker than the hydrogen-bond-representing terms, to which the task of maintaining the secondary arrangement is imputed. As in the case of the helices, different types of sheets can be identified and distinguished on the basis of hydrogen bond topology, and their parametrization can be addressed, reproducing the directionality and anisotropy with a combination of three bonds for each real hydrogen bond. An extension of the FF to include other kinds of networks is currently under evaluation.[46]

At this level, the model requires the knowledge of the secondary structure, which however enters only in the hydrogen bonding topology definition. This can be assigned on the basis of the primary structure using available primary to secondary structure prediction algorithms.[35] Then there would be the possibility of predicting the folding, or assigning "uncertain" secondary hierarchical levels by tuning the relative weight of the hydrogen bonding networks pertaining to different secondary structures.

In this work, the FF parametrization is performed with an iterative user-driven BI. However, the process of parametrization of accurate and predictive FFs requires a huge amount of work, as it is often based on a manual process of trial and error. Furthermore, the extension of the potential will introduce additional parameters, thus hampering the method. An automatic procedure could help reduce the time and amount of work and, at the same time, take into account the redundancy between certain terms and their relative weight in the overall FF. The building of an automatic parametrization tool based on the principles here described is currently in progress.[47]

In conclusion, we propose a model able to reproduce accurately the helical polypeptides and outline a general strategy for extending it to the other secondary structures. The very simple terms involved allow at the same time a clear physical interpretation and a parametrization based on a (even probabilistic) knowledge of the secondary structure. Given its simplicity, the FF is easily implementable in different molecular dynamics software and amenable to being included in multiscale procedures, both "serial" and "parallel," such as for instance the inclusion of the peptide in a "cytoplasm system" where the crowders are represented with large soft spheres,[39] though reaching the macroscopic scale in the simulations.

■ ASSOCIATED CONTENT

Supporting Information

Topology and structural parameters of the different kind of helices; (θ^{+},φ) and (θ^{-},φ) correlation plots comparison; distribution parameters for the three different types of helices for the CG structures; comparison of (θ^{+-},φ) experimental and simulation correlations plots of a 28mer without hydrogen bonds; analysis of thermalization and production run of the three helices equilibrium dynamics; unfolding events in folding dynamics; and a parameterization flowchart. These materials are available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Phone: +39-050-509.699. Fax: +39-050-509.417. E-mail: giuseppe.maccari@iit.it.

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank Dr. Fabio Trovato and Dr. Paolo Mereghetti for useful discussions and for help in the software design and programming. We acknowledge the support of Platform "Computation" of IIT (Italian Institute of Technology).

■ REFERENCES

(1) Karplus, M.; McCammon, J. A. *Nat. Struct. Biol.* **2002**, *9*, 646−52.

(2) Shaw, D. E.; Chao, J. C.; Eastwood, M. P.; Gagliardo, J.; Grossman, J. P.; Ho, C. R.; Lerardi, D. J.; Kolossváry, I.; Klepeis, J. L.; Layman, T.; McLeavey, C.; Deneroff, M. M.; Moraes, M. A.; Mueller, R.; Priest, E. C.; Shan, Y.; Spengler, J.; Theobald, M.; Towles, B.; Wang, S. C.; Dror, R. O.; Kuskin, J. S.; Larson, R. H.; Salmon, J. K.; Young, C.; Batson, B.; Bowers, K. J. *Commun. ACM* **2008**, *51*, 91.
(3) Freddolino, P. L.; Liu, F.; Gruebele, M.; Schulten, K. *Biophys. J.* **2008**, *94*, L75−7.
(4) Lindorff-Larsen, K.; Maragakis, P.; Piana, S.; Eastwood, M. P.; Dror, R. O.; Shaw, D. E. *PLoS One* **2012**, *7*, e32131.
(5) Chaudret, R.; Gresh, N.; Cisneros, G. A.; Scemama, A.; Piquemal, J.-P. *Can. J. Chem.* **2013**, 1−7.
(6) Zhao, D.-X.; Liu, C.; Wang, F.-F.; Yu, C.-Y.; Gong, L.-D.; Liu, S.-B.; Yang, Z.-Z. *J. Chem. Theory Comput.* **2010**, *6*, 795−804.
(7) Tozzini, V. *Curr. Opin. Struct. Biol.* **2005**, *15*, 144−50.
(8) Maupetit, J.; Tuffery, P.; Derreumaux, P. *Proteins* **2007**, *69*, 394−408.
(9) Ha-Duong, T. *J. Chem. Theory Comput.* **2010**, *6*, 761−773.
(10) Marrink, S. J.; Risselada, H. J.; Yefimov, S.; Tieleman, D. P.; de Vries, A. H. *J. Phys. Chem. B* **2007**, *111*, 7812−24.
(11) Tozzini, V. *Q. Rev. Biophys.* **2010**, *43*, 333−71.
(12) Trovato, F.; Tozzini, V. In *AIP Conference Proceedings*; American Institute of Physics: Pavia, Italy, 2012; Vol. *1456*, pp 187−200.
(13) Smith, J. C.; Roux, B. *Structure* **2013**, *21*, 2102−5.
(14) Tadmor, E. B.; Elliott, R. S.; Sethna, J. P.; Miller, R. E.; Becker, C. A. *JOM* **2011**, *63*, 17−17.
(15) Chaimovich, A.; Shell, M. S. *J. Chem. Phys.* **2011**, *134*, 094112.
(16) Lyubartsev, A.; Laaksonen, A. *Phys. Rev. E: Stat. Phys., Plasmas, Fluids, Relat. Interdiscip. Top.* **1995**, *52*, 3730−3737.
(17) Reith, D.; Pütz, M.; Müller-Plathe, F. *J. Comput. Chem.* **2003**, *24*, 1624−36.
(18) Bahar, I.; Lezon, T. R.; Yang, L.-W.; Eyal, E. *Annu. Rev. Biophys.* **2010**, *39*, 23−42.
(19) Clementi, C. *Curr. Opin. Struct. Biol.* **2008**, *18*, 10−5.
(20) Trylska, J.; Tozzini, V.; McCammon, J. A. *Biophys. J.* **2005**, *89*, 1455−63.
(21) Voltz, K.; Trylska, J.; Tozzini, V.; Kurkal-Siebert, V.; Langowski, J.; Smith, J. *J. Comput. Chem.* **2008**, *29*, 1429−39.
(22) Sorenson, J. M.; Head-Gordon, T. *J. Comput. Biol.* **2000**, *7*, 469−81.
(23) Tozzini, V.; Rocchia, W.; McCammon, J. A. *J. Chem. Theory Comput.* **2006**, *2*, 667−673.
(24) Tozzini, V. *Acc. Chem. Res.* **2010**, *43*, 220−30.
(25) Alemani, D.; Collu, F.; Cascella, M.; Dal Peraro, M. *J. Chem. Theory Comput.* **2010**, *6*, 315−324.
(26) Yap, E.-H.; Fawzi, N. L.; Head-Gordon, T. *Proteins* **2008**, *70*, 626−38.
(27) Tozzini, V.; Trylska, J.; Chang, C.; McCammon, J. A. *J. Struct. Biol.* **2007**, *157*, 606−15.
(28) Crisma, M.; Formaggio, F.; Moretto, A.; Toniolo, C. *Biopolymers* **2006**, *84*, 3−12.
(29) Weaver, T. M. *Protein Sci.* **2000**, *9*, 201−6.
(30) Cooley, R. B.; Arp, D. J.; Karplus, P. A. *J. Mol. Biol.* **2010**, *404*, 232−46.
(31) Todorov, I. T.; Smith, W. *Philos. Trans. R. Soc., A* **2004**, *362*, 1835−52.
(32) Svanberg, M. *Mol. Phys.* **1997**, *92*, 1085−1088.
(33) Hanwell, M. D.; Curtis, D. E.; Lonie, D. C.; Vandermeersch, T.; Zurek, E.; Hutchison, G. R. *J. Cheminform.* **2012**, *4*, 17.
(34) Trovato, F.; Tozzini, V. Under review.
(35) Pirovano, W.; Heringa, J. *Methods Mol. Biol.* **2010**, *609*, 327−48.
(36) Amadei, A.; Linssen, A. B.; Berendsen, H. J. *Proteins* **1993**, *17*, 412−25.
(37) Trovato, F.; Tozzini, V. *J. Phys. Chem. B* **2008**, *112*, 13197−200.
(38) Leonarski, F.; Trovato, F.; Tozzini, V.; Leś, A.; Trylska, J. *J. Chem. Theory Comput.* **2013**, *9*, 4874−4889.
(39) Trovato, F.; Nifosì, R.; Di Fenza, A.; Tozzini, V. *Macromolecules* **2013**, *46*, 8311−8322.
(40) Rose, P. W.; Bi, C.; Bluhm, W. F.; Christie, C. H.; Dimitropoulos, D.; Dutta, S.; Green, R. K.; Goodsell, D. S.; Prlic, A.; Quesada, M.; Quinn, G. B.; Ramos, A. G.; Westbrook, J. D.; Young, J.; Zardecki, C.; Berman, H. M.; Bourne, P. E. *Nucleic Acids Res.* **2013**, *41*, D475−82.
(41) Maccari, G.; Spampinato, G. L. B.; Tozzini, V. *Bioinformatics* **2014**, *30*, 668−74.
(42) Eaton, J. W.; Bateman, D.; Hauberg, S. *{GNU Octave} Version 3.0.1 Manual: A High-Level Interactive Language for Numerical Computations*; CreateSpace Independent Publishing Platform: Seattle, 2009.
(43) Giuntoli, A.; Tozzini, V. In preparation.
(44) Smith; Wood, E. J. *Biological Molecules; Molecular and Cell Biochemistry*, 1st ed.; Springer: New York, 1991; Chapter 2.
(45) Wang, Y.-T.; Hsu, H.-J.; Fischer, W. B. *Springerplus* **2013**, *2*, 324.
(46) Pesce, L.; Tozzini, V. In preparation.
(47) Mereghetti, P.; Maccari, G.; Spampinato, G. L. B.; Tozzini, V. In preparation.