Modeling the Free Energy of Polypeptides in Different Environments

Sara Furlan and Giovanni La Penna*

Consiglio nazionale delle ricerche, Istituto di chimica dei composti organo-metallici, Via Madonna del Piano 10, 50019 Sesto Fiorentino, Firenze, Italy

Angelo Perico

Consiglio nazionale delle ricerche, Istituto per lo studio delle macromolecole, Via De Marini 6, 16149 Genova, Italy

Received October 4, 2007; Revised Manuscript Received December 27, 2007

ABSTRACT: An estimate of the free energy for increasing the content of helical and elongated backbone segments in polypeptides is given. Computer simulations of "reasonable" random walks of isolated single chain all-atom models of a polypeptide are performed, and the potential energy of the obtained configurations is refined by using mean-field models for the molecular environment. The entropic contribution is computed assuming that entropy depends only on the enumeration of configurations (density of states) for each single chain. The free energy estimate is applied on several homo-polypeptides (X_{40} with X = G, A, V, T, K, E) and to the Aβ(1–40) peptide involved in Alzheimer's disease. Also, several peptides with the Aβ(1–40) sequence randomly scrambled and the mini-protein villin headpiece HP(1–36) are analyzed.

1. Introduction

It has been recognized that protein aggregation, and, in many cases, the consequent formation of amyloid fibrils, is a generic property of polypeptide chains.[1,2] Once formed, amyloid fibrils are very stable because of the large number of intermolecular hydrogen bonds that must be broken in order to extract a single polypeptide chain from the fibril. In the living cell, the formation of unwanted materials, like amyloid fibrils, is prevented through protection of those proteins designed for specific interactions and for being functional in a soluble form. This protection is provided by a defined structure (the native structure), where the presence of a hydrophobic core, formed by residues hidden in the interior of a globular shape, avoids the exposure of the protein backbone to a-specific intermolecular interactions. The formation of amyloid fibrils is, therefore, a second fundamental way to organize proteins, an alternative to the folding into the native structure, when this latter is defined by the protein sequence.

Some amino acid sequences are more prone to aggregation than others. Within the above perspective, the aggregation propensity can be the indirect effect of the absence of a folding propensity. This effect has been called the "inverse side chain effect"[3] or "natively disordered model":[2] some sequences, when merged into a given environment, do not find a pathway to fold into protected structures and are therefore more prone to form a-specific backbone interactions that can eventually drive protein oligomerization and interactions with the many other proteins present nearby.

One of the peptides that are most extensively studied, because abundant in the fibril formed in the Alzheimer's disease, is the Aβ(1–40) peptide. Despite the large amount of information available on this and similar peptides,[4] there is no clear understanding of which are the reasons of its propensity to be included in amyloid fibrils. In this work we aim at designing a model that is able to monitor the native disorder of a given sequence when this is merged in water or in a hydrophobic environment. We compare the conformational landscape of a single chain of the Aβ(1–40) peptide with several other sequences, including some sequences that are known to adopt a defined structure in given molecular environments. By using a single-chain model, we find that the Aβ(1–40) peptide is more disordered in the water environment than in a hydrophobic environment where the chain preferentially adopts an extended helical structure.

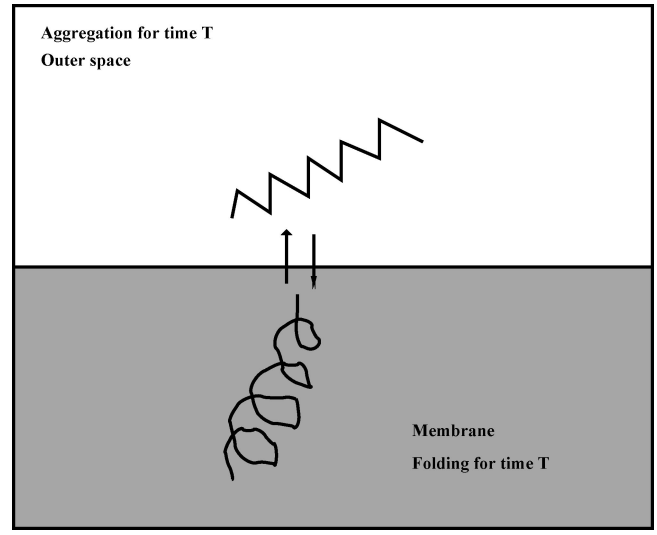

Figure 1. Schematic picture of the simulation of a random walk for a given molecular chain with $c_\alpha < 0$ (lower region) and with $c_\beta < 0$ (upper region) switched on, alternatively, in eqs 11 and 12 for the fictitious time T ($M_c N_T$ saved configurations during N_T cycles of length $2T$ of switching on/off the external potential).

In the model, collective properties that monitor the secondary structure of peptides are used. Because of the relevance of helical and elongated segments in the structure of the Aβ peptides, we focused on the construction of extended helical segments or, alternatively, of extended elongated segments in different sequences of 40 amino acids.

2. Theory and Methods

The atomic configuration for molecule a is defined by the position of the constituting N_a atoms, $\mathbf{r} = \vec{r}_1, ..., \vec{r}_{Na}$. Given a configuration-dependent property $X(\mathbf{r})$, we want to compute

* Corresponding author. E-mail: glapenna@iccom.cnr.it.

10.1021/ma7022155 CCC: $40.75
Published on Web 03/20/2008

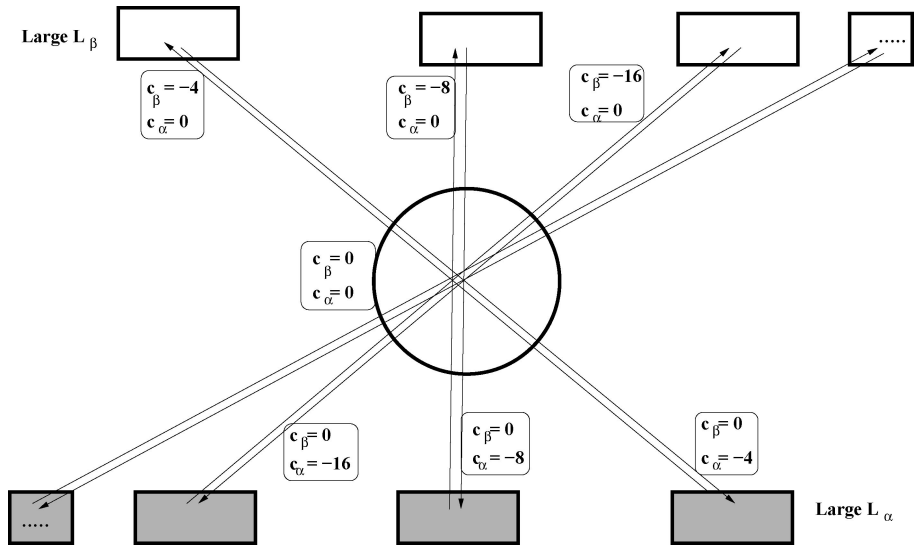

Figure 2. Schematic picture of the set of random walk simulations for a given molecular chain with different pairs of c_α/c_β values in eqs 11 and 12. Arrows connect two sets of configurations selected according to $c_\alpha < 0$ (lower regions) and $c_\beta < 0$. The inner circle represents the set of configurations obtained with $c_\alpha = c_\beta = 0$.

the free energy to construct a state with the value x for the configurational variable $X(\mathbf{r})$. This free energy, for chain a, is

$$f_a(x) = u_a(x) - Ts_a(x) \tag{1}$$

where u_a is the total energy in state x and s_a is related to the number of configurations with $X(\mathbf{r}) = x$, $n_a(x)$, characteristic of chain a

$$s_a(x) = R \log[n_a(x)] \tag{2}$$

with R the ideal gas constant. The average total energy is

$$u_a(x) = u_{0,a} + \langle K_a \rangle + \frac{\int_v U_a(\mathbf{r})\delta[X(\mathbf{r}) - x]\Omega_a(\mathbf{r})\, d\mathbf{r}}{n_a(x)} \tag{3}$$

with $u_{0,a}$ the formation energy of the chain a in a given reference conformation, U_a the potential energy change with respect to the reference conformation, and δ is the usual Dirac δ function. The function Ω_a is the density of states for chain a in configuration $\mathbf{r}$, and therefore the degeneracy $n_a(x)$ is

$$n_a(x) = \int_v \delta[X(\mathbf{r}) - x]\Omega_a(\mathbf{r})\, d\mathbf{r} \tag{4}$$

The entropy thus becomes

$$s_a(x) = R \log[n_a(x)] = R \log\{\int_v \Omega_a(\mathbf{r})\delta[X(\mathbf{r}) - x]\, d\mathbf{r}\} \tag{5}$$

The function $\Omega_a(x)$ is defined within a multiplicative constant, which is always canceled in configurational averages like that in eq 3. As for entropy, the constant defines the entropy of a reference state. The kinetic part $\langle K_a \rangle$ depends on moments only: $\langle K_a \rangle = d_a RT/2$, with d_a the number of degrees of freedom of chain a.

The evaluation of the energy contribution to f_a (eq 3) can be performed once we assume a given force field $U_a(\mathbf{r})$ for the interactions between the N_a atoms and some model for the interactions between the molecule and a mean-field representation of the environment (for instance, the water solvent, a membrane, or an aggregate). The entropic contribution and the presence of the denominator n_a in eq 3 deserve more calculations that are summarized in the following.

In general, one has to compute integrals of the type

$$I(x) = \int_v F(\mathbf{r})\Omega_a(\mathbf{r})\delta[X(\mathbf{r}) - x]\, d\mathbf{r} \tag{6}$$

with $F = U_a$ in eq 3 and $F = 1$ in eq 4. For estimating $\Omega_a(\mathbf{r})$ we resort to a Monte Carlo (MC) method. A computational method for generating a steady-state flow through relevant molecular configurations has been recently proposed.[5] A random walk for a molecular chain is generated by randomly changing dihedral angles in the molecular chain. For each suggested torsional move, a Metropolis test with a temperature randomly chosen in a given range is performed. The probability of accepting a move is, in the Metropolis scheme

$$P_A(r_1 \rightarrow r_2) = \min\{1, \exp[-\beta(U(r_2) - U(r_1))]\} \tag{7}$$

with $\beta = 1/RT$ the inverse temperature when the movement occurs and U the potential energy of the configuration $\mathbf{r}$. If T is constant, the set of configurational moves selected on the basis of the conditional probability encoded in the standard Metropolis MC scheme determines a steady-state trajectory which samples the configurations with the canonical configurational probability

$$P(\mathbf{r}) = 1/Z \exp[-\beta U(\mathbf{r})] \tag{8}$$

with

$$Z = \int_v \exp[-\beta U(\mathbf{r})]\, d\mathbf{r} \tag{9}$$

and v the space spanned by configurations $\mathbf{r}$. Alternatively, if β is extracted from a constant distribution, the conditional probability still determines a steady-state trajectory, but sampling an unknown configurational probability. If $T = 1/R\beta$ is uniformly and randomly extracted within all the possible values, i.e. in the range $(0, \infty)$, the configurational probability tends to the density of states. The enumeration of all possible states would include also states with very high energy that do not contribute to average properties at room temperature. In practice, the very high energy configurations are not included in the statistics if the temperature range is limited to a high number, like, for instance, 10 000 K. With this choice only "reasonable" configurations are explored, and if the trajectory is very long, all the states $\mathbf{r}$ with U in a limited range are enumerated.

The statistics associated with this random walk is called metastatistics and the associated configurational probability is

indicated with $\tilde{P}(\mathbf{r})$. We assume that for a wide range of reasonable configurations

$$\tilde{P}(\mathbf{r}) \sim \Omega(\mathbf{r}) \tag{10}$$

with Ω the density of states in eq 4.

Such metastatistics do not sample adequately configurations with low potential energy because of the intrinsic low number of configurations with such property.[6] To explore these configurations, a biasing external potential, V, is added to the intramolecular potential. This potential can be constructed with some useful properties, depending on the problem under study. Here it is chosen as a function of a collective variable describing molecular states very different from each other, which are known to be critically influenced by the molecular environment. In order to study the transition from configurations rich in helix content toward a state rich in elongated β-strands, two variables are defined. The maximum length of an helical segment, L_α, is defined as the maximum number of consecutive residues with $260° < \phi < 320°$ and $297° < \psi < 353°$. Similarly, the maximum length of β-strand, L_β, is defined as the maximum number of consecutive residues with $150° < \phi, \psi < 210°$.

The choice of Φ and Ψ dihedral angles as the basic variables to define the extent of helical and elongated motifs in the chains is an approximation. Typical α-helix and 3_{10}-helix domains display hydrogen bonds involving atoms O(i)...H($i + 4$) – N($i + 4$) and O(i)...H($i + 3$) – N($i + 3$), respectively. These requirements are not necessarily fulfilled when residues i and $i + 4$ (or i and $i + 3$) display Φ and Ψ dihedral angles in the above range. Nevertheless, since the configurations describing the state of each residue are obtained with an atomistic potential, the information about the required hydrogen bonds is partially included in the model, and indeed, most of the configurations extracted with Φ and Ψ in the defined range display also the required backbone hydrogen bonds. The correlation between the chosen dihedral angles and other geometrical properties required by the given secondary motifs will be reported and discussed in the first part of the Results section.

Each variable L_α and L_β is coupled with the external potential by a linear relationship:

$$\begin{aligned} RTV_\alpha &= RTc_\alpha L_\alpha \\ RTV_\beta &= RTc_\beta L_\beta \end{aligned} \tag{11}$$

The potentials V_α and V_β are dimensionless, with $c_{\alpha,\beta} \leq 0$. The two variables L_α and L_β are antagonist: the increase in L_α that occurs when V_α is acting on the molecule decreases the value of L_β and vice versa. When V_α is switched off and V_β is switched on, long α-helical segments are progressively destroyed and long β-strands are constructed. Different choices of c_α/c_β pairs change the pathway of this transition. The scheme in Figure 1 gives a physical description of the process: when V_α acts on the molecule for a given fictitious time T (given in terms of accepted configurational moves), the molecule is in a membrane-like environment; when V_β acts on the molecule for the same fictitious time T, the molecule is in an aggregate-like environment, where elongated configurations are preferred because they efficiently interact in extended intermolecular β-strands. When an external potential L_γ ($\gamma = \alpha, \beta$) is applied, the acceptance probability in eq 7 is changed to

$$P_A(r_1 \rightarrow r_2) = \min\{1, \exp[-\beta(U(r_2) - U(r_1)) - (V_\gamma(r_2) - V_\gamma(r_1))]\} \tag{12}$$

Separated random walks performed with different pairs of c_α/c_β values, from regions with large L_α (L_β) to regions with large L_β (L_α), represent separated $\alpha \rightarrow \beta$ (and vice versa) processes, all going through configurations with L_α, $L_\beta \sim 0$: this intersection area where all the trajectories overlap is due to the antagonism between the two variables. The scheme in Figure 2 is a representation of the ensemble of these trajectories: different trajectories share the same set of transient configurations, the ensemble of which is represented as a circle in Figure 2. This ensemble is the set of configurations with $c_\alpha = c_\beta = 0$. In theory, the metastatistics obtained with $c_\alpha = c_\beta = 0$ is the desired $\tilde{P}(\mathbf{r}) \sim \Omega(\mathbf{r})$. But, as observed above, in the absence of an external potential important configurations are not sampled in limited real times. When the external potential V_γ is applied, the configurational probability $\tilde{P}_\gamma$ is related to the density of states by

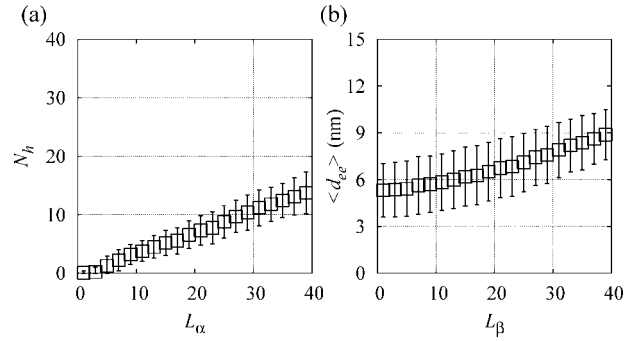

Figure 3. Average number of residues in helical domains (N_h) for each L_α value (panel a) and average distance ($\langle d_{ee} \rangle$) for each L_β value (panel b). Error bars were obtained as root-mean-square errors on collected values.

$$\tilde{P}_\gamma(\mathbf{r}) = 1/Z_\gamma \tilde{P}(\mathbf{r}) \exp[-V_\gamma(\mathbf{r})] \tag{13}$$

where V_γ is either V_α or V_β. Z_γ is the normalization factor

$$Z_\gamma = \int_v \tilde{P}_\gamma(\mathbf{r})\, d\mathbf{r} \tag{14}$$

Inverting eq 13 for $\tilde{P}$ and using the assumption of eq 10, one obtains

$$\Omega_a(\mathbf{r}) \sim \tilde{P}_a(\mathbf{r}) = Z_{\gamma,a} \tilde{P}_{\gamma,a}(\mathbf{r}) \exp[V_\gamma(\mathbf{r})] \tag{15}$$

where the label a related to the chain has been introduced. If simulations with a given external potentials V_γ are performed for the same number of collective moves, M_c, for different chains, the values of $Z_{\gamma,a}$ are the same for every chain α, being $Z_{\gamma,a} = Z_{\gamma,b} = M_c$. The above equation becomes

$$\Omega_a(\mathbf{r}) \sim \tilde{P}_a(\mathbf{r}) = M_c \tilde{P}_{\gamma,a}(\mathbf{r}) \exp[V_\gamma(\mathbf{r})] \tag{16}$$

Equation 16 gives an estimate of the density of states for a given molecular chain in terms of quantities that can be calculated from a single simulation of metastatistics in the presence of an external potential V_γ. V_γ is, alternatively, V_α and V_β of eq 11, with different values of c_α and c_β, respectively.

The function $\tilde{P}_{\gamma,a}(\mathbf{r})$ is a biased estimate of the same density of states $\Omega_a(\mathbf{r})$ of eq 10, but in a region where L_α or L_β is large and where an unbiased $\tilde{P}_a$, constructed with no biasing external potential, does not contribute significantly. Therefore, in this form, i.e. with a unique value for V_γ, the metastatistics sample a limited region of the configurational space, that associated with the biased values of X imposed by the defined value of V_γ. In order to explore configurations with a large range of X (L_α or L_β) values, many V_γ (c_α or c_β values, respectively) must be used and combined. A question then arises: is there the possibility of estimating $\Omega_a(\mathbf{r})$ with contributions of many choices of V_γ? To explore this possibility, let us consider a further approximation for $\Omega_a(\mathbf{r})$, expanding eq 16 in the form

$$\Omega_a(\mathbf{r}) = Z \sum_\gamma \tilde{P}_{\gamma,a}(\mathbf{r}) \exp[V_\gamma(\mathbf{r})] \tag{17}$$

with Z the arbitrary constant that defines the reference for entropy of eq 5. This constant is omitted, hereafter, because it

does not affect the comparisons between $f_a(x)$ functions between different chains a.

Inserting this equation into eq 6 we have

$$I(x) = \int_v F(\mathbf{r})\{\sum_\gamma \exp[V_\gamma(\mathbf{r})]\widetilde{P}_{\gamma,a}(\mathbf{r})\}\delta[X(\mathbf{r}) - x]\,\mathrm{d}\mathbf{r} \quad (18)$$

We make the following observations.

1. The function

$$\widetilde{P}'_a(\mathbf{r}) = 1/N_\gamma \sum_\gamma \widetilde{P}_{\gamma,a}(\mathbf{r}) \quad (19)$$

with N_γ the number of used biasing potential, is the averaged configurational probability of the biased metastatistics. This function is the biased form of $\tilde{P}_a$ that appears in eq 16, which is unbiased. The distribution associated with these metastatistics are smooth functions of the collective variable X used for the bias when the density of c_γ values is constant. For instance, the biased degeneracy of x

$$\tilde{n}'_a(x) = \int_v \widetilde{P}'_a(\mathbf{r})\delta[X(\mathbf{r}) - x]\,\mathrm{d}\mathbf{r} \quad (20)$$

is a smooth function of x, as we shall see in the examples shown in the Results section.

2. Because of the δ function in the integrand function of eq 18, values of $I(x)$ are contributed by the sum

$$\sum_\gamma \exp[V_\gamma(\mathbf{r}')]\widetilde{P}_{\gamma,a}(\mathbf{r}') \quad (21)$$

for those configurations $\mathbf{r}'$ having $X(\mathbf{r}') = x$. This sum is related to an average over the values for the external potential, which can be computed on the basis of the metastatistics. In fact, the identity

$$\langle \exp(V)\rangle'_a(x) = \frac{\sum_\gamma \exp[V_\gamma(\mathbf{r}')]\widetilde{P}_{\gamma,a}(\mathbf{r}')}{\sum_\gamma \widetilde{P}_{\gamma,a}(\mathbf{r}')} \quad (22)$$

is the average, over the choices of biasing external potential, of (V_γ) computed using the subset of $X(\mathbf{r}')$ values obtained via the metastatistics. Inserting eq 22 in eq 18, we have

$$\begin{aligned} I(x) &= \int_v F(\mathbf{r})\langle \exp((V)\rangle'_a(x)\left\{\sum_\gamma \widetilde{P}_{\gamma,a}(\mathbf{r})\right\}\delta[X(\mathbf{r}) - x]\,\mathrm{d}\mathbf{r} \\ &= \langle \exp((V)\rangle'_a(x)\int_v F(\mathbf{r})\left\{\sum_\gamma \widetilde{P}_{\gamma,a}(\mathbf{r})\right\}\delta[X(\mathbf{r}) - x]\,\mathrm{d}\mathbf{r} \\ &= N_\gamma\langle \exp((V)\rangle'_a(x)\int_v F(\mathbf{r})\widetilde{P}'_a(\mathbf{r})\delta[X(\mathbf{r}) - x]\,\mathrm{d}\mathbf{r} \end{aligned} \quad (23)$$

where eq 19 has been used in the last step. The number N_γ is a constant that is absorbed in the entropy of a reference state. We notice that the averages $\langle \exp(V)\rangle'$ are averages of small numbers (because V is negative by definition, eq 11, and it is large in absolute value in practice), but the weight of these small numbers is approximately exponentially enhanced by the biased metastatistics.

If $F = 1$ in eq 23, we recognize in the integral the biased degeneracy of the quantity x (eq 20) for chain a obtained from the biased metastatistics. Therefore, the entropy of eq 5 becomes

$$s_a(x) = R\log\{\tilde{n}_a(x)\} + R\log\{\langle \exp(V)\rangle'_a(x)\} \quad (24)$$

As for the energetic contribution to f_a, by using eq 23 (with $F = U_a$ in the numerator and $F = 1$ in the denominator) and eq 19 in the last term on the right-hand side in eq 3, this latter becomes

$$u_a(x) = u_{0,a} + d_aRT/2 + \frac{\int_v U_a(\mathbf{r})\widetilde{P}'_a(\mathbf{r})\delta[X(\mathbf{r}) - x]\,\mathrm{d}\mathbf{r}}{\tilde{n}'_a(x)} \quad (25)$$

The term $\langle \exp(V)\rangle'_a(x)$, both in numerator and denominator, is canceled.

The entropy $s_a(x)$ is expected to be a function determined by the sequence of the molecular chain, rather than by the details of the interatomic potential, when x is a collective variable like L_α or L_β. The entropy as a function of x, therefore, is almost independent from the value of U. Therefore, eq 25 can be improved by refining the potential function: where

$$\begin{aligned} u_a(x) &= u_{0,a} + d_aRT/2 + u'_a(x) \\ &= u_{0,a} + d_aRT/2 + \frac{\int_v U'_a(\mathbf{r})\widetilde{P}'_a(\mathbf{r})\delta[X(\mathbf{r}) - x]\,\mathrm{d}\mathbf{r}}{\tilde{n}'_a(x)} \end{aligned} \quad (26)$$

$$U'_a(\mathbf{r}) = U_a(\mathbf{r}) + U_{\mathrm{corr},a}(\mathbf{r}) \quad (27)$$

The correction $U_{\mathrm{corr},a}$ introduces the difference between a realistic potential for the molecular chain and the intramolecular potential U_a used for the metastatistics. This correction includes a more realistic distance cutoff for pair interactions ($U_{\mathrm{intra},a}$) and the mean-field interaction with the solvent or with the molecular environment characteristic of a molecular aggregate ($u_{\mathrm{solv},a}$):

$$U_{\mathrm{corr},a}(\mathbf{r}) = U_{\mathrm{intra},a}(\mathbf{r}) + u_{\mathrm{solv},a}(\mathbf{r}) \quad (28)$$

In this study, the water solvent is taken into account by computing the non polar and polar contribution to the formation of the solute−water interface of each molecular configuration (see below).

Summarizing, the free energy of chain a for a given L_α (or L_β) is

$$f_a(L_\alpha) = u_{0,a} + d_aRT/2 + u'_a(L_\alpha) - RT\log\{\tilde{n}'_a(L_\alpha)\} - RT\log\{\langle \exp(V)\rangle'_a(L_\alpha)\} \quad (29)$$

By inserting L_β in place of L_α the free energy difference for increasing the β-strand content is obtained. All the quantities in eq 29 are calculated directly from the biased metastatistics via eqs 26, 22, and 20. The details of the biased metastatistics and of the energy correction are reported in the following section.

A similar estimate of $s_a(x)$ can be obtained using the generalized ensemble method.[7–9] This latter method is based on a factorization of the density of states $n(x) = \exp[s(x)]$ as in eq 24. Indeed, tests on ideal systems of the two methods give very similar results.

3. Computational Details

3.1. Models of Chains. Atomistic models for the polypeptide chains have been constructed by using the Amber PARM94 force field,[10] with the modifications proposed to decrease the α-helical propensity of the force field.[11] These modifications impose a zero torsional potential for the Φ and Ψ dihedral angles in the protein backbone. All the chains have been terminated with neutral groups (acetyl at the N-terminus and N-methylamide at the C-terminus). The protonation state of residues is that usually assumed for amino acids at pH = 7; i.e., Glu and Asp residues are deprotonated, Lys and Arg residues are protonated, and His is in the Nϵ2 protonated (neutral) form.

In all cases, the initial configurations has been generated assigning to the Φ and Ψ dihedral angles in the backbone the values of 180°; i.e., all the chains started from an all-trans conformation in the backbone. The only exception is for Pro residue in the villin headpiece HP(1–36), where the Φ dihedral angle was fixed at –60° and never changed in any random walk.

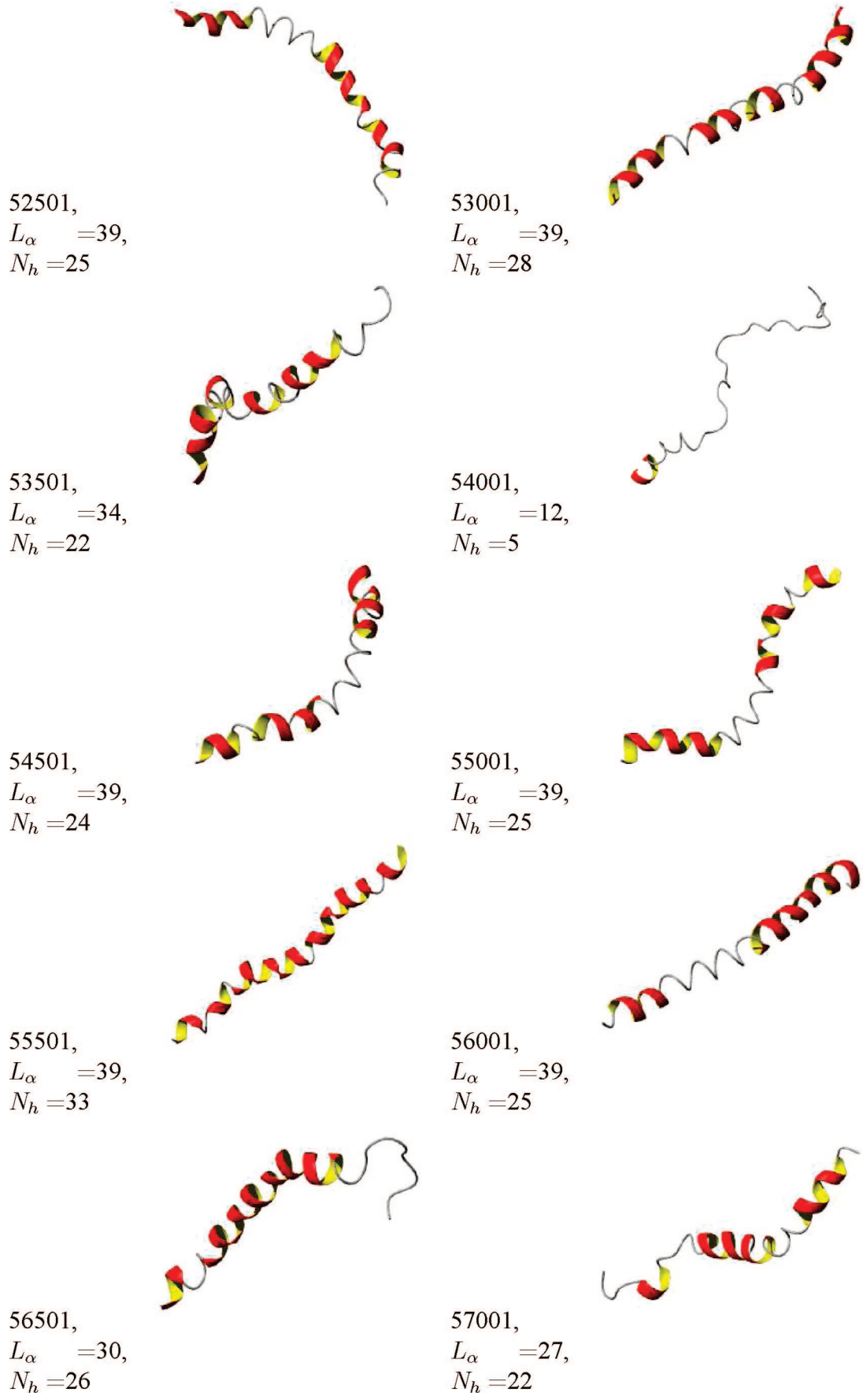

Figure 4. Ten structures obtained with $c_\alpha = -28$. Structures are displayed as ribbons where helical residues are identified, by using the program MolMol.[19] Numbers indicate the snapshot within the 60 000 configurations used for the statistics with $L_\alpha > 0$, the value of L_α, and the value of N_h (see text for details).

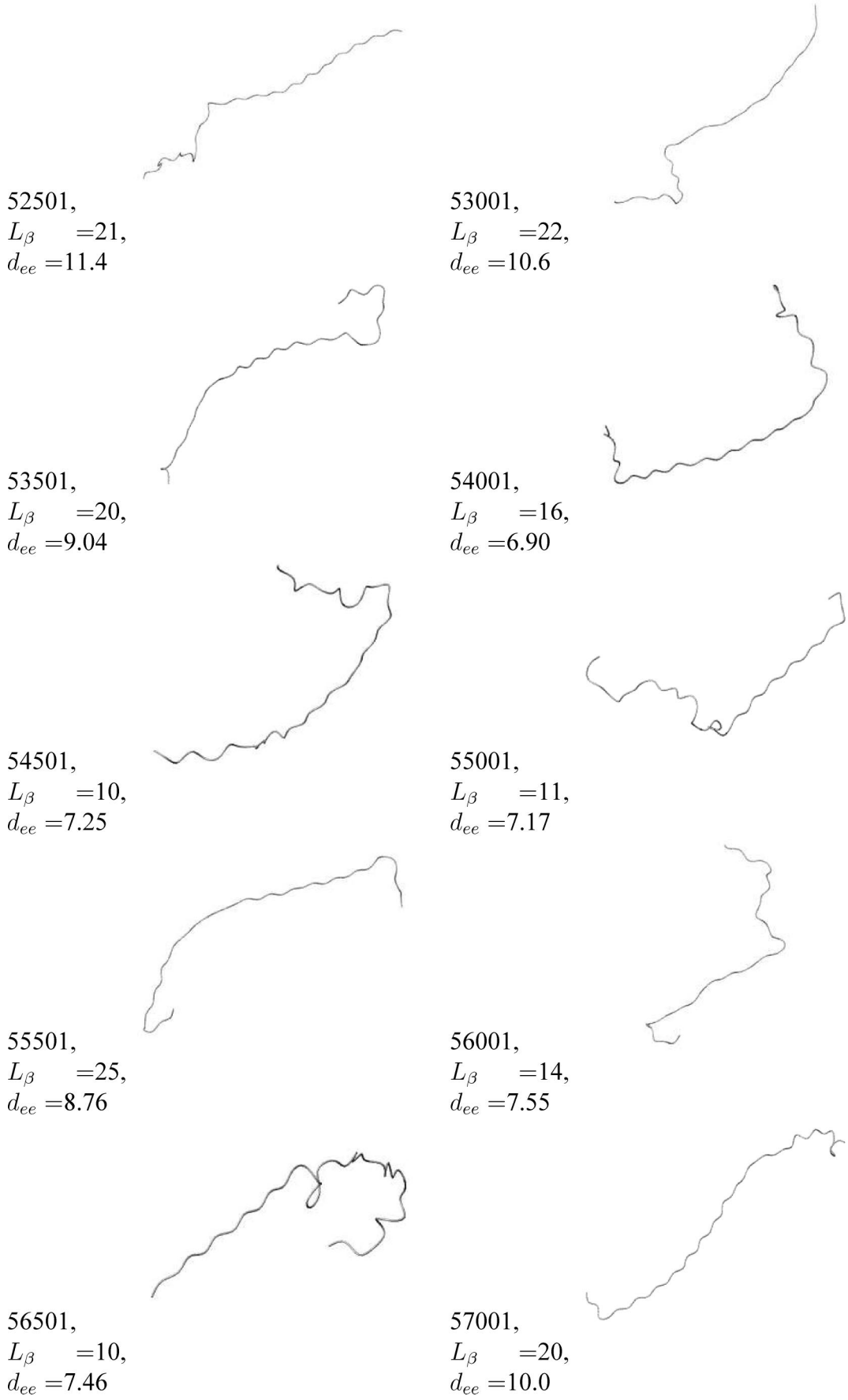

Figure 5. Ten structures obtained with $c_\beta = -28$. Structures are displayed by using the program MolMol.[19] Numbers indicate the snapshot within the 60 000 configurations used for the statistics with $L_\beta > 0$, the value of L_β, and the value of d_{ee} in nm (see text for details).

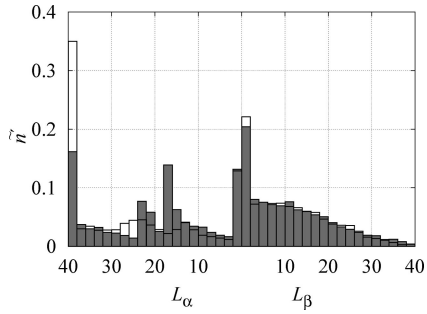

Figure 6. Probability of L_α (x-axis to the left) and L_β (x-axis to the right) obtained with biased random walks for Thr (filled bars) and for Val (empty bars).

The homopolymeric sequences studied in this work are Ace-Gly_{40}-Nme, Ace-Ala_{40}-Nme, Ace-Val_{40}-Nme, Ace-Thr_{40}-Nme, Ace-Lys_{40}-Nme, and Ace-Glu_{40}-Nme. The homopolymer chains with Lys and Glu were also modeled by neutralizing the side chains, by adding a charge $+^1/_3|e|$ to the Cϵ and Oϵ atoms of Glu and $-^1/_4$ $|e|$ to the Nζ and Hζ atoms of Lys (indicated as Glun and Lysn, respectively).

The Aβ(1–40) peptide and three sequences of 40 amino acids derived by randomly scrambling the Aβ(1–40) sequence were modeled. The Aβ(1–40) sequence followed by the three scrambled sequences (sequences 1, 2, and 3, respectively) are Ace-DAEFRHDSGY_{10}EVHHQKLVFF$_{20}$AEDVGSNKGA$_{30}$IIG-LMVGGVV$_{400}$-NMe, Ace-AKEDALVAFI$_{10}$NKGFLYVRGV$_{20}$ EGEGVFHVMG$_{30}$HGQHDIDSSV$_{40}$-NMe, Ace-VRHFFVGAIV$_{10}$ GDFGHLIVDE$_{20}$QGSAVVSEKL$_{30}$EGMKAGHDNY$_{40}$-NMe, and Ace-AKAVVSVIVR$_{10}$GYEGGHDAFG$_{20}$HMVLDKGIEE$_{30}$ HQFDLFNGVS$_{40}$-NMe. The villin headpiece HP(1–36) sequence is Ace-MLSDEDFKAV$_{10}$FGMTRSAFAN$_{20}$LPLWK-QQNLK$_{30}$KEKGLF-NMe.

3.2. Metastatistics. Configurational moves are proposed by randomly choosing one torsional angle among the set of all torsional angles involving non-hydrogen atoms. The chosen torsional angle is changed by rotating all the dihedral angles associated with the torsion of an angle randomly chosen in the range $[0, 2\pi]$. The new configuration is then used in the Metropolis test with a temperature T randomly chosen within 0 and 10 000 K. A collective move is assumed when the number of attempted torsional moves reaches 5 times the number of movable torsions in the chain. For each chain, 8 separated random walks have been acquired, each of 30 000 collective moves. Once each trajectory starts with a given V_β, after 1000 collective moves the V_β and V_α fields are switched off and on, respectively. The process is reversed after other 1000 collective moves, and the two steps are repeated 15 times. Every trajectory corresponds to a given pair of values for c_α and c_β in eq 11. The values of c_α and c_β are decreased in steps of –4 starting from a value of zero for the first trajectory and reaching the minimum value of –28 for the eighth trajectory. Summarizing, for each chain the biased random walk, as it is schematically represented in Figures 1 and 2, consists of 240 000 attempted collective torsional moves. The acceptance ratio was about 1/2. Among all of the acquired configurations, the second half of those obtained during the application of each external field of eq 16 were collected for further analysis, as they are expected approximately equilibrated in the respective external potential. A total amount of 60 000 configurations with $L_\alpha > 0$ ($c_\alpha < 0$) and 60 000 configurations with $L_\beta > 0$ ($c_\beta < 0$) were acquired for each chain.

3.3. Potential Refinement and Mean-Field Correction. For all the calculations of corrected $U_a'(\mathbf{r})$ values (eq 27), the cutoff was infinite; i.e., all the pairwise electrostatic interactions are included in the refined energy potential. This is particularly important for highly charged chains.

The Lennard-Jones interactions have been discarded from the calculation of the refined potential energy, in both U_a and $U_{\mathrm{corr},a}$ of eq 27, for the following reason. The loss of van der Waals interactions between solvent and solute atoms, when these latter are moved from the molecular surface inside the solute macromolecule, has to be included in a volume-dependent fashion in the mean field.[12] Such a correction to the surface-dependent mean field is not yet fully available. Nevertheless, configurations included in the calculation of free energy do not overlap because short-range Lennard-Jones interactions were included in the potential energy of the random walks.

A mean-field model for the solvent was used to measure the potential energy of the polypeptide in water. For each chain a, the contribution to $U_{\mathrm{corr},a}$of eq 27 is

$$u_{\mathrm{solv},a}(\mathbf{r}) = \Delta u_{\mathrm{nonpol}}(\mathbf{r}) + \Delta u_{\mathrm{pol}}(\mathbf{r}) \tag{30}$$

The first term on the right-hand side, $\Delta u_{\mathrm{nopol}}$, is the contribution to the solvation free energy due to the formation of a cavity of zero charge density with the shape of the solute and to the creation of the solute−solvent interface. The introduction of a charge density in the space occupied by the solute gives the second contribution, Δu_{pol}. The charge density is given in terms of the point charges q_i of atom i in the Amber force field, where i runs over the N_a atoms in the molecule. These point charges are located in the positions $\mathbf{r}_i$ of the given molecular configuration $\mathbf{r}$.

The term $\Delta u_{\mathrm{nopol}}$ is calculated as a linear combination of the solvent accessible surface areas (SASAs) of each group in the solute molecule:[13]

$$\Delta u_{\mathrm{nonpol}} = \sum_i^{N_a} \sigma_i \mathrm{SASA}_i \tag{31}$$

with the coefficients σ_i positive and negative for hydrophobic and hydrophilic groups, respectively.

The electrostatic contribution to the solvation free energy, Δu_{pol}, is the electrostatic energy for charging the solute molecule with an arbitrary shape, and it is obtained by numerical finite difference solution of the Poisson−Boltzmann equation (PBE).[14] A grid of about 100 points per side was used, and the size of the finite element was 0.0875 nm. Each configuration of the polypeptide molecule was placed in the center of an orthorhombic cell providing at least 1 nm of boundary dielectric in every direction of space. The Debye boundary conditions were adopted; i.e., the electrostatic potential decreased exponentially at the cell boundaries. The solvation energy is obtained in terms of the interactions between the atomic point charges and the polarization surface charge density placed on the solute−water interface.[15,16]

The evaluation of SASA of each atom was performed using the "Numerical Surface Calculation" (NSC) code[17] with a density of 122 points per unit sphere. The SASA of each group was calculated using the van der Waals radii of the same Amber PARM94 force field used in MD simulations. The probe radius was 0.14 nm, and it was kept fixed for all the following calculations. The σ_i coefficients of eq 31 were taken from the literature,[13] and for Δu_{pol}, the relative dielectric permittivities of the molecule and of the solvent were 1 and 80, respectively. In the assumed condition (bulk 1:1 physiological salt concentration $C_s = 0.1$ M) the linear approximation of the PBE can be used.

4. Results

The correlation between Φ/Ψ state for each residue, as defined in the Methods section, and backbone hydrogen bonds, typical of helical domains, or molecular elongation, typical of

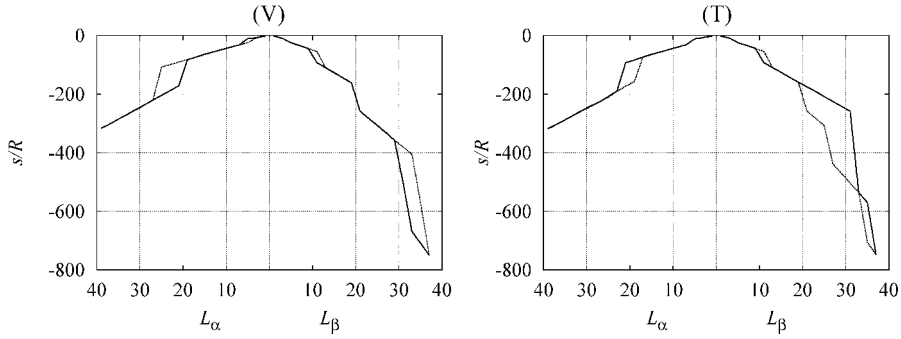

Figure 7. Entropy of Val_{40} (panel V) and Thr_{40} (panel T) as a function of L_α (*x*-axis to the left) and L_β (*x*-axis to the right) in units of *R*, computed with random walks using a 0.5 nm cutoff (solid line) and 1 nm cutoff (dotted line) for nonbonding interactions.

β-strands, was measured in the collected statistics. When $L_\alpha > 0$, the average number of helical residues (N_h, both α and 3_{10} helices) as they are determined with the DSSP algorithm[18] implemented in the MolMol program,[19] was measured for each represented value of L_α . When $L_\beta > 0$, the average end-to-end distance ($\langle d_{ee} \rangle$, here defined as the distance between N of the first aminoacid and C of the last amino acid) was measured for each value of L_β. In Figure 3 these two quantities are plotted in panels a and b, respectively, for Val_{40}. The variable L_α is a suitable tool for changing N_h and the two variables appear linearly correlated, even if the correlation between the two variables is low. For instance, when L_α is at maximum (40), only about one-third of the residues (13) are, on average, in helical domains. Despite the fact that only part of the efforts spent in increasing L_α translate into an effective increase in N_h, the use of L_α is more manageable than N_h for biasing the statistics. Moreover, all of the configurations with $L_\alpha > 8$ display at least one helical domain.

The same analysis holds for the relationship between L_β and the average end-to-end distance: only part of the increase in L_β produces an increase in the average elongation, but the correlation appears, also in this case, almost linear.

The efficiency of L_α and L_β in changing the average properties of the chains can be further investigated by looking at some of the configurations collected when V_α, or alternatively V_β, are negative and maximum in absolute value. In Figures 4 and 5, 10 configurations obtained with $c_\alpha = -28$ and with $c_\beta = -28$, respectively, are displayed. The secondary motifs are represented according to the same DSSP algorithm used to measure N_h. With few exceptions, the large content of helical (Figure 4) and elongated (Figure 5) configurations is displayed with helical and stretched domains, respectively, located in positions statistically distributed along the molecular chain. As can be observed, for instance, in configuration 56001 of Figure 4, the breaking of DSSP identified domains is due to a segment that closely resembles a helical domain, thus showing that the whole chain represents a nascent long helix.

From the above analysis we conclude that our choice of *x* as L_α or L_β allows to change the molecular statistics from a draft of long helical segments (large L_α, represented by configurations like those displayed in Figure 4) toward elongated chains (large L_β, represented by configurations like those in Figure 5).

In Figure 6, the density of *x*, $\tilde{n}'(x)$, i.e., the biased metastatistics probability of finding configurations with given $x = L_\alpha$ (left *x*-axis) or $x = L_\beta$ (right *x*-axis), is plotted for Thr_{40} (filled boxes) and Val_{40} (empty boxes). The density of segments with length $L_\alpha \sim 20$ is emphasized for Thr, while a complete helical draft for Val is more likely than for Thr. The probability of L_β is almost the same for the two oligomers. Except for the most extended helical draft, Val_{40} displays a low probability for any helical or stretched conformation, thus suggesting that this oligomer can indifferently sample the conformational space.

In order to show that the entropy of eq 24, $s_a(x)$, does not depend significantly from the potential used in the metastatistics, we performed the random walks for Val_{40} and Thr_{40} with a short cutoff (0.5 nm) and a large cutoff (1 nm) for the nonbonding interactions. Then the two resulting entropy functions are compared in Figure 7 for Val_{40} (panel V) and Thr_{40} (panel T). The entropy of these chains is almost identical in the two different random walks in the range of $L_\alpha, L_\beta < 20$. As expected, entropy decreases increasing both L_α and L_β compared to unstructured configurations ($L_\alpha \sim L_\beta \sim 0$), i.e., no secondary motifs are present. The decrease in entropy by increasing L_α is smaller than by increasing L_β. The leading contribution to *s* is the second term in the right-hand side of eq 24: when L_α (or, respectively, L_β) is large, this contribution almost cancels the bias induced by the external potential in n'.

The $s(x)$ profiles for all the homopolypeptides are very similar to those reported for Val and Thr, with the exception of Gly. Gly displays a completely symmetric profile: the *s* function decreases equally while increasing L_α and L_β (data not shown).

In Figures 8, 9, and 10 the free energy of eq 29 is plotted as a function of L_α (left *x*-axis) and L_β (right *x*-axis). The temperature $T = 300$ K has been chosen for convenience. The values obtained in water solution and with no solvent (i.e., Δu_{pol} and Δu_{npol} are zero) are reported on the left and right *y*-axis, respectively. The all-trans conformation (maximum L_β) has been chosen as the zero-energy conformation $u_{0,a}$. Therefore, the zero free energy displayed in the figures (when L_β is maximum) is simply due to the number of torsional degrees of freedom in the chain (d_a in eq 29) and to $s_a(L_{\beta,max})$.

For Gly_{40} (panel G) the behavior of *f* in water is very different from that in the vacuum: the chains with no helical and stretched domains display the minimal *f* in water, while in the vacuum *f* is minimal for maximal L_α. Experimental results on Gly are controversial because of the low solubility in water of poly-Gly sequences that readily form aggregates.[3,20] Since in the vacuum the increase in *f* by increasing L_β is moderate, a small stabilization due to intermolecular hydrogen bonds can compensate the *f* increase for the single chain, thus supporting the experimentally observed aggregation propensity of polyglycine sequences. Ala_{40} displays propensity for helices (panel A), mainly because of the large energetic contribution u' to *f*. The negative energetic contribution largely compensates the entropic contribution to $f\,(-Ts)$ that for Ala_{40} displays a profile like that plotted in Figure 7 for Val and Thr. Val_{40} and Thr_{40} chains display an *f* profile that is dominated by entropy (see Figure 7 for comparison). Noticeably, for Ala, Val, and Thr chains there is no significant effect of the solvent. Thr displays a slightly smaller propensity of helical segments in water compared to Ala and Val chains.

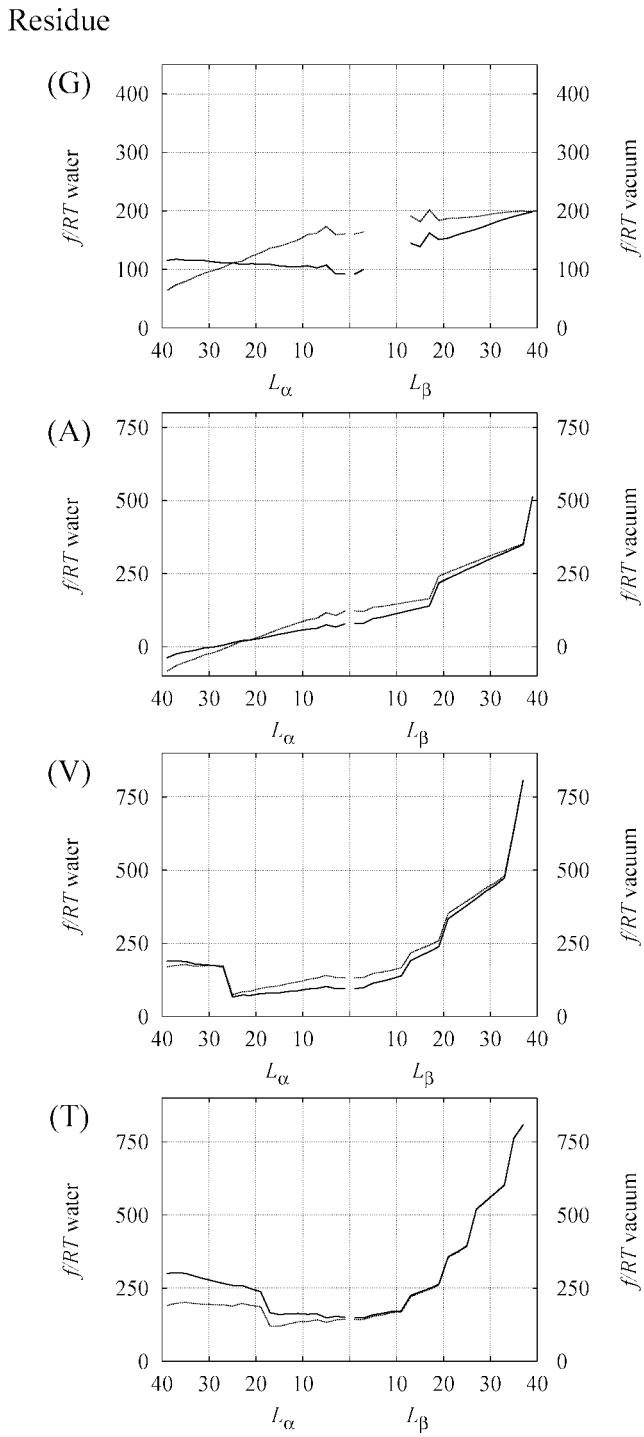

Figure 8. Free energy (eq 29) as a function of L_α (*x*-axis to the left) and L_β (*x*-axis to the right) in units of RT ($T = 300$ K). All the plots are shifted to have $u'(L_{\beta,\max}) = 0$, with $L_{\beta,\max} = 39$. Gly_{40} (panel G), Ala_{40} (panel A), Val_{40} (panel V), Thr_{40} (panel T). Solid line is in the water solvent (left *y*-axis); dotted line is in the vacuum (right *y*-axis).

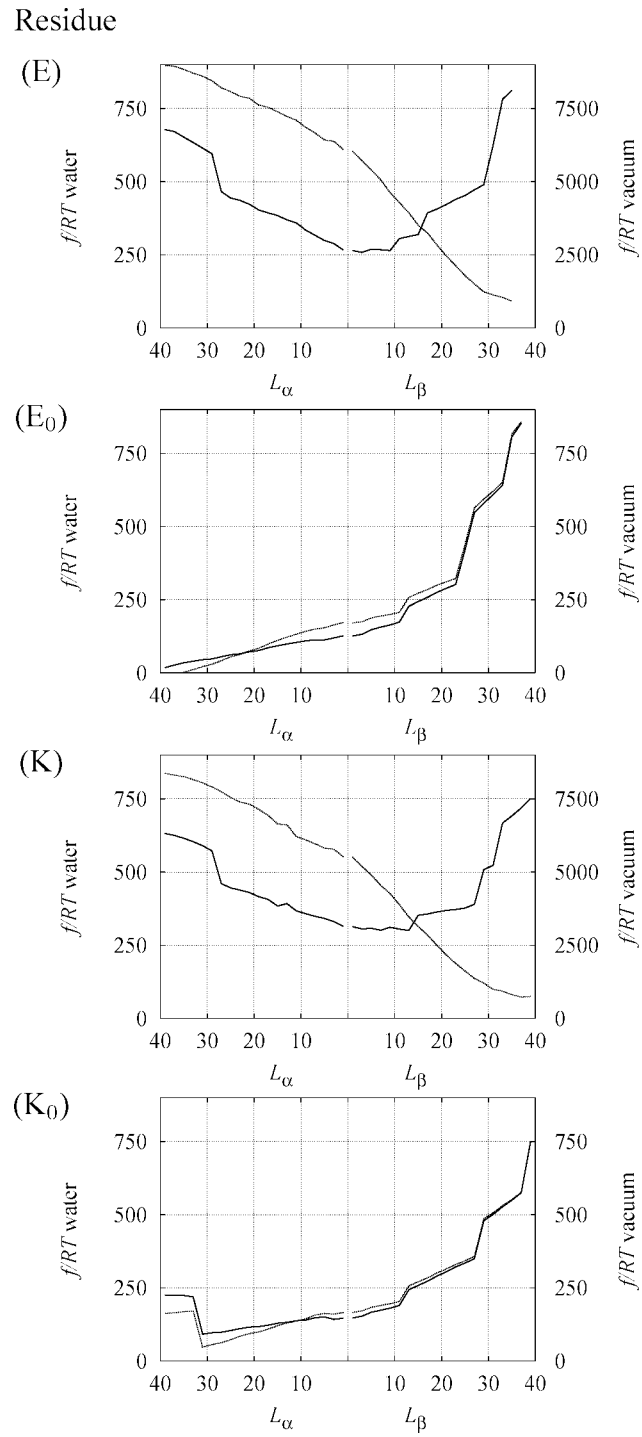

Figure 9. Same as for Figure 8 with Glu_{40} (panel E), Lys_{40} (panel K), $Glun_{40}$ (panel E_0) (Glun is charge-neutralized Glu), $Lysn_{40}$ (panel K_0) (Lysn is charge-neutralized Lys). Solid line is in the water solvent (left *y*-axis), dotted line is in the vacuum (right *y*-axis).

In Figure 9, the free energy is plotted for the highly charged homopolypeptides Glu_{40} (panel E) and Lys_{40} (panel K). When there are long helical segments the net charges on the side chains are statistically closer in space and the free energy becomes large in the vacuum. (The expansion of the right *y*-axis is 10 times larger than that of the left *y*-axis when the chains are charged.) Therefore, in the vacuum the energetic contributions largely dominate over the entropy variation. For charged chains, the eventual aggregation, favored by the propensity for elongation, is, however, contrasted by the strong electrostatic repulsion between different chains. If the side chains are charge-neutralized (as in the cases of neutralized Glu, panel E_0, and neutralized Lys, panel K_0), the behavior is inverted and the long helical segments are stabilized. For instance, the behavior of $Glun_{40}$ (panel E0), becomes very similar to Ala_{40} (Figure 8A). The Glu and Lys residues, even if neutralized, are more soluble in water than Ala and therefore Glu_{40} at low pH and Lys_{40} at high pH are expected to be structured in long stable helical segments when in water solution. The nonhelix to helix transition for polyglutammate and polylysine has been studied experimentally:[21] poly-Glu at pH < 4.5 and poly-Lys at pH > 10.5 are structured as α-helices. The charged model we used

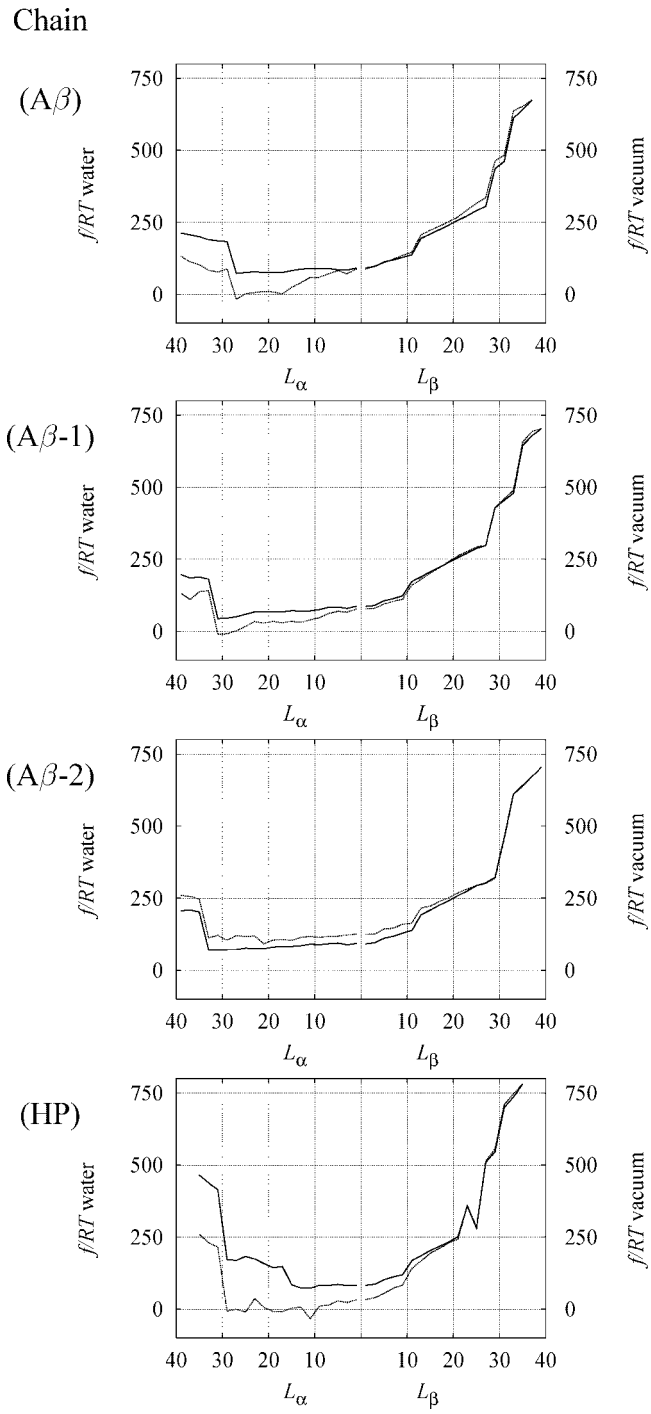

Figure 10. Same as for Figures 8 and 9 with Aβ(1–40) (panel Aβ), random Aβ(1–40), sequence 1 (panel Aβ –1), random Aβ(1–40), sequence 2 (panel Aβ–2), and villin headpiece (1–36) (panel HP). Solid line is in the water solvent (left *y*-axis); dotted line is in the vacuum (right *y*-axis).

indicates that as far as the configurations deviate from the entirely stretched all-trans structure, the repulsion between charges increases dramatically the potential energy and the free energy compared to Ala_{40} and Val_{40}. The free energy profile estimated, for instance, for Lys_{40} is expected to change, by increasing pH, from that of Figure 9K to that of Figure 9K_0, crossing a flat *f* profile for some intermediate pH. In this latter case the chain becomes structure indifferent, that is, the translation of the polymeric random coil in the frame of biopolymers. Since it has been observed that at pH 10.3 polylysine is still 35% charged,[22] a flat *f* profile in water is

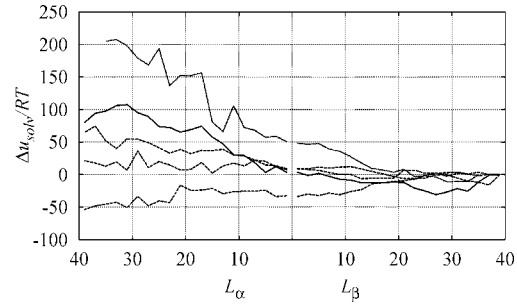

Figure 11. Solvation contribution to u' (Δu_{solv}) (the difference between the solid and dotted curves in Figure 10) plotted for Aβ(1–40) (solid line), random Aβ(1–40) (dashed lines), and villin headpiece (1–36) (dotted line).

expected to occur for models with a charge larger than our entirely neutralized K_0 model.

In Figure 10, the free energy for the Aβ(1–40) peptide (panel Aβ), for two random sequences (1 and 2) with the same composition of the Aβ(1–40) peptide (panels Aβ–1 and Aβ–2), and for the villin headpiece 1–36 (HP(1–36), panel HP) is displayed. In all cases and in both conditions (in water solvent and in the vacuum, with the exception of HP in water), relatively long helical drafts are more stable than stretched configurations. The water solvent is more effective in changing the *f* profile in the vacuum for HP(1–36) and Aβ(1–40). However, the peptides are almost indistinguishable in the scale of these plots, especially when the peptides are in water.

For the Aβ(1–40) peptide, for the three randomized Aβ sequences and for the villin headpiece, the total solvation free energy of eq 30 was also analyzed. Most of the solvation free energy is given by the electrostatic contribution (Δu_{pol}): the nonpolar contribution is about –150 *RT* units when L_α is at maximum and displays, by moving toward high L_β values and for all the peptides, a decrease contained within 25 *RT* units; the polar contribution varies around a value of –650 *RT* with oscillations of about 200 *RT*. The sum of the two contributions is, therefore, dominated by the polar contribution. In Figure 11, the solvation energy (i.e., the difference between solid and dotted *f* plotted in Figure 10 for each peptide) is plotted for Aβ(1–40) (solid line), HP(1–36) (dotted line), and the three randomized Aβ(1–40) sequences (dashed lines). In all cases, the increase in Δu_{solv} for increasing L_α almost compensates the decrease of *f* in the vacuum when the two terms are added together for computing the *f* in water. This large compensation makes the decrease of *f* in water much smaller than in the vacuum. Despite the large extent of cancelation, in water a significant increase of *f* by increasing L_β can be observed, and remarkably, this increase is the same for all the five chains here analyzed. This increase is largely due to the entropic contribution to *f* that is not sufficiently compensated by the inverse change of Δu_{solv}. This result shows that for none of these model chains there is large propensity to extend the β-strand content, when the peptides are both in the vacuum and in the water solvent. For a comparison, Gly_{40} in the vacuum clearly displays an intrinsic propensity to extend L_β (Figure 8G, dotted line).

5. Conclusion

An atomistic model has been used to generate reasonable configurations for polypeptide chains of sequence X_{40} with X = G, A, V, T, K, E, and with the sequence of Aβ(1–40), three Aβ(1–40) randomly scrambled sequences, and the villin headpiece HP(1–36).

The set of reasonable configurations has been used to compute the entropy of a chain in a macrostate with given maximum length *x* of either helical or elongated segments ($x = L_\alpha$ or $x =$

L_β, respectively). The total energy for the chains has been computed by using a realistic potential function for the interatomic interactions, including solvation, and assuming that the entropy is independent from the details of the potential energy function. The total energy of a chain in the same macrostate x has been computed and indicated as $u'(x)$. A free energy estimate has been given as $f(x) = u'(x) - Ts(x)$, i.e. the free energy change for the chain in the macrostate x with respect to an arbitrary reference state, which we choose as $x = L_{\beta,\max}$.

The method resembles generalized ensemble statistics but has the advantage of relying on any set of arbitrary configurational variables x in place of the potential energy U. This variable can be chosen as the most convenient for analyzing free energy changes of chain molecules, like polypeptides, in different environments. Moreover, the use of random temperature allows a wide exploration of configurational landscape.

By comparing free energy for the studied polypeptides, it has been observed that the $A_\beta(1–40)$ peptide sequence behaves like Val_{40} in water and displays a slight preference for helical segments in the vacuum, this latter environment assumed to mimic a hydrophobic environment.

Remarkably, Aβ(1–40), its scrambled sequences, and HP(1–36), at the same extent, display no propensity to extend β-strand conformations when compared to the behavior of Gly_{40}, indeed very sensitive to these conformations in the vacuum. This confirms the so-called inverse effect of side chains for polypeptides in water: aggregation is more the result of a low propensity for intramolecular folding than of a specific propensity for extending β-strands. In a hydrophobic environment, Aβ(1–40) displays significant helical propensity, as expected being this peptide part of a membrane protein, the amyloid precursor protein (APP).

References and Notes

(1) Dobson, C. M. *Trends Biochem. Sci.* **1999**, *24*, 329–332.
(2) Nelson, R.; Eisenberg, D. *Adv. Protein Chem.* **2006**, *73*, 235–282.
(3) Fändrich, M.; Dobson, C. M. *EMBO J.* **2002**, *21*, 5682–5690.
(4) Chiti, F.; Dobson, C. M. *Annu. Rev. Biochem.* **2006**, *75*, 333–366.
(5) La Penna, G. *J. Chem. Phys.* **2003**, *119*, 8162–8174.
(6) La Penna, G.; Morante, S.; Perico, A.; Rossi, G. C. *J. Chem. Phys.* **2004**, *121*, 10725–10741.
(7) Berg, B. A. *J. Stat. Phys.* **1996**, *82*, 323–342.
(8) Mitsutake, A.; Sugita, Y.; Okamoto, Y. *Biopolymers* **2001**, *60*, 96–123.
(9) Kinnear, B. S.; Jarrold, M. F.; Hansmann, U. H. E. *J. Mol. Graph. Mod.* **2004**, *22*, 397–403.
(10) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M. J.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.
(11) Garcia, A. E.; Sanbonmatsu, K. Y. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 2782–2787.
(12) Gallicchio, E.; Zhang, L. Y.; Levy, R. M. *J. Comput. Chem.* **2002**, *23*, 517–529.
(13) Ooi, T.; Oobatake, M.; Némethy, G.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **1987**, *84*, 3086–3090.
(14) Nicholls, A.; Honig, B. *J. Comput. Chem.* **1991**, *12*, 435–445.
(15) Rocchia, W.; Sridharan, S.; Nicholls, A.; Alexov, E.; Chiabrera, A.; Honig, B. *J. Comput. Chem.* **2002**, *23*, 128–137.
(16) Furlan, S.; La Penna, G.; Perico, A.; Cesaro, A. *Macromolecules* **2004**, *37*, 6197–6209.
(17) Eisenhaber, F.; Lijnzaad, P.; Argos, P.; Sander, C.; Scharf, M. *J. Comput. Chem.* **1995**, *16*, 273–284.
(18) Kabsch, W.; Sander, C. *Biopolymers* **1983**, *22*, 2577–2637.
(19) Koradi, R.; Billeter, M.; Wüthrich, K. *J. Mol. Graphics* **1996**, *14*, 51–55. (http://hugin.ethz.ch/wuthrich/software/molmol/index.html).
(20) Ohnishi, S.; Kamikubo, H.; Onitsuka, M.; Kataoka, M.; Shortle, D. *J. Am. Chem. Soc.* **2006**, *128*, 16338–16344.
(21) Doty, P.; Imahori, K.; Klemperer, E. *Proc. Natl. Acad. Sci. U.S.A.* **1958**, *44*, 424–431.
(22) Vorobjev, Y. N.; Scheraga, H. A.; Honig, B. *J. Phys. Chem.* **1995**, *99*, 7180–7187.

MA7022155