Structure and mechanism of maximum stability of isolated alpha-helical protein domains at a critical length scale

Eur. Phys. J. E (2013) **36**: 53
DOI 10.1140/epje/i2013-13053-8

THE EUROPEAN PHYSICAL JOURNAL E

Regular Article

Structure and mechanism of maximum stability of isolated alpha-helical protein domains at a critical length scale

Zhao Qin[1], Andrea Fabre[1,2], and Markus J. Buehler[1,3,4,a]

[1] Laboratory for Atomistic and Molecular Mechanics (LAMM), Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, 77 Massachusetts Ave., Room 1-235 A&B, Cambridge 02139, MA, USA
[2] Department of Chemical Engineering, Massachusetts Institute of Technology, 77 Massachusetts Ave., Cambridge, MA 02139, USA
[3] Center for Computational Engineering, Massachusetts Institute of Technology, 77 Massachusetts Ave., Cambridge, MA 02139, USA
[4] Center for Materials Science and Engineering, Massachusetts Institute of Technology, 77 Massachusetts Ave., Cambridge, MA 02139, USA

Received 29 April 2012 and Received in final form 12 September 2012
Published online: 29 May 2013 –

Abstract. The stability of alpha helices is important in protein folding, bioinspired materials design, and controls many biological properties under physiological and disease conditions. Here we show that a naturally favored alpha helix length of 9 to 17 amino acids exists at which the propensity towards the formation of this secondary structure is maximized. We use a combination of thermodynamical analysis, well-tempered metadynamics molecular simulation and statistical analyses of experimental alpha helix length distributions and find that the favored alpha helix length is caused by a competition between alpha helix folding, unfolding into a random coil and formation of higher-order tertiary structures. The theoretical result is suggested to be used to explain the statistical distribution of the length of alpha helices observed in natural protein structures. Our study provides mechanistic insight into fundamental controlling parameters in alpha helix structure formation and potentially other biopolymers or synthetic materials. The result advances our fundamental understanding of size effects in the stability of protein structures and may enable the design of *de novo* alpha-helical protein materials.

1 Introduction

The alpha helix (fig. 1a) is a very widely observed and the most prevalent secondary structure (which reflects $\sim 30\%$ of the entire Protein Data Bank) of proteins, characterized by a right-handed coil stabilized by hydrogen bonds between backbones of every 3.6 residues of the polypeptide chain [1] as distinct from the much less frequently found 3_{10} helices (3 residues per turn) and π-helices (4.1 residues per turn) [2–5]. The stability of the alpha helix secondary structure is important since in its absence protein domains may misfold, which results in compromised structural, mechanical, binding and other biologically functional properties that play an important role in the emergence of disease states [6–9]. A statistical analysis of experimental data for variegated amino acid sequences shows that most naturally existing alpha-helical domains are composed of segments of $\sim$10 residues in length [6, 10–12]. Since this overarching structural feature emerges from analyzing many of functionally very different protein molecules, we hypothesize that the favored length scale of alpha helices is driven by more fundamental principles than the specific amino acid sequence, solvent property or other biochemical features.

Even though the accessible structural data of protein molecules has increased rapidly during recent years, overarching models that explain the generic driving forces behind structure formation in protein materials remain limited [13]. This is partly because computer simulations of biological molecules are often confined to relatively short time scales, making it difficult to reach the native folded state by a conventional search method (*e.g.* using classical molecular dynamics simulation) [14]. Existing statistical models, combined with empirical parameters, provide predictive power toward understanding the general principles of folding and unfolding of alpha helices [15–17]. Their parameters, however, are usually empirically fitted, adding difficulty to integrate them with molecular models with parameters derived from *ab initio* calculations. Moreover, these models focus on the collective behavior of helices and do not account for the mechanistic insight of a single isolated alpha helix. Indeed, single alpha he-

[a] e-mail: mbuehler@MIT.EDU

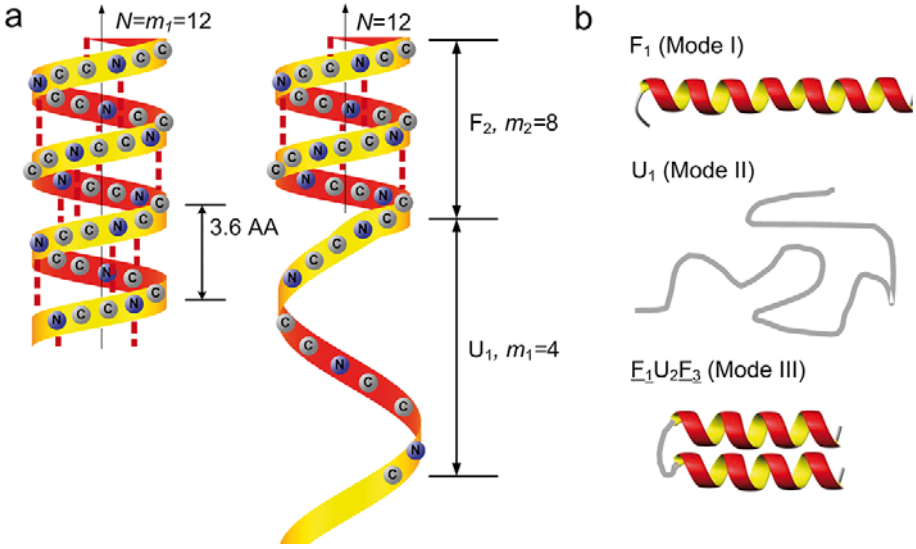


Fig. 1. Representative geometry of an alpha helix structure. a) The coiled structure of an alpha helix, displayed via a ribbon with 3.6 residues per turn, defined by the backbone nitrogen and carbon atoms. Each hydrogen bond forms between two of the amino acids are indicated by dashed lines. In the left structure shown here $N = m_1 = 12$ since all residues form an alpha helix secondary structure; in the right structure shown here the chain is composed of two segments represented by U_1F_2 as $m_1 = 4$ and $m_2 = 8$. b) Examples of the three modes of all the possible structures as summarized in table 1, and each of which is represented by F_1, U_1 and $\underline{F}_1U_2\underline{F}_3$, respectively.

lix formation is weakly cooperative, resulting in complex distributions of helical segment size [18]. For example, the Lifson-Roig helix-to-coil transition models predict the propensities for different chain lengths to form an alpha helix, but they only suggest an increasing propensity to form an alpha helix with increasing length [19]. This implies that longer polypeptide chains are always more inclined to form alpha helices, preventing the models from explaining complex higher-order folding behaviors of alpha helices [20]. In contrast, other models predict that the alpha helix strength decreases for increasing length, implying a less stable structure that unfolds more easily [21]. Such coarse-grained models, however,do not incorporate terms that can capture the details of the amino acid sequence and do not include the mechanistic interplay of alpha helix folding, unfolding and assembly. Zimm and Bragg's study predicts the average length of alpha helix in an infinite polyalanine repeat as 15–30 residues [15,22]. Again, this study does not account for the complex higher-order folding and this estimated length is much longer than most naturally existing alpha-helical segments of 5 to 14 residues [6,10–12]. The predictions made by these models are contradictory and form a paradox, and each of them alone fails to explain the observation of a favored length of alpha helices.

Here we explain the naturally favored alpha helix length based on a combination of a theoretical and computational model. The theoretical model is based on two key concepts that describe three possible structures that can be assumed by the helix-prone polypeptide (see fig. 1b). The first one is that the folded region of the polypeptide chain can either adopt the geometry of an alpha helix secondary structure or an unfolded structure that is free to adopt any conformation without constraints, resembling a random coil. The second concept is that higher-order tertiary structures can be formed by self-folding of longer alpha helix segments onto one another, where the total energy is lowered by the formation of additional inter-helix adhesion when alpha-helical regions align [23]. It is noted that helix bundles have been theoretically investigated in the literature [20,24]. We emphasize that our study focuses on the physical mechanism that leads to the critical length of an alpha helix with maximum stability by itself, and not stabilized by interactions with other proteins. We will not focus on any specific protein folding problem because that requires the sequence information, the full atomic model, the accurate force field and a large sampling space. This distinguishes our work from earlier analyses [25–28].

2 Model and methods

2.1 Thermodynamic model

Without considering the interactions among amino acids within an alpha helix, a polypeptide chain of N amino acids yields a total 2^N states with different distributions of amino acids in alpha helix state and disordered state. To facilitate the calculation, we describe the conformation state of the polypeptide chain by the combination of states of each segment as "F" for a segment of all amino acids in an alpha helix state and "U" for a segment of all amino acids in a disordered state. We use m_i to denote the number of amino acids within each segment as illustrated in fig. 1. Because of the geometry character of alpha helix

Table 1. Summary of all states of a helix-prone polypeptide. The entire polypeptide is composed of up to three alpha helix segments (noted by F_1 to F_3) connected by unfolded sections (noted by U_i) as illustrated in fig. 1b. Each alpha helix is of length larger than the minimum length requirement $N_{\min} = 5$. The formula for the internal energy of each state is given in the table, with a detailed definition of E given in eq. (3).

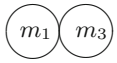
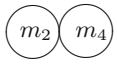
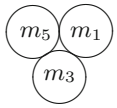
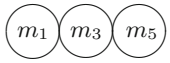
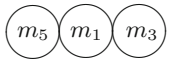
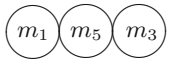
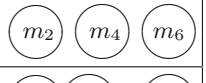
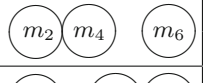
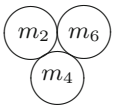
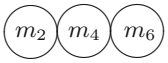
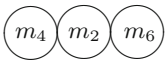
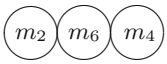
S	Conformation	Mode	Internal energy E_f	Illustration
1	F_1	I	$E(N)$	
2	F_1U_2	II	$E(m_1)$	
3	U_1	II	0	--
4	U_1F_2	II	$E(m_2)$	
5	$U_1F_2U_3$	II	$E(m_2)$	
6	$F_1U_2F_3$	II	$E(m_1) + E(m_3)$	
7	$F_1U_2F_3U_4$	II	$E(m_1) + E(m_3)$	
8	$\underline{F}_1U_2\underline{F}_3$	III	$E(m_1) + E(m_3) - \varepsilon d(\min(m_1, m_3))$	
9	$\underline{F}_1U_2\underline{F}_3U_4$	III	$E(m_1) + E(m_3) - \varepsilon d(\min(m_1, m_3))$	
10	$U_1F_2U_3F_4$	II	$E(m_2) + E(m_4)$	
11	$U_1F_2U_3F_4U_5$	II	$E(m_2) + E(m_4)$	
12	$U_1\underline{F}_2U_3\underline{F}_4$	III	$E(m_2) + E(m_4) - \varepsilon d(\min(m_2, m_4))$	
13	$U_1\underline{F}_2U_3\underline{F}_4U_5$	III	$E(m_2) + E(m_4) - \varepsilon d(\min(m_2, m_4))$	
14	$F_1U_2F_3U_4F_5$	II	$E(m_1) + E(m_3) + E(m_5)$	
15	$F_1U_2F_3U_4F_5U_6$	II	$E(m_1) + E(m_3) + E(m_5)$	
16	$\underline{F}_1U_2\underline{F}_3U_4F_5$	III	$E(m_1) + E(m_3) + E(m_5) - \varepsilon d(\min(m_1, m_3))$	
17	$\underline{F}_1U_2\underline{F}_3U_4F_5U_6$	III	$E(m_1) + E(m_3) + E(m_5) - \varepsilon d(\min(m_1, m_3))$	
18	$F_1U_2\underline{F}_3U_4\underline{F}_5$	III	$E(m_1) + E(m_3) + E(m_5) - \varepsilon d(\min(m_3, m_5))$	
19	$F_1U_2\underline{F}_3U_4\underline{F}_5U_6$	III	$E(m_1) + E(m_3) + E(m_5) - \varepsilon d(\min(m_3, m_5))$	
20	$\underline{F}_1U_2F_3U_4\underline{F}_5$	III	$E(m_1) + E(m_3) + E(m_5) - \varepsilon d(\min(m_1, m_5))$	
21	$\underline{F}_1U_2F_3U_4\underline{F}_5U_6$	III	$E(m_1) + E(m_3) + E(m_5) - \varepsilon d(\min(m_1, m_5))$	
22	$\underline{F}_1U_2\underline{F}_3U_4\underline{F}_5$	III	$E(m_1) + E(m_3) + E(m_5)$ $-\varepsilon d(\min(m_1, m_3)) - \varepsilon d(\min(m_3, m_5))$ $-\varepsilon d(\min(m_1, m_5))$	
23	$\underline{F}_1U_2\underline{F}_3U_4\underline{F}_5U_6$	III	$E(m_1) + E(m_3) + E(m_5)$ $-\varepsilon d(\min(m_1, m_3)) - \varepsilon d(\min(m_3, m_5))$ $-\varepsilon d(\min(m_1, m_5))$	
24	$\underline{F}_1U_2\underline{F}_3U_4\underline{F}_5$	III	$E(m_1) + E(m_3) + E(m_5) - \varepsilon d(\min(m_1, m_3))$ $-\varepsilon d(\min(m_3, m_5))$	
25	$\underline{F}_1U_2\underline{F}_3U_4\underline{F}_5U_6$	III	$E(m_1) + E(m_3) + E(m_5) - \varepsilon d(\min(m_1, m_3))$ $-\varepsilon d(\min(m_3, m_5))$	
26	$\underline{F}_1U_2\underline{F}_3U_4\underline{F}_5$	III	$E(m_1) + E(m_3) + E(m_5) - \varepsilon d(\min(m_1, m_3))$ $-\varepsilon d(\min(m_1, m_5))$	
27	$\underline{F}_1U_2\underline{F}_3U_4\underline{F}_5U_6$	III	$E(m_1) + E(m_3) + E(m_5) - \varepsilon d(\min(m_1, m_3))$ $-\varepsilon d(\min(m_1, m_5))$	
28	$\underline{F}_1U_2\underline{F}_3U_4\underline{F}_5$	III	$E(m_1) + E(m_3) + E(m_5) - \varepsilon d(\min(m_3, m_5))$ $-\varepsilon d(\min(m_1, m_5))$	
29	$\underline{F}_1U_2\underline{F}_3U_4\underline{F}_5U_6$	III	$E(m_1) + E(m_3) + E(m_5) - \varepsilon d(\min(m_3, m_5))$ $-\varepsilon d(\min(m_1, m_5))$	
30	$U_1F_2U_3F_4U_5F_6$	II	$E(m_2) + E(m_4) + E(m_6)$	
31	$U_1F_2U_3F_4U_5F_6U_7$	II	$E(m_2) + E(m_4) + E(m_6)$	
32	$U_1\underline{F}_2U_3\underline{F}_4U_5F_6$	III	$E(m_2) + E(m_4) + E(m_6) - \varepsilon d(\min(m_2, m_4))$	
33	$U_1\underline{F}_2U_3\underline{F}_4U_5F_6U_7$	III	$E(m_2) + E(m_4) + E(m_6) - \varepsilon d(\min(m_2, m_4))$	
34	$U_1F_2U_3\underline{F}_4U_5\underline{F}_6$	III	$E(m_2) + E(m_4) + E(m_6) - \varepsilon d(\min(m_4, m_6))$	
35	$U_1F_2U_3\underline{F}_4U_5\underline{F}_6U_7$	III	$E(m_2) + E(m_4) + E(m_6) - \varepsilon d(\min(m_4, m_6))$	

Table 1. Continued.

S	Conformation	Mode	Internal energy E_f	Illustration
36	$U_1\underline{F}_2U_3F_4U_5\underline{F}_6$	III	$E(m_2)+E(m_4)+E(m_6)-\varepsilon d(\min(m_2,m_6))$	
37	$U_1\underline{F}_2U_3F_4U_5\underline{F}_6U_7$	III	$E(m_2)+E(m_4)+E(m_6)-\varepsilon d(\min(m_2,m_6))$	
38	$U_1\underline{F}_2U_3\underline{F}_4U_5\underline{F}_6$	III	$E(m_2)+E(m_4)+E(m_6)-\varepsilon d(\min(m_2,m_4))$ $-\varepsilon d(\min(m_4,m_6))$ $-\varepsilon d(\min(m_2,m_6))$	
39	$U_1\underline{F}_2U_3\underline{F}_4U_5\underline{F}_6U_7$	III	$E(m_2)+E(m_4)+E(m_6)-\varepsilon d(\min(m_2,m_4))$ $-\varepsilon d(\min(m_4,m_6))$ $-\varepsilon d(\min(m_2,m_6))$	
40	$U_1\underline{F}_2U_3\underline{F}_4U_5\underline{F}_6$	III	$E(m_2)+E(m_4)+E(m_6)-\varepsilon d(\min(m_2,m_4))$ $-\varepsilon d(\min(m_4,m_6))$	
41	$U_1\underline{F}_2U_3\underline{F}_4U_5\underline{F}_6U_7$	III	$E(m_2)+E(m_4)+E(m_6)-\varepsilon d(\min(m_2,m_4))$ $-\varepsilon d(\min(m_4,m_6))$	
42	$U_1\underline{F}_2U_3\underline{F}_4U_5\underline{F}_6$	III	$E(m_2)+E(m_4)+E(m_6)-\varepsilon d(\min(m_2,m_4))$ $-\varepsilon d(\min(m_2,m_6))$	
43	$U_1\underline{F}_2U_3\underline{F}_4U_5\underline{F}_6U_7$	III	$E(m_2)+E(m_4)+E(m_6)-\varepsilon d(\min(m_2,m_4))$ $-\varepsilon d(\min(m_2,m_6))$	
44	$U_1\underline{F}_2U_3\underline{F}_4U_5\underline{F}_6$	III	$E(m_2)+E(m_4)+E(m_6)-\varepsilon d(\min(m_4,m_6))$ $-\varepsilon d(\min(m_2,m_6))$	
45	$U_1\underline{F}_2U_3\underline{F}_4U_5\underline{F}_6U_7$	III	$E(m_2)+E(m_4)+E(m_6)-\varepsilon d(\min(m_4,m_6))$ $-\varepsilon d(\min(m_2,m_6))$	

($t = 3.6$ residues per turn for alpha helix), the minimum length for each alpha helix segment is 5 residues (stabilized by at least a single hydrogen bond), and the minimum length of each disordered segment is one residue. Therefore, for each possible conformation denoted by S, we can estimate the minimum length requirement as summarized in table 1. The canonical partition function of each conformation state with a known length for each segment as m_i is given by $Z(N|S|\{m_i\}) = \Omega_{\rm tot}\exp[-\beta E_f]$ [29], where $\beta = 1/(K_BT)$ is the thermodynamic factor, E_f is the total internal energy of those conformations, m_i is the length of each segment that only subjects to geometric limits and $\Omega_{\rm tot}$ is the statistical weight defined as the number of conformations within the state. The canonical partition function of each conformation state without any limit by segment length is given by

$$Z_0(N) = \sum_S \sum_{\{m_i,\, i=1\ldots\}} Z(N|S|\{m_i,\, i=1\ldots\}) \qquad (1)$$

to sum up all the possible combinations of different segment number and length of each of them as summarized in table 1 with the only constraint that $\sum m_i = N$.

The total internal energy can be calculated via the sum of the energy of each segment, as well as the interacting term given by

$$E_f = \sum E(m_i) - \varepsilon d \sum \delta_{ij}\min(m_i,\, m_j), \qquad (2)$$

where $E(m_i)$ is the internal energy of the segment with the length of m_i amino acids, ε is the non-bonded interaction of unit length between two self-folded alpha helices, d is the helix rise along the helix axis for each residue, and δ_{ij} is the mark that equals to one as the two alpha helix segments are self-folded and equals zero when they are not. We include the second term on the right side of eq. (2) because many protein structures seen in the Protein Data Bank [13] show the structural characteristic that helical segments within the same polypeptide are separated by several amino acids that have a random coil structure. For example, the pore helix structures at the center of ion channels are composed of bundles of self-folded alpha helices. By considering the internal energy in the form of eq. (2) for all the possible segment lengths for all the possible conformations as summarized in table 1, we calculate the internal energy of intact alpha helix (mode I), partial unfolded alpha helix (mode II) and self-folded alpha helix (mode III), including symmetric and asymmetric folding, and their partition functions are given by $Z(N \mid \text{mode I}) = \sum_{S=1}\sum_{\{m_i,i=1\}} Z(N|S|\{m_i, i=1\}) = Z(N|1|m_1)$, $Z(N \mid \text{mode II}) = \sum_{S=\{2,3,4\ldots\}}\sum_{\{m_i,i=1\ldots\}} Z(N|S|\{m_i, i=1\ldots\})$ and $Z(N|\text{mode III}) = \sum_{S=\{8,9,12,13\ldots\}}\sum_{\{m_i,i=1\ldots\}} Z(N|S|\{m_i, i=1\ldots\})$.

2.2 Internal energy of alpha-helical segment

We set up a homogeneous alpha-helical model that treats each residue equally to calculate the internal energy of each alpha-helical segment. The internal energy includes the energy stored in hydrogen bonds, and the deformation

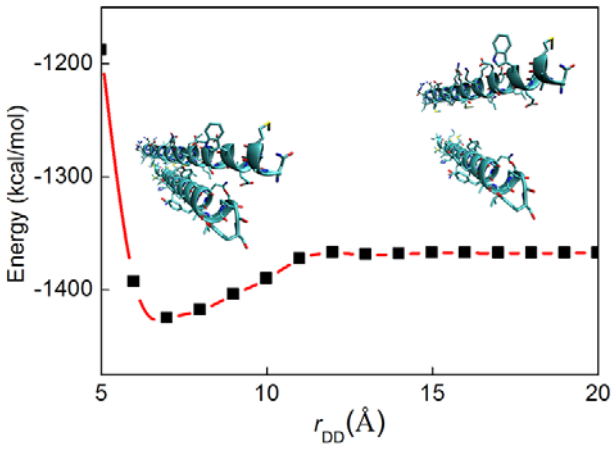


Fig. 2. We use the natural form of a coiled coil that is composed of two alpha helices to calculate the non-bonded interaction of unit length between two self-folded alpha helices. The atomic model is obtained from protein data bank (with PDB ID 2FB5, domain A) and each of the alpha helices has a length of 36 amino acids. We change the distance r_{DD} between the two helix axis and measure the energy as a function. The adhesion energy is thereby given by the energy difference between the lowest point and the energy of two helices far away from each other. The non-bonded interaction for this model is given by $\varepsilon = (E_{x=20} - E_{x=7})/(36d)$.

energy of the backbone, given by

$$E(m_i) = -\max(m_i - t - 2, 0)E_{b\_\mathrm{int}} + (m_i - t)E_{\mathrm{hyr}} - \min(2, m_i - t)E_{b\_\mathrm{ext}} + (m_i - t)E_e, \quad (3)$$

where E_e is the bending energy of each residue's backbone within the helix conformation, E_{hyr} is the energy difference for each internal amino acid in alpha helix state and in disordered state caused by solvent effect, E_{b_ext} is the energy of the hydrogen bond at the end region, E_{b_int} is the energy of each hydrogen bond at the interior of each helical segment and $t = 3.6$ is the number of residues per turn for an alpha helix. It is noted that in many former studies the bending term is not explicitly considered, but it is explicitly included here to make the physical meaning of each term more clear. We include two different hydrogen bonding energy terms here to describe a chemical cooperative effect as has been shown in protein structures. The terminal effect is included by $(m_i - t)$ in each energy term.

We obtain the reference value of each of those energy terms as follows. We first measure the adhesion energy between two alpha helices in their natural state within a coiled coil obtained from the protein data bank (with PDB ID 2FB5, domain A) and model the protein structure by using the CHARMM19 all-atom force field with an effective Gaussian model for the water solvent as shown in fig. 2 [30]. By changing the distance between two coils we obtain the reference value of $\varepsilon = 1.0\,\mathrm{kcal/mol/Å}$. $E_e = 2K_BTL_pb/D^2$ is the average bending energy of an amino acid within the helix conformation, where L_p is the persistence length of a polypeptide chain with a reference value of $L_{p0} = 5\,\mathrm{Å}$ at a reference temperature of $T_0 = 300\,\mathrm{K}$ [31], D is the average diameter of the alpha helix measured by doubling the distance from the backbone atoms to the helix axis (with its value as $D = 3.2\,\mathrm{Å}$ [32]) and b is the contour length of an amino acid within the polypeptide chain with its reference value of 3.7 Å [31]. These parameters yield the reference value of $E_e = 2.2\,\mathrm{kcal/mol}$ at 300 K. For E_{hyr} which includes only the solvent effect and no other energy components, we use a difference-in-difference measurement. We first set up two polypeptide chains composed of 30 amino acids with the same type; one of an alpha helix conformation and the other one of a fully unfolded conformation. We then measure the energy difference of the two models in solvent as given by ΔG_{wt} (this energy difference stems from hydrogen bonds, deformation of the polypeptide chain and solvent energy at the surface), and in vacuum as given by ΔG_{va} (this energy difference comes from hydrogen bonds and deformation of polypeptide chain), as summarized in table 2. The energy difference caused by solvent for an amino acid of this specific type in different states is thereby given by $(\Delta G_{wt} - \Delta G_{va})/30$. We repeat this process of all the amino acid types and then obtain the average value E_{hyr} by using a statistical weight of each amino acid type.

We use the well-tempered metadynamics method [33, 34] to measure the free energy landscape of each hydrogen bond in the interior region and the edge region of an alpha helix to give estimations of E_{b_int} and E_{b_ext}, respectively. The simulation includes a polyalanine chain of 20 amino acids with an initial conformation of alpha helix after equilibration using the CHARMM27 all-atom force field and an explicit solvent (TIP3P water molecule model) box with dimension of $100 \times 40 \times 40\,\mathrm{Å}^3$. We set the collective variable as the distance between hydrogen and oxygen atoms within a hydrogen bond. To eliminate the angular effect, we apply a restraint to the system that only allows the hydrogen bond angle to fluctuate within 60° of range. The evolvements of the free energy landscapes of hydrogen bond at the edge and interior are as illustrated in fig. 3, from which we can obtain the bond energy by measuring the converged barrier height. All reference values of the energy parameters are as summarized in table 3.

2.3 Statistical weight of each state

The Worm-Like Chain (WLC) model [35] is adopted here as the elastic description of the polypeptide backbone. The conformation energy of each unfolded segment is

$$E_c(m_i|\mathrm{U}) = -\int_0^R \frac{K_BT}{L_p}\left[\frac{1}{4(1-h/L)^2} - \frac{1}{4} + \frac{h}{L}\right]\mathrm{d}h, \quad (4)$$

where $L = bm_i$ is the contour length of the polypeptide of the m_i unfolded amino acids, and R is the expected end-to-end length of the polypeptide with the value [36] of $R^2 = 2L_pL\{1 - L_p[1 - \exp(-L/L_p)]/L\}$ for the unfolded segment. We thereby have the integral as $E_c(m_i|\mathrm{U}) = -K_BT\{L/[4(1 - R/L)] - (R + L)/4 + R^2/(2L)\}/L_p$ and

Table 2. Summary of ΔG_{wt} (energy difference which comes from hydrogen bond, deformation of polypeptide chain and solvent energy at surface) and ΔG_{va} (energy difference which comes from hydrogen bond and deformation of polypeptide chain) for polypeptide chains with same type of amino acids in the state of alpha helix and fully unfolded state ($\Delta G_{wt} = G_{wt\_F} - G_{wt\_U}$, $\Delta G_{va} = G_{va\_F} - G_{va\_U}$). $P_{\text{a.a.}}$ gives the statistic weight of each amino acid type by calculating their potions in natural protein segments which is deposited in PDB with alpha helix as the secondary structure.

Type	ΔG_{wt} (kcal/mol)	ΔG_{va} (kcal/mol)	$(\Delta G_{wt} - \Delta G_{va})/30$ (kcal/mol)	$P_{\text{a.a.}}$
ALA	−49.6	36.3	−2.9	0.118
ARG	−55.3	50.7	−3.5	0.060
ASN	12.2	41.5	−1.0	0.032
ASP	−43.2	50	−3.1	0.047
CYS	−52.7	−42.1	−0.3	0.009
GLU	−85.1	−213.6	4.3	0.090
GLN	−77.1	58.1	−4.5	0.046
GLY	4.5	84.9	−2.7	0.035
HIS	−7.6	−2	−0.2	0.021
ILE	−40.5	−141.4	3.4	0.063
LEU	−23.2	−99.6	2.5	0.121
LYS	−21.4	73.3	−3.2	0.066
MET	−77.5	−120	1.4	0.027
PHE	−33.9	−122.2	2.9	0.040
PRO	88.9	88.3	0.02	0.020
SER	−109.1	−113.6	0.15	0.045
THR	−130	−7.3	−4.1	0.043
TRP	−54.6	−160.6	3.5	0.015
TYR	−19.6	−108	2.9	0.034
VAL	−43.5	−83.6	1.3	0.067

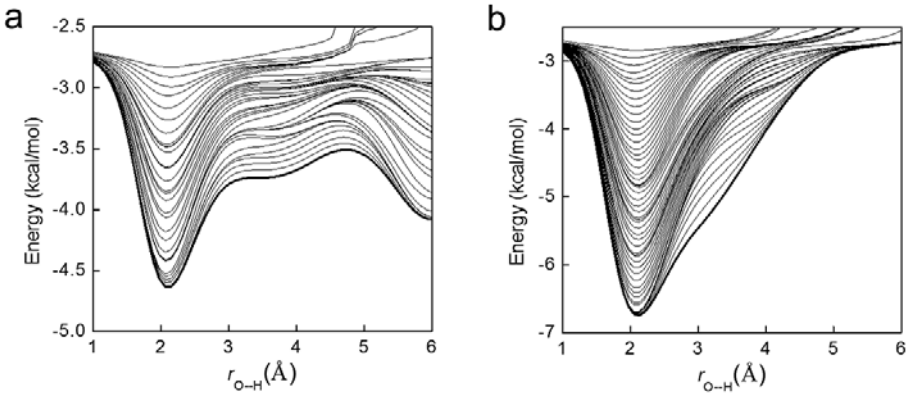


Fig. 3. The free-energy landscape of each single hydrogen bond as a function of the distance between the oxygen atom in the amino acid backbone and the hydrogen atom of the next amino acid in the neighbor turn at the corresponding position. The starting energy deposition rate for metadynamics is set to be $\omega = 0.01\,\text{kcal}\,\text{mol}^{-1}\,\text{ps}^{-1}$ with a Gaussian width of 0.35 Å. Different curves in each panel correspond to different time with different amount of bias potential applied to develop the energy landscape and the time difference between two neighbor curves is 100 ps. The converged curves at the bottom of each panel gives the physical energy landscapes of the hydrogen bond at different positions. The hydrogen bond energy is measured by the height of energy barrier from the lowest energy point around $r_{\text{O–H}} = 2.1$ Å to the length corresponding to bond breaking length $r_{\text{O–H}} > 5$ Å. The calculation is performed by metadynamics for hydrogen bond energy at the edge region in panel a) and interior region in panel b) of an alpha helix.

Table 3. Summary of geometric and energy parameters for the theoretical model (values as given in the literature). The effect of the values of E_b and ε is investigated within the sensitivity analysis as shown in figs. 6a and b.

Variable		Value
Temperature	T_0	300 K
Persistence length of peptides in T_0	L_{p0}	5.0 Å [31]
Contour length per residue	b	3.7 Å [31]
Number of residue per turn in alpha helix	t	3.6 [2]
Helix rise per residue.	d	1.5 Å[(a)]
Average diameter of an alpha helix	D	3.2 Å[(b)]
Non-bonded interaction between unit length alpha helices	ε	1.0 kcal/mol/Å[(c)] [37]
Average value of solvent effect on each amino acid	E_{hyr}	−0.123 kcal/mol[(d)]
Energy of each hydrogen bond at the end region of alpha helix	E_{b_ext}	1.12 kcal/mol[(e)]
Energy of each hydrogen bond at the interior of alpha helix	E_{b_int}	4.00 kcal/mol[(e)]

[(a)] Based on the fact that the pitch of an alpha helix is 5.4 Å, and 3.6 amino acid per turn [32].
[(b)] Average diameter of an alpha helix based on its backbone.
[(c)] Obtained from fig. 2 in this study.
[(d)] Calculated from table 2 in this study.
[(e)] Obtained from fig. 3 in this study.

the number of conformations of each segment is obtained via [29]

$$-K_B T \ln(\Omega_i(m_i)) = E_c(m_i), \tag{5}$$

where Ω_i is the statistical weight as the number of conformations of the segment i. We have $E_c(m_i|\mathrm{F}) = 0$ because there is a single conformation for an alpha helix segment. The above analysis yields a general form of the statistical weight Ω_i of each segment m_i as

$$\Omega_i(m_i) = \begin{cases} 1, & \text{for F,} \\ \exp(\{L/[4(1-R/L)] & \\ -(R+L)/4 + R^2/(2L)\}/L_p), & \text{for U,} \end{cases} \tag{6}$$

and the total statistical weight can be calculated by

$$\Omega_{\mathrm{tot}} = \Omega_0 \prod \Omega_i(m_i), \tag{7}$$

where $\Omega_0 = \exp(-E_0/(K_B T))$ is a ground state constant and E_0 is the ground state conformational energy of the polypeptide before unfolding.

We select typical values for the geometric parameters as summarized in table 3 to obtain a quantitative estimate of the probability distribution as summarized in table 1. It is noted that L_{p0} is given by the persistence length of the polypeptide in the reference temperature of T_0, and for other temperatures T, the persistence length is $L_p = L_{p0} T_0/T$. This relation means that a higher temperature introduces more fluctuations to the polypeptide and makes it easier to deform. The probability of the polypeptide to form an intact alpha helix without unfolding can be derived from the canonical partition function by

$$P(N) = Z(N|\text{mode I})/Z_0(N), \tag{8}$$

where $Z_0(N) = Z(N \mid \text{mode I}) + Z(N \mid \text{mode II}) + Z(N \mid \text{mode III})$ is the total partition function equivalent to eq. (1). This formula incorporates the mechanism of the competition between alpha helix folding and unfolding as well as the effect of self-folding. The reference values of all the parameters needed to calculate $P(N)$ are obtained from molecular dynamics simulations and literature sources and are summarized in table 3.

3 Results and discussion

3.1 Length effect from thermodynamic model

We now apply this model and calculate $P(N)$ as a function of the polypeptide length as shown in fig. 4. We observe that $P(N)$ features a plateau between 9 to 17 amino acids with a polypeptide length at its middle of $C_N = 13$ amino acids (at 300 K), which corresponds to the polypeptide length with the maximum probability C_P to form an intact alpha helix. For polypeptides shorter than C_N, P decreases with decreasing N. Similarly, for polypeptides longer than C_N, an increasing length also leads to a decreasing P. We also notice that C_N does not change significantly as the temperature increases from 300 to 400 K as shown in fig. 4, suggesting that C_N is not very sensitive to temperature. The phenomenon that the increasing temperature leads to a decreased C_P agrees with our experience that increasing temperature leads to increased fluctuations in the polypeptide that break the hydrogen bonds, and thus decreases the content of alpha helix as the protein structure is denaturalized. The existence of C_N in our model is caused by the changing probability of each

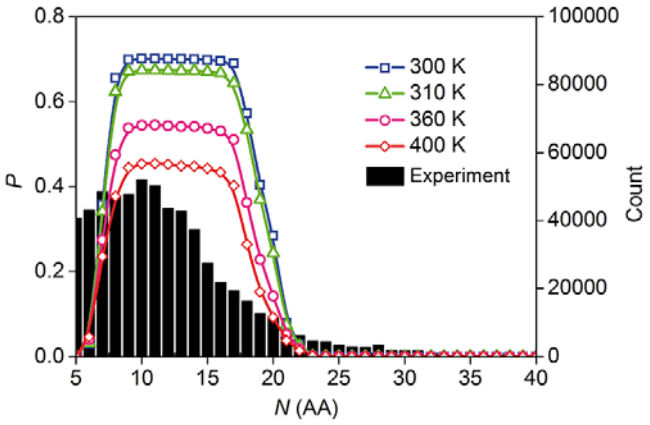


Fig. 4. The probability of the polypeptide to form an intact alpha helix as a function of its length and the number of alpha helices as a function of length in Protein Data Bank. Histogram that shows the number of alpha helices with that particular length, revealing that alpha helices with a length of 10 are most common. The data is the result of a statistic calculation based on 46030 high-resolution protein crystal structures obtained from the Protein Data Bank. Those structures are typically obtained by X-ray crystallography or nuclear magnetic resonance spectroscopy and cover a very broad variety of biological molecules. The secondary structures of each amino acid within the proteins are determined by the DSSP algorithm [12]. The continuous curves are the results obtained from our theoretical model at different temperatures. The peak point of each distribution refers to the critical polypeptide length (C_N) with a maximum probability (C_P) in forming an intact alpha helix. The reason our model predicts no alpha helix for peptides smaller than 5 is that our model assumes the minimum length of $N_{\min} = 5$ to form alpha helix (3.6 amino acids per turn and we need at least one hydrogen bond forming).

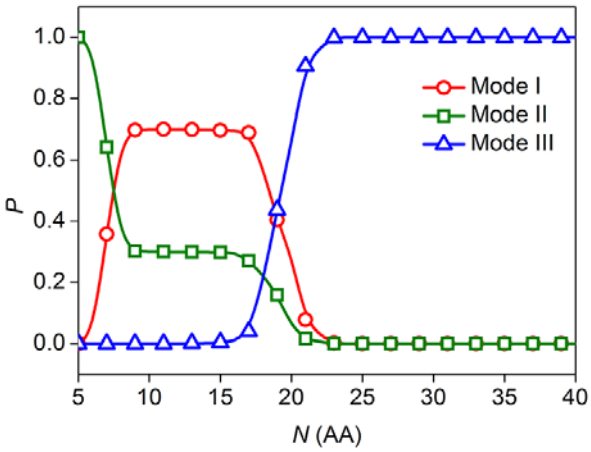


Fig. 5. The probability distribution of the polypeptide for each of the three modes as summarized in table 1 under 300 K temperature, *i.e.*, intact alpha helix (mode I), partial unfolded alpha helix (mode II) and self-folded alpha helix (mode III). The probability distribution of mode I forms a plateau between $N = 9$ and $N = 17$, we use the weight center of the plateau as $C_N = 13$ for the peak point. The probability of mode III significantly increases for $N > C_N$, while the probability of mode II decreases outside the plateau region.

folding mode for increasing N as shown in fig. 5. The probability of mode III is almost 0 as constant for $N < C_N$ region while it significantly increases for $N > C_N$. However, the probability of mode II decreases outside the plateau region. The combination of the behaviors of those two modes leads to the peak value of the probability of mode I since $\sum Z(N|\text{mode I, II, III}) = Z_0(N)$.

The values for E_{b_int} and ε are typically found in a range of possible values depending on the sequence and solvent conditions [2,37,38]. To examine the effect of varying these parameters we alter their values in our model and calculate the probability distribution for these cases. For each set of parameters we identify the critical length that leads to the maximum probability to form an alpha helix. The result depicted in fig. 6a shows that the maximum probability C_P for the formation of an alpha helix decreases with decreasing E_{b_int} and it shows no strong dependence on ε, while the critical polypeptide length C_N decreases with increasing ε and decreasing E_{b_int} (fig. 6b). We find that the critical polypeptide length C_N varies from 10 to 17 amino acids for a wide energy range (E_{b_int} from 3 to 5 kcal/mol and ε from 1 to 1.5 kcal/mol/Å). This indicates that the critical length depends on those energy terms. We observe that the smaller ε and greater E_{b_int} lead to the greater C_N. This result may be important for the understanding and design of alpha helix assembly processes because both ε and E_{b_int} are weak interactions in protein structures that can be directly controlled by external factors such as the temperature, solvent polarity and *p*H value. We also test the dependence of C_N on E_{b_ext} and E_{hyr} as summarized in figs. 6c and d, respectively. It is observed that increasing E_{b_ext}, which means a smaller cooperative effect as E_{b_ext} and E_{b_int} become more similar, leads to decreasing C_N. It is also observed that decreasing E_{hyr} (more hydrophobic) leads to a greater C_N.

3.2 alpha helix stability in metadynamics

We now carry out direct molecular simulations to test whether the length effect revealed by our theoretical model is also observed in a chemistry-based molecular model. We calculate the free-energy landscape for the alpha helix similarity parameter of a polyalanine chain by using the well-tempered metadynamics method [33,34]. The reason why we use polyalanine in our simulations is that alanine is regarded as the most stabilizing residue within alpha helices (47% of all the alanine residues are within alpha helices for all the protein structures we surveyed). Our metadynamics simulations are performed by NAMD implemented in the PLUMED package [33,34,39]. We study the stability of polyalanine chains in explicit solvent environment with different lengths. The initial geometry of the alpha helix is set up according to the standard geometry of an alpha helix of $(\phi^{\text{ref}}, \psi^{\text{ref}}) = (-58°, -47°)$ [32], the atomic interaction is modeled using the CHARMM27 force field and the solvent environment includes explicit TIP3P water molecule model and ionic concentration of

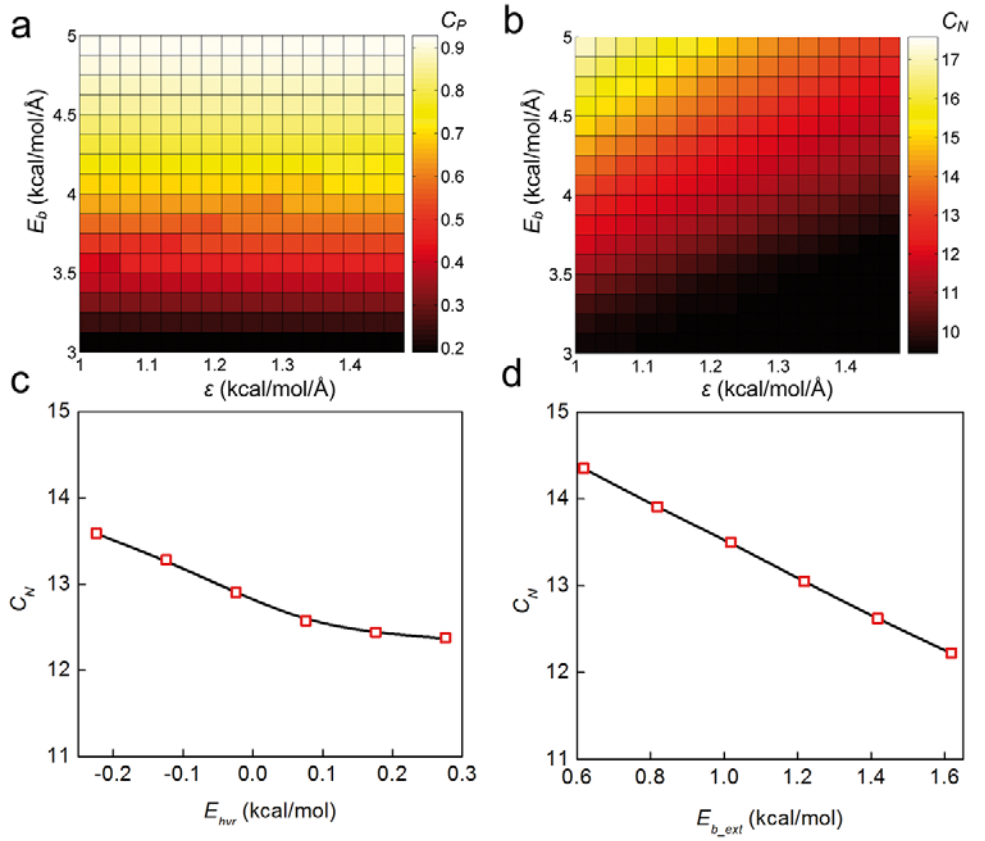


Fig. 6. Maximum probability and critical length of the polypeptide to form an intact alpha helix for varied hydrogen bond energies E_{b_int} and non-bonded adhesive energies ε. a) Maximum probability C_P, which refers to the P-axis value of the peak point in fig. 4, of the polypeptide to adopt the alpha-helical conformation. b) Critical length C_N, which refers to the N-axis value of the peak point in fig. 4, of the polypeptide with a maximum probability to form alpha helix. The values of the maximum probability and critical length for each combination of the hydrogen bond energies and non-bonded adhesive energies are defined by the color bars in panels a) and b) respectively. c) Critical length C_N of the polypeptide as a function of the hydrophobic effect given by E_{hyr}. d), Critical length C_N of the polypeptide as a function of the cooperative effect given by E_{b_ext}.

0.308 mol/L (equals to physiological saline environment of 0.154 mol/L sodium chloride concentration) with the numbers of cations and anions carefully controlled to neutralize the total charge of the system.

We use an NPT ensemble controlled by a Langevin thermostat and barostat (constant number of particles, constant pressure (1 atm) and constant temperature (300 K)) through the equilibrium stage. The typical system is of the size $100 \times 40 \times 40\,\text{Å}^3$ composed of ∼5000 water molecules and 30 ions. The system is set to be periodic in all directions and Particle Mesh Ewald method (with a lattice size of 1 Å) is used to accurately compute the electrostatic interactions. The integration time step is select to be 2 fs and rigid bond model is applied for the hydrogen atoms. We first equilibrate the solvent box by fixing the polyalanine for 200 ps as the volume and energy of the system converges, and then equilibrate the entire system without constraints for 5.4 ns to ensure that the protein structure has been fully equilibrated by examining the convergence of the root mean square deviation of the atoms within the protein structure.

After equilibration we perform a Well-Tempered Metadynamics calculation until convergence of the free-energy landscape is reached. The collective variable we use to analyze the free-energy landscape is the alpha helix similarity, which is defined as

$$S = \frac{1}{2N}\sum_{i=1}^{N}[\cos(\phi - \phi^{\mathrm{ref}}) + \cos(\psi - \psi^{\mathrm{ref}})], \qquad (9)$$

where (ϕ, ψ) are the two dihedral angles of the backbone of two neighborhood amino acids within the internal region of the peptide, and $(\phi^{\mathrm{ref}}, \psi^{\mathrm{ref}})$ defines the conformation of standard alpha helix. It is noted that as $(\phi, \psi) \to (\phi^{\mathrm{ref}}, \psi^{\mathrm{ref}})$, $S \to 1$, meaning that the conformation of the peptide has the geometry of an alpha helix. Other parameters include the enhanced temperature of 1500 K where the collective variable is sampled, the starting Gaussian height is 0.1 kcal/mol and the deposition interval is 200 fs, corresponding to a deposition rate of $0.5\,\text{kcal mol}^{-1}\,\text{ps}^{-1}$. The probability for the polypeptide to form alpha helix conformation is given by

$$P_{\mathrm{meta}}(N) = \int_{S_0}^{1} \exp\left(-\frac{F(S)}{K_B T}\right) \mathrm{d}s \Big/ \int_{-1}^{1} \exp\left(-\frac{F(S)}{K_B T}\right) \mathrm{d}s, \qquad (10)$$

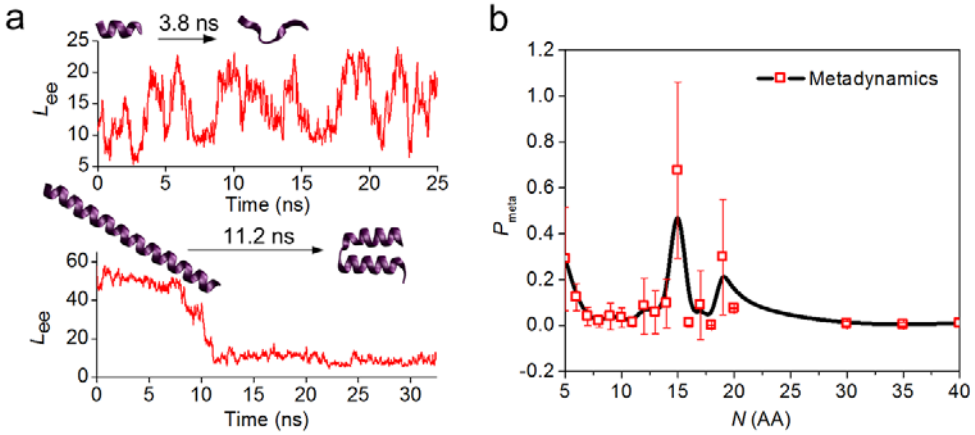


Fig. 7. Length effect on the folding process of polyalanine revealed by Well-Tempered Metadynamics molecular simulations. a) Conformation of a short polyalanine with 9 amino acids and the conformation of a long polyalanine with 40 amino acids. The plots in this panel present the end-to-end lengths of the short polyalanine(top) and long polyalanine (bottom) as a function of the simulation trajectories (in units of ns). b) Probabilities of the polyalanin chains to form an intact alpha helices, for chain lengths from 5 to 40 amino acids. The solid curve, with the purpose of guidance, is fitted to the data according to a B-spline function. The peak point corresponds to the maximum probability of the polyalanin chain to form an intact alpha helix and is found at a critical length of 16.3 ± 2.1 amino acids. This number and range is obtained by fitting the data points with a Gaussian function.

where $F(S)$ is the free-energy surface obtained in Well-Tempered Metadynamics calculation [34] and $S_0 = 0.6$ is the boundary value used in this study to define alpha helix conformation. We alter S_0 by ± 0.1 and recalculate $P_{\rm meta}$ to obtain the error bar of each data point.

To examine the length effect we compute the probability of each polypeptide chain length to form an intact alpha helix conformation by calculating the proportion of structures of alpha helix conformation among all conformational samplings, by integrating over its free-energy landscape (as given in eq. (10)). We carry out these simulations for polypeptide chains with lengths ranging from 5 to 40 amino acids and calculate their probability to adopt intact alpha-helical geometries. Figure 7a shows that in the simulations of the short polypeptide chain with 9 amino acids and less the structure unfolds rapidly and adopts a completely random coil conformation. In contrast, longer polypeptide chains such as the one with 40 amino acids behave very differently and are seen to self-fold into a helix hairpin. Both cases, the short and the very long polypeptide chains, display a rather small probability to remain in their initial straight alpha helix conformation. Notably we also observe that there exists an intermediate length that maximizes the probability to form an alpha helix length. These observations, including the findings made for the two extreme cases, agree with the assumptions made earlier in the development of the theoretical model.

The probabilities $P_{\rm meta}(N)$ of all polypeptide chains considered in our simulations to form an intact alpha helix conformation is plotted as a function of their length in fig. 7b. We find that the probability of the critical polypeptide length $C_N = 16.3 \pm 2.1$ amino acids shows a significant portion of alpha helix without self-folding. This critical length identified here is slightly larger than the value predicted by the theoretical model. It is also shown that very short peptides ($N = 5$ and 6) have $P_{\rm meta} > 0$. We check the simulation trajectories and confirm that those values are caused by the limits of applying alpha helix similarity (eq. (9)) to very short peptide. The hydrogen bonds in these peptides are not stable from the beginning of the simulation and $P_{\rm meta} > 0$ simply reflects the fact that the peptide backbone during free fluctuations visits many conformations similar to $(\phi^{\rm ref}, \psi^{\rm ref})$, albeit these conformations are not stable. Nevertheless, the direct computational results suggest that the existence of a critical length at which the alpha helix stability is maximized. The difference between the model and the simulation results can be explained by the hydrophobic effect of polyalanine chain. This effect (as given by $E_{\rm hyr(ALA)} = -2.9\,$kcal/mol for a polyalanine chain as summarized in table 2) is much stronger than that of the statistic average, and also by considering that the stronger $E_{\rm hyr}$ leads to a favoring of longer alpha helices as summarized in fig. 6c. We explain this by the fact that a pure polyalanine chain should have a longer C_N than the statistical value. We also find that the actual probability of short polypeptide chains to form a stable alpha helix geometry is smaller than predicted by the theoretical model, indicating that the specific choice of alanine side chains may make the length prevalence more pronounced. It is noted that the prevalent length range given by the simulation result is narrower than the theoretical result because the atomic simulations are performed only for polyalanine, and they do not cover other sequences. There is evidence given by data in the Protein Data Bank as that for continuous polyalanine segments with medium length (> 6 amino acids) only one single length at $N = 11$ gives 13 helix structures. For each of the other lengths only less than two helix structures (one or none) can be found. We anticipate that other protein sequences may lead to different prevalent lengths, making the prevalent length of the alpha helix structures vary in a range and thus broadening the peak.

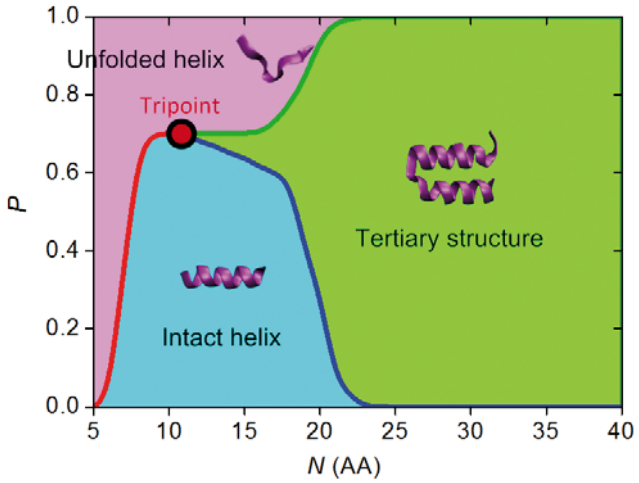


Fig. 8. Schematic of the conformation diagram of alpha helix structures. Three dominant conformations including unfolded helix, intact alpha helix and tertiary structure are depicted by three regions with different colors within the figure. The probability of forming the specific structure for a peptide is measured by the height of the specific region at the corresponding peptide length N. The exact probability of each mode is given in fig. 5. The tripoint at the intersection point of the three regions is highlighted with a circle. It yields a favorite alpha helix length as well as the transition point for higher-order structures, and it also corresponds to the co-existing area with all the three conformations possible. For peptides shorter than the length corresponding to the tripoint, the increasing length increases the helix stability, while for longer peptides, the increasing length decreases the helix stability because of the possibility to form tertiary structures.

3.3 A perspective from length dependent alpha helix stability

Our finding that there exists a critical length for maximum alpha helix stability agrees well with the statistical distribution of the number of alpha helices observed in natural protein structures as a function of the peptide length, as shown in fig. 4. In future work the model could be adopted to study the stability of π-helices and 3_{10} helices by altering parameter values to adapt to the geometries and energy terms of those helix types [40]. In preliminary calculations our model suggests that there exist larger critical lengths for π-helices (18 amino acids by taking $t = 4.1$, $D = 3.6\,\text{Å}$ and $d = 1.3\,\text{Å}$) and smaller critical lengths for 3_{10} helices (10 amino acids by taking $t = 3$, $D = 2.7\,\text{Å}$ and $d = 1.8\,\text{Å}$) than that of the alpha helix (13 amino acids). This result suggests that long chains have an intrinsic preference for π-helical structures while short chains have the intrinsic preference for 3_{10}-helical structures. This suggestion is supported by arecent simulation study of a long helical chains composed one hundred amino acids. In that study it was reported that an applied tensile force induces a transition from an alpha helix to a π-helix at an early deforming stage, and that the content of the 3_{10}-helix in this system under varied force conditions is always low [41]. We are also aware that in order to obtain the critical length of any specific alpha helix, the effect of the amino acid sequence needs to be included for example by applying inhomogeneous weak interactions along the polypeptide. As it is suggested by our model (as shown in fig. 6c) that a stronger hydrophobicity caused by side chains may lead to a longer alpha helix with greater stability.

The critical length and probability distribution discovered here have important implications in our understanding of the folding mechanism and stability of alpha helix protein material. The new insight uncovered here is that the characteristic length scale of alpha helices could be driven by more fundamental principles than the specific amino acid sequence, hydrogen bond energy, or solvent properties. Our results, as illustrated by fig. 8, show that the favored alpha helix length —given by the tripoint of the three possible conformation regions— is governed by the thermal equilibrium between entropy-driven unfolding and free-energy-driven folding in forming secondary and higher-order structures. Polypeptides longer than this length do indeed form alpha helices, but also tend to self-fold to form higher-order structures. In contrast, polypeptides shorter than the critical length are less stable and form random coils. This result may explain why there are no single long alpha helices found in nature, but that they are almost exclusively found in the forms of coiled-coils, triple helices, and even higher-order helical structures [5, 42,43]. Moreover, our results show how the favored alpha helix length is altered by changes to energy terms that define the molecular interactions. This knowledge may potentially help us to design simplified control models that may aid researchers in manipulating molecular and cellular processes [44].

We acknowledge support from AFOSR-YIP, NSF and a DOD-MURI. We appreciate support from the MIT UROP Program. Helpful discussions with the authors of the PLUMED simulation package (Prof. Bonomi and Prof. Parrinello) are much appreciated.

References

1. J.C. Kendrew, R.E. Dickerson, B.E. Strandberg, R.G. Hart, D.R. Davies, D.C. Phillips, V.C. Shore, Nature **185**, 422 (1960).
2. T. Ackbarow, X. Chen, S. Keten, M.J. Buehler, Proc. Natl. Acad. Sci. U.S.A. **104**, 16410 (2007).
3. L. Pauling, R.B. Corey, H.R. Branson, Proc. Natl. Acad. Sci. U.S.A. **37**, 205 (1951).
4. Z. Qin, M.J. Buehler, Phys. Rev. Lett. **104**, 198304 (2010).
5. Z. Qin, S. Cranford, T. Ackbarow, M.J. Buehler, Int. J. Appl. Mech. **1**, 85 (2009).
6. M.J. Buehler, Y.C. Yung, Nat. Mater. **8**, 175 (2009).
7. S.B. Prusiner, Science **216**, 136 (1982).
8. M.A. DePristo, D.M. Weinreich, D.L. Hartl, Nat. Rev. Genet. **6**, 678 (2005).
9. K.M. Pan, M. Baldwin, J. Nguyen, M. Gasset, A. Serban, D. Groth, I. Mehlhorn, Z.W. Huang, R.J. Fletterick, F.E. Cohen, S.B. Prusiner, Proc. Nat. Acad. Sci. U.S.A. **90**, 10962 (1993).

10. S. Kumar, M. Bansal, Biophys. J. **75**, 1935 (1998).
11. S. Penel, R.G. Morrison, R.J. Mortishire-Smith, A.J. Doig, J. Mol. Biol. **293**, 1211 (1999).
12. W. Kabsch, C. Sander, Biopolymers **22**, 2577 (1983).
13. H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, Nucl. Acids Res. **28**, 235 (2000).
14. A. Sali, E. Shakhnovich, M. Karplus, Nature **369**, 248 (1994).
15. P.A. Thompson, W.A. Eaton, J. Hofrichter, Biochemistry **36**, 9200 (1997).
16. V. Munoz, L. Serrano, Nat. Struct. Biol. **1**, 399 (1994).
17. U.R. Doshi, V. Munoz, J. Phys. Chem. B **108**, 8497 (2004).
18. V. Munoz, R. Ramanathan, Proc. Natl. Acad. Sci. U.S.A. **106**, 1299 (2009).
19. S. Lifson, A. Roig, J. Chem. Phys. **34**, 1963 (1961).
20. A. Vitalis, A. Caflisch, J. Chem. Theor. Comput. **8**, 363 (2012).
21. J. Bertaud, J. Hester, D.D. Jimenez, M.J. Buehler, J. Phys.: Condens. Matter **22**, 035102 (2010).
22. B.H. Zimm, J.K. Bragg, J. Chem. Phys. **31**, 526 (1959).
23. E. Shakhnovich, Chem. Rev. **106**, 1559 (2006).
24. K. Ghosh, K.A. Dill, J. Am. Chem. Soc. **131**, 2306 (2009).
25. R. Burioni, D. Cassi, F. Cecconi, A. Vulpiani, Proteins **55**, 529 (2004).
26. M. de Leeuw, S. Reuveni, J. Klafter, R. Granek, PLoS ONE **4**, e7296 (2009).
27. D.U. Ferreiro, A.M. Walczak, E.A. Komives, P.G. Wolynes, Plos Comput. Biol. **4**, e1000070 (2008).
28. M. Karplus, Y.Q. Zhou, D. Vitkup, J. Mol. Biol. **285**, 1371 (1999).
29. L.D. Landau, E.M. Lifshitz, L.P. Pitaevskii, *Statistical Physics*, 3d revision and English edition (Pergamon Press, Oxford, New York, 1980).
30. T. Lazaridis, M. Karplus, Proteins-Struct. Funct. Genet. **35**, 133 (1999).
31. M. Bertz, M. Wilmanns, M. Rief, Proc. Natl. Acad. Sci. U.S.A. **106**, 13307 (2009).
32. C.-I. Brändén, J. Tooze, *Introduction to Protein Structure*, 2nd edition (Garland Pub., New York, 1999).
33. J.C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R.D. Skeel, L. Kale, K. Schulten, J. Comput. Chem. **26**, 1781 (2005).
34. A. Barducci, G. Bussi, M. Parrinello, Phys. Rev. Lett. **100**, 020603 (2008).
35. J.F. Marko, E.D. Siggia, Macromolecules **28**, 8759 (1995).
36. J.R.C.v.d. Maarel, *Introduction to Biopolymer Physics* (World Scientific, Hackensack, NJ, 2008).
37. J. Bertaud, Z. Qin, M.J. Buehler, J. Strain Anal. Engin. Design **44**, 517 (2009).
38. S. Keten, M.J. Buehler, Phys. Rev. Lett. **100**, 198301 (2008).
39. M. Bonomi, D. Branduardi, G. Bussi, C. Camilloni, D. Provasi, P. Raiteri, D. Donadio, F. Marinelli, F. Pietrucci, R.A. Broglia, M. Parrinello, Comput. Phys. Commun. **180**, 1961 (2009).
40. D.J. Barlow, J.M. Thornton, J. Mol. Biol. **201**, 601 (1988).
41. S.M. Kreuzer, R. Elber, T.J. Moon, J. Phys. Chem. B **116**, 8662 (2012).
42. P. Palencar, T. Bleha, Macromol. Theory Simul. **19**, 488 (2010).
43. M.J. Buehler, Nature Nanotechnol. **5**, 172 (2010).
44. P.R. LeDuc, W.C. Messner, J.P. Wikswo, Annu. Rev. Biomed. Engin. **13**, 369 (2011).