Supplementary Information

Is it possible for short peptide composed of positively- and negatively-charged "hydrophilic" amino acid residue-clusters to form metastable "hydrophobic" packing ?

Hiroshi Nishigami[1], Jiyoung Kang[1,2], Ryu-ichiro Terada[1], Hiori Kino[3], Kazuhiko Yamasaki[4], and Masaru Tateno[1]*

[1]Graduate School of Life Science, University of Hyogo, 3-2-1 Kouto, Kamigori, Hyogo 678-1297, Japan

[2] Center for Systems and Translational Brain Sciences, Institute of Human Complexity and Systems Science, System Science Center for Brain and Cognition, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul 03722, Republic of Korea.

[3] National Institute for Material Science, 1-2-1 Sengen, Tsukuba, Ibaraki 305-0047, Japan

[4] Biomedical Research Institute, National Institute of Advanced Industrial Science and Technology (AIST), 1-1-1 Higashi, Tsukuba, Ibaraki 305-8566, Japan

*To whom correspondence should be addressed.

E-mail: tateno1611@gmail.com

Fax: (+81)-791-58-0347

Phone number: (+81)-791-58-0347

Contents

S1. Nomenclature employed in the present report (Table S1) and preliminary structural analysis of SDR peptide

Table S1 provides a reference of the nomenclature employed in the previous[1] and present reports. For example, the supercoiled DNA-recognition domain was denoted by SRD in the previous report,[1] while in the present report, it is denoted by SDR domain.

Table S1. Reference of the nomenclature used in the previous and present reports.

	Previous Report	Present Report	Definition
Supercoiled DNA-recognition domain	SRD	SDR domain	Domain (residues 200-336 of LEDGF) recognizing supercoiled DNA
Supercoiled DNA-recognition peptide	$K_9E_9K_9$	SDR peptide	Peptide that preferentially binds to supercoiled DNA
$Lys_9Glu_9Lys_9$ peptide	$K_9E_9K_9$ peptide	$Lys_9Glu_9Lys_9$ peptide	Peptide with the amino acid sequence of $(Lys)_9(Glu)_9(Lys)_9$

Preliminary structural analysis of SDR peptide by NMR spectroscopy

As shown in Figure S1, the chemical shift values of the peaks in the ^{1}H-NMR spectrum of the SDR peptide were almost identical to those of Glu or Lys protons in the random-coil state[2]. This indicates that the SDR peptide basically adopts an unfolded structure, under the NMR measurement conditions.

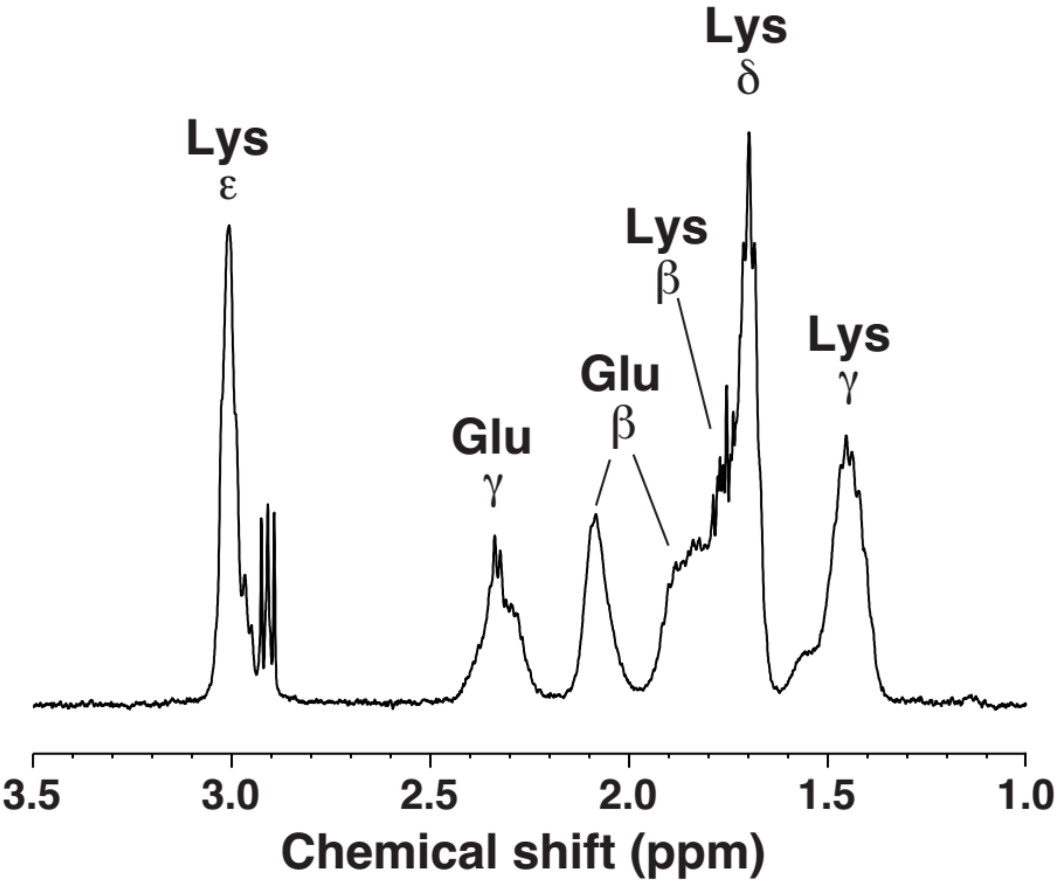

Figure S1. ^{1}H-NMR spectrum of the SDR peptide at the frequency of 500 MHz. The peaks were assigned on the basis of the chemical shifts of random-coil peptides[2].

Functional role of the SDR peptide

Interestingly, the $Lys_9Glu_9Lys_9$ peptide was previously revealed to selectively bind to supercoiled DNA. This was achieved because its amino acid sequence was artificially designed based on that of the DNA-binding domain of lens epithelium-derived growth factor (LEDGF) (also known as SBP75/p75), which was originally identified as a factor essential for selective binding to supercoiled DNA.[1, 3] Thus, in this report, we refer to this DNA-binding domain and the $Lys_9Glu_9Lys_9$ peptide as the supercoiled DNA-recognition (SDR) domain and the SDR peptide, respectively (for details on the terminology, see Supplementary Table S1). LEDGF tethers numerous partner proteins to chromatin, through which many types of transcriptional regulation occur. One of the partners is the integrase of human immunodeficiency virus type 1 (HIV-1), which catalyzes the insertion of the HIV-1 cDNA into the host cell genome, and thus LEDGF plays a pivotal role in triggering HIV-1 infection.[4]

Previous analyses of effects of charged amino acid residues to folding of peptides

The formation of hydrophobic packing is a critical factor in the folding of most proteins.[5-7] The Trp-cage is an example of such a structure in peptides, and its folding process has undergone intensive investigations, which revealed that a hydrophobic core is formed by some hydrophobic amino acid residues, and both short- and long-range inter-residue interactions participate in the core packing.[8-10] Conversely, high contents of charged amino acid residues tend to induce unfolding at neutral pH. In fact, poly-glutamic acid (PGA) and poly-lysine (PL) are unfolded at neutral pH, because of the electrostatic repulsions of their charged side chains.[11-13] PGA and PL fold into α-helices under lower and higher pH conditions, respectively, where their side chains are uncharged and stabilized via hydrophobic interactions.[11-13]

When both of positively and negatively charged residues are present in amino acid sequences, the electrostatic interactions can stabilize proteins.[14-16] For example, the previous studies have revealed that α-helix can be stabilized by a single pair of Glu and Lys residues placed in a helical peptide when the spacing between the residues in amino acid sequence is close to the helical repeat of 3.6 residues per turn; i.e., (i, i+3) or (i, i+4).[14] However, the Glu-Lys spacing of (i, i+1) or (i, i+2) was found to destabilize the α-helix in the peptides.[14]

S2. Validation of the REMD simulation

A productive REMD simulation was performed, and generated the trajectories of the total simulation time of 24.32 μs (190 ns × 128 replicas) with explicit solvent water molecules. Numerous replicas (128 replicas) were derived from the fully-solvated system, which was quite large (166,981 atoms) as compared with those of the previous studies that also employed REMD simulations.[17, 18] The number of replicas was sufficient to connect each neighboring replica with the closest temperature (the temperature range was 300.00 K to 406.66 K). In fact, the potential energy distributions significantly overlapped between the neighboring replicas, thereby giving an exchange rate of 0.34 (Figure S2(a)). During the 190 ns simulation time, the temperature of each replica increased and decreased from the lowest to the highest temperatures several times, and thus we successfully obtained a well-converged structural ensemble at 300.00 K.

To confirm the convergence of the present REMD simulation, we calculated the secondary structure profiles of the SDR peptide at 300 K (see below) by employing two distinct datasets, corresponding to the trajectories from 0 to 11.52 μs and from 11.52 to 24.32 μs (Figure S2(b)). These two profiles were almost identical, indicating that the resultant ensemble was converged.

In the present REMD simulation, 128 replicas were employed within a temperature range of 300.00 to 406.66 K. The temperatures were set as follows: 300.00, 300.74, 301.48, 302.22, 302.97, 303.71, 304.46, 305.20, 305.95, 306.70, 307.46, 308.21, 308.97, 309.72, 310.48, 311.24, 312.00, 312.76, 313.53, 314.30, 315.07, 315.84, 316.61, 317.38, 318.15, 318.92, 319.70, 320.48, 321.26, 322.02, 322.81, 323.59, 324.38, 325.16, 325.95, 326.74, 327.53, 328.33, 329.12, 329.91, 330.71, 331.51, 332.31, 333.11, 333.92, 334.73, 335.54, 336.35, 337.16, 337.97, 338.79, 339.60, 340.42, 341.24, 342.06, 342.88, 343.71, 344.53, 345.36, 346.19, 347.02, 347.85, 348.69, 349.52, 350.36, 351.20, 352.04, 352.88, 353.73, 354.57, 355.42, 356.27, 357.12, 357.97, 358.83, 359.69, 360.54, 361.40, 362.26, 363.13, 363.99, 364.86, 365.73, 366.60, 367.47, 368.34, 369.22, 370.10, 370.97, 371.85, 372.74, 373.62, 374.51, 375.39, 376.28, 377.17, 378.07, 378.96, 379.86, 380.75, 381.65, 382.56, 383.46, 384.36, 385.27, 386.18, 387.09, 388.01, 388.92, 389.84, 390.75, 391.67, 392.60, 393.53, 394.45, 395.38, 396.31, 397.24, 398.17, 399.11, 400.05, 400.98, 401.93, 402.87, 403.81, 404.76, 405.71, and 406.66 K.

In previous studies, similar temperature ranges were also adopted for REMD calculations, even for larger proteins.[19-23] Thus, in order to obtain a structural ensemble of the SDR peptide, we emzployed the standard methodology that is shared by researchers in the field. The distributions of potential energy of all of the replicas are shown in Figure S2(a). To confirm the convergence of the present REMD simulation, we calculated the secondary structure profiles of the SDR peptide at 300 K (see the text) by employing two distinct datasets, corresponding to the trajectories from 0 to 11.52 μs and from 11.52 to 24.32 μs (Figure S2(b)).

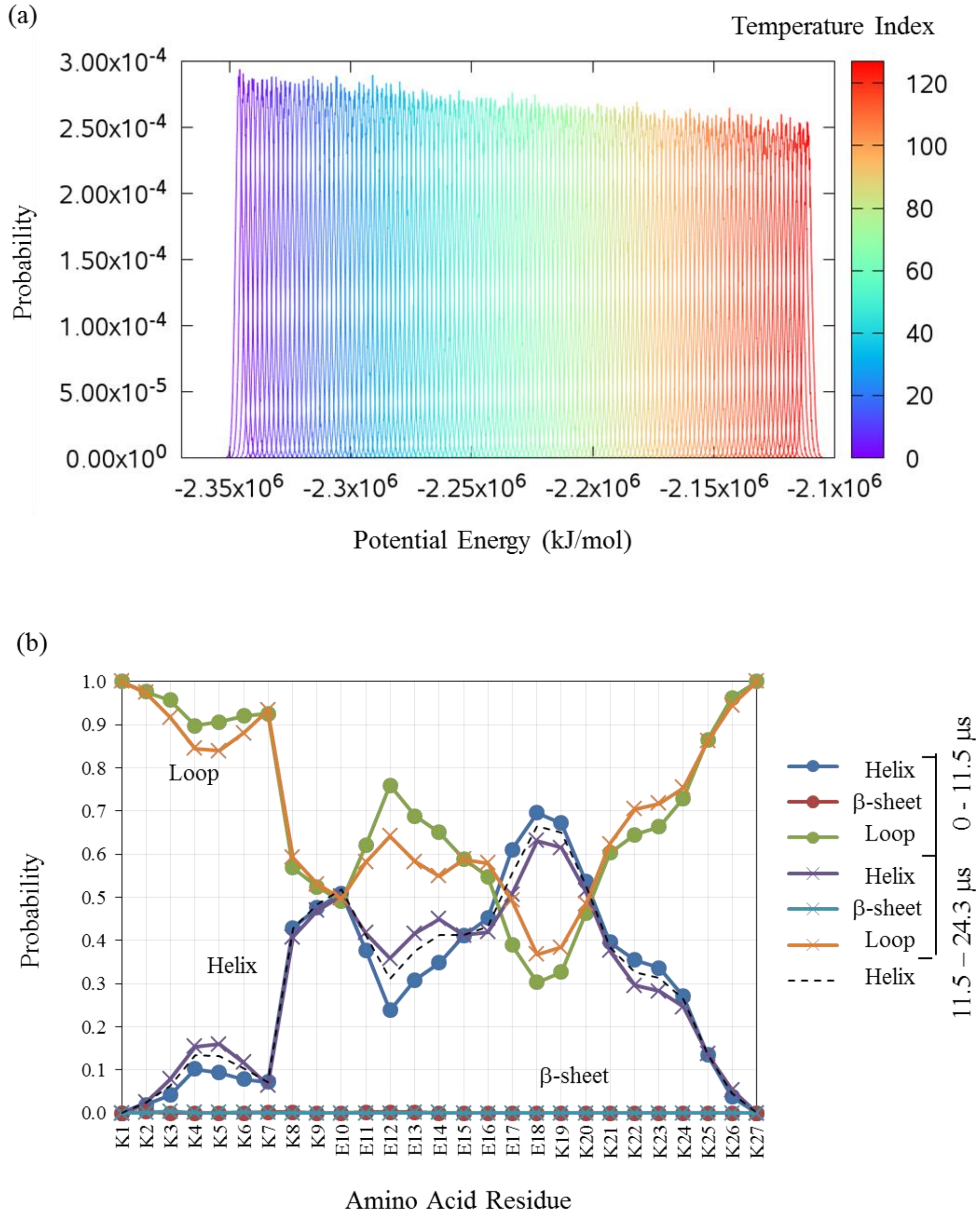

Figure S2. (a) Probability distributions in terms of potential energy, involving all replicas at 128 distinct temperatures (300.00-406.66 K) in the REMD simulation. The lines are colored by the indices of the temperatures, as shown by the color bar in the figure. On average, the MD simulation for 20 ns consumed 5 hours with 32 nodes (total 512 cores) of SGI ICE X (64 Intel Xeon E5-2670

CPU), or 10 hours with 2 nodes (total 48 cores and 4 × 2880 CUDA cores) of SGI ICE XA (×4 Intel Xeon E5-2680v3 CPU and ×4 NVIDIA Tesla K40). (b) Secondary structure profiles of the structural ensemble at 300 K constituted by employing distinct datasets 1 (from 0 to 11.5 μs), 2 (from 11.5 to 24.3 μs), and the merged dataset (from 0 to 24.3 μs) (for the latter, only the profile in terms of the helix type structures is depicted by the dashed line. This profile is identical to that in Figure 1). Comparison of the profiles revealed that the three datasets provided almost equivalent profiles, thus indicating the convergence of the present REMD simulation.

Evaluation of protonation states of Glu residues in the SDR peptide

It is well known that 99.9% of carboxylates of Glu residues are deprotonated (*i.e.*, negatively charged) in solution at pH 7. In fact, pKa for Glu is 4.07,[24] and thus the protonation of the Glu residue is an extremely rare event. Accordingly, we adopted positively- and negatively-charged side chains of Lys and Glu residues, respectively.

The followings are additional discussions relevant to the protonation of Glu residue: In S6 ribosomal protein, sharing of a proton between nearby carboxylates on the protein surface was identified (the Laser Induced Proton Pulse technique was employed) (the proton transfer was also found to be involved in this process). However, the probability for the formation of such configurations (where nearby carboxylates share a proton) is low.[25]

In addition, the crystal structure of S6 ribosomal protein (PDB accession code: 1RIS) definitely shows that the aforementioned proton sharing is independent of the protein folding. Actually, the relevant side chains (*i.e.*, Glu22 and Asp83) are completely dissociated (and thus exposed as their charged states) in the crystal structure. Nevertheless, the proton sharing by these carboxylates occurs as a rare event, just for the proton transfer as experimentally shown in the cited paper. This is due to the aforementioned reason (*i.e.*, 99.9% of carboxylates of Glu residues are negatively charged in solution at pH 7).

In another study, the pKa shift was observed for a Glu residue, which was actually introduced by the mutagenesis technique into a stable protein, to examine the capability of globular proteins to tolerate the presence of buried charges.[26] This analysis showed that the protonation of the Glu side chain occurs just for a Glu residue that was buried in the hydrophobic core of a very stable protein (*i.e.*, a highly stable form of staphylococcal nuclease). Notably, ionizable groups buried in the hydrophobic interior of proteins are essential for the catalysis and energy transduction mediated by other proteins (such as ATPase, cytochrome c oxidase, and bacteriorhodopsin), and thus the protonation of the Glu residue occurs just only when the Glu residue works as a functional residue.

In this manner, the protonation of Glu residues occurs “tentatively” together with the functional processes such as proton transfer, catalytic reaction, and energy transduction. We do not know any reports showing (suggesting) that Glu clusters in any (metastable) peptides are protonated even at pH 7. Similarly, for the SDR peptide, the protonation of the Glu side chains is a markedly rare event. In fact, the structural data obtained by our MD simulations was comparable with our CD spectroscopy measurement data, which is an evidence to show that the protonation states of the SDR peptide in our MD simulations were appropriate (also see Results and Discussion).

Effects of ionic strength on the conformations of the SDR peptide

As is well known, ionic strength can modulate the electrostatic interactions. For example, in a single pair of Glu and Lys residues found in a helical peptide where those residues are located as (*i*, *i*+3) or (*i*, *i*+4) in the amino acid sequence, a side-chain interaction between the Lys and Glu residues is stabilized and destabilized at low (10 mM NaCl) and high (2,500 mM NaCl) ionic strength, respectively.[14] This shows that such extremely higher concentrations of NaCl (*e.g.* 2,500 mM NaCl) destabilize the electrostatic interactions in the peptide. Conversely, the physiological ranges of salt concentrations marginally influence the electrostatic interactions in peptides.

Furthermore, the helical contents of the peptide where a single pair of Glu and Lys residues was included with the spacing of (*i*, *i*+4) were measured at pH 7, and found to be 39%, 40%, and 37% at NaCl concentrations of 10, 1,000, and 2,500 mM, respectively.[14] Moreover, varying ionic strength did not induce conformational changes in the helical or random-coil states of poly-glutamic acid, whereas pH values exclusively modulated the helical and random-coil conformations.[11] Thus, these previous reports clearly indicated that the effect of physiological ionic strength to the peptide conformations was marginal, and also the employed ionic concentration is appropriate for the current analysis.

Dependence of structural properties of IDPs on force field parameters

Although the force field parameters have recently been updated, structural propensities of disordered peptides observed in MD simulations have been shown to be dependent on force fields.[27-33] In addition, the combinations of force fields and water models were also shown to be crucial to preserve consistencies with experimental data.[34] Thus, for IDPs, appropriate force fields are also depending on the systems analyzed currently.

Accordingly, in the present study, we employed the combination of AMBER ff99SB and TIP3P water model, because these have been most tested for various biological macromolecular systems for many years. Actually, in a previous study, REMD and (total) 70 μs conventional MD simulations of a disordered protein (*i.e.*, 37-residue human islet amyloid polypeptide; hIAPP) were performed with this combination, and were employed as the reference of statistical modeling of a structural ensemble.[35] Moreover, as a validation of our theoretical results, we compared our computational data with the experimental CD measurements, and thereby showed that the (total) helical propensities were found to be comparable in both data (see Results and Discussion). Thus, our computational analysis was indicated to be reasonable.

By contrast, in a previous study, the stability of helical structures in a (helical) peptide, Ac-$(AAQAA)_3$-NH_2, was underestimated by the combination of the AMBER ff99SB force field and TIP3P water, and that of the AMBER ff03* force field (revised parameters) and TIP3P water was most consistent with the experimental data.[36] In another case, the CHARMM 22* force field (revised parameters) and the charmm-modified TIP3P agrees best with all of available experimental data.[34] These analyses show that the appropriate force fields are currently depending on the systems. So, despite the good agreement of our MD simulation with the experimental data, a bias that may be caused on details of the data by the force field (and the combination with the water model) should also be considered in the present study. Still, it should be noted here that our aim of the present study is to obtain the substantial structural data of the SDR peptide that are independent of minute details relevant to the methodologies.

S3. Distribution of each type of secondary structure identified by employing the DSSP algorithm

The β-sheet contents were almost zero in all of the amino acid residues of the SDR peptide, and thus the total ratio was 0.05%. Notably, we obtained asymmetric secondary structure profiles, although the amino acid sequence of the SDR peptide was completely symmetric (this asymmetry may be due to the differences in the effects of the N- and C-termini; see the text).

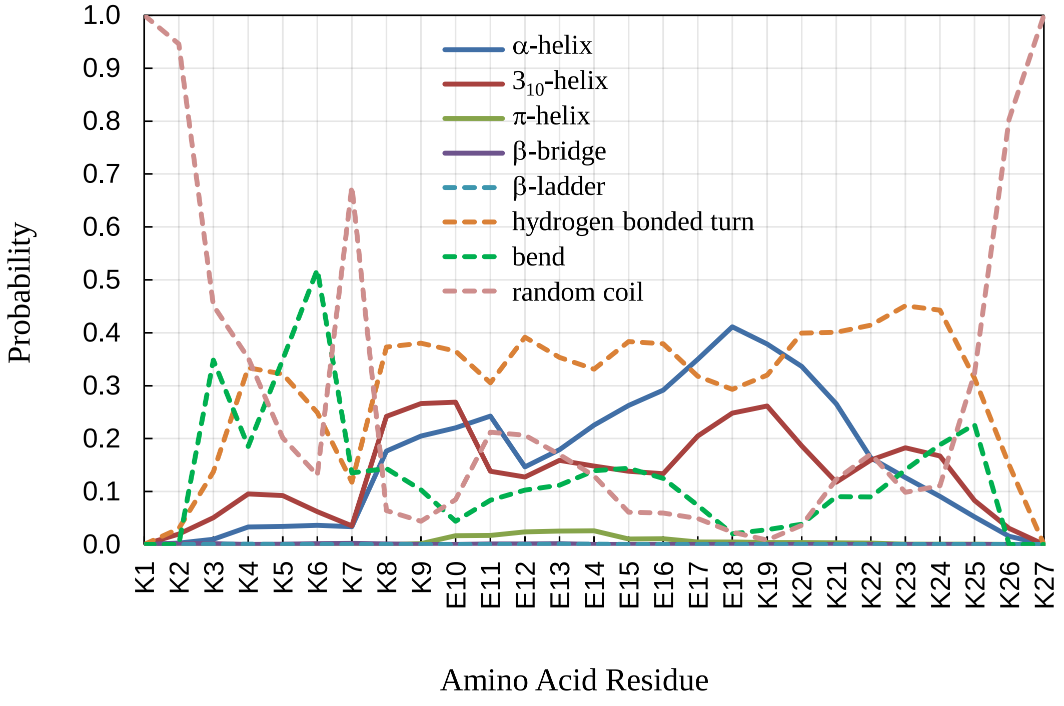

Figure S3. Distribution of each type of secondary structure identified by employing the DSSP algorithm. Each line corresponds to each secondary structure type, as follows: α-helix (blue line), 3_{10}-helix (red line), π-helix (green line), β-bridge (purple line), β-ladder (dashed cyan line), hydrogen bonded turn (dashed orange line), bend (dashed green line), and random coil, where no secondary structures were assigned (dashed pale red line).

S4. Free energy landscapes in N_{helix}, R_g, and ASA spaces

One-dimensional (1D) free energy landscapes in N_{helix}, R_g, and ASA spaces are shown in Figure S4. In the N_{helix} space, three distinct minima were found in the free energy landscape (Figure S4(a)). The lowest free energy conformation was identified at 7 of N_{helix} (this state was used as the standard of the free energy value, as exactly 0 kcal/mol), which corresponds to almost 2-turns of one contiguous helix type structure or two distinct 1-turns of helix type structures, such as a combined α-helix and 3_{10}-helix (for both types, the detailed structural features are described below). The second most stable state was found at 4 of N_{helix} (for which the free energy value was 0.25 kcal/mol larger than the most stable conformation), and thus the structure corresponded to a 1-turn α-helix. The third most stable state was at 11 of N_{helix} (0.26 kcal/mol), which involved an almost contiguous 3-turn α-helix. The free energy values increased gradually toward over ~11 of N_{helix}, and so the probabilities of the conformations with a longer contiguous α-helix decreased.

The conformations that lack any helical structures (0 of N_{helix}) were relatively lower in the free energy landscape in the N_{helix} space (~1 kcal/mol) (Figure S4(a)). This means that the experimental determination of the 3D structure of the SDR peptide in solution is difficult, since the metastable structures are mixed with the random coil state.

For the R_g space (Figure S4(b)), the minimum of the free energy exhibited a deep and wide basin at 1.05 nm of R_g, which means that most of the SDR peptide molecules existing within this free energy subspace were folded into compact structures. For the ASA space, the free energy landscape was almost identical to that of the R_g space, as the most stable state was located at 33.0 nm^2 of ASA, which also corresponded to the compact structures of the SDR peptide molecules. The free energy difference between the folded and unfolded structures was ~4.5 kcal/mol in the free energy landscapes in the R_g and ASA spaces.

The free energy landscapes were also constituted in two- and three-dimensional (2D and 3D) spaces, with respect to N_{helix}, R_g, and/or ASA (Figure S5)). In the 2D free energy landscape, the calculated correlation coefficient between R_g and ASA was 0.70 (Figure S5(c)). In contrast, for the parameters of N_{helix} and R_g, and N_{helix} and ASA (Figure S5(a) and (b)), no significant correlations were found (the correlation coefficients were 0.24 and 0.05, respectively). Thus, these are the independent parameter sets (*i.e.*, N_{helix} and R_g, and N_{helix} and ASA).

In the 2D free energy landscapes in terms of R_g and ASA (Figure S5(c)), a single dominant free energy basin existed, and the most stable states were located at 1.02 nm of R_g and 32.4 nm^2 of ASA. The existence of a single free energy minimum in both the R_g and ASA spaces suggests that the major packing mode and its surrounding fluctuated structures are commonly involved in the conformations in the free energy-minimum structural ensemble, which could be relevant to the hydrophobicity (discussed below), and operating in the folding processes of the SDR peptide to achieve the compactness specified by the free energy landscapes in the R_g and ASA spaces. In contrast, the free energy landscape in the N_{helix} space included several distinct minima, thus suggesting that the compact structures involved some distinct conformations, including different lengths of the helix type structures, as shown in the free energy landscape in the N_{helix} space. To reveal the driving factors that induce the compactness of the SDR peptide, this crucial issue will be discussed further (see the text).

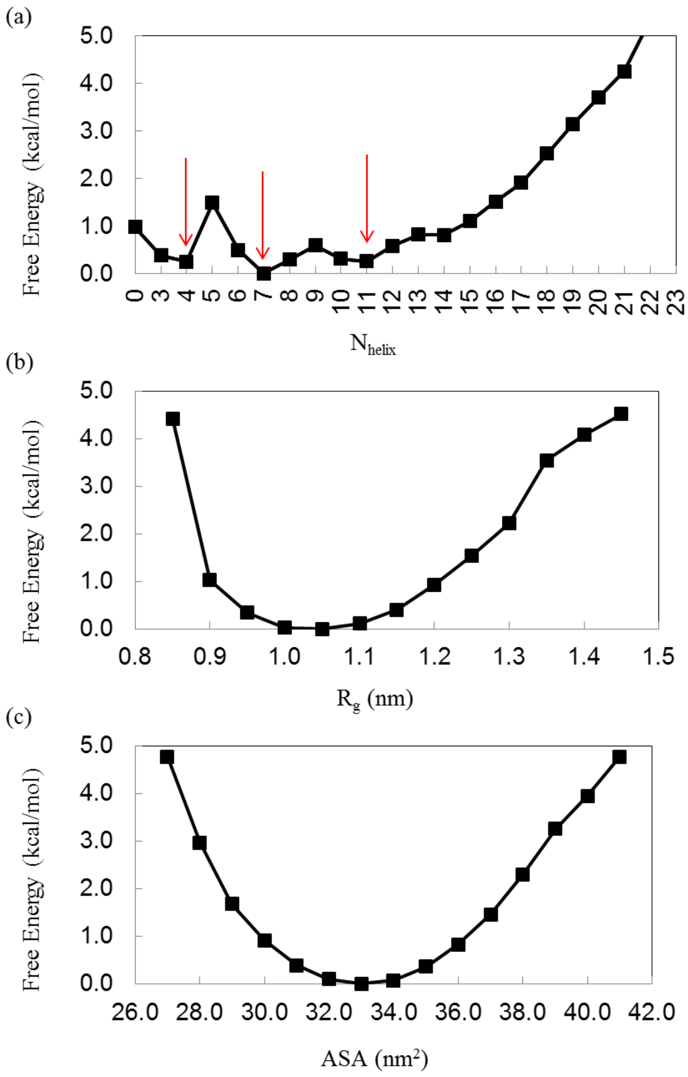

Figure S4. The free energy landscapes (1D types) with respect to (a) the number of helical residues (N_{helix}), (b) the radius of gyration (R_g), and (c) the solvent accessible surface area (ASA) of the SDR peptide. Note that N_{helix} is an integer larger than 3 (the shortest helix type structure is a 3_{10}-helix).

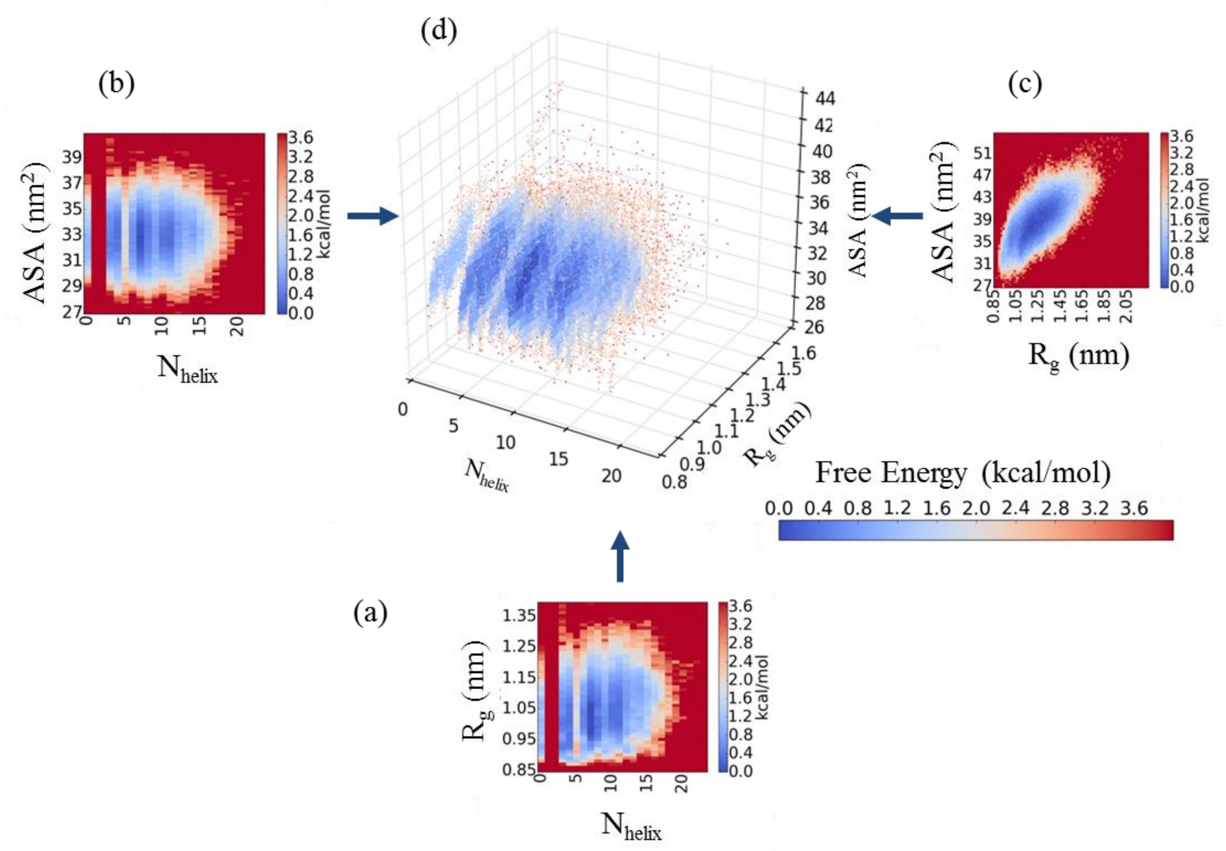

Figure S5. The free energy landscapes (2D and 3D types) with respect to the three parameters employed in Figure S4: N_{helix}, R_g, and ASA. (a-c). The 2D free energy landscapes of (a) N_{helix} and R_g, (b) N_{helix} and ASA, and (c) R_g and ASA. (d) The 3D free energy landscape in the space of N_{helix}, R_g, and ASA. The free energy value is mapped according to the color index. Notably, N_{helix} is either an integer larger than 3 or 0.

S5. Summary of the properties of the representative structures classified by our cluster analysis

The cluster analysis indicated that the conformations with the helix type structures in the C-terminal Glu_9Lys_9 boundary are more favorable, in terms of the free energy, than those in the N-terminal Lys_9Glu_9 boundary (Figure 3(a)). This is consistent with the secondary structure profile of each amino acid residue (Figure 1), where the peak of the helix type structures in the C-terminal Glu_9Lys_9 boundary is higher than that in the N-terminal Lys_9Glu_9 boundary. The C-terminal Glu_9Lys_9 boundary may have preferentially formed the helix type structures because of the negative charge on the carboxyl group of K27 (the C-terminal residue), which facilitated the interactions with the Lys side chains. Thus, the interactions of the K27 residue with the central Glu_9 segment would be hindered. Conversely, the additional positive charge of the N-terminal residue (K1) would preferentially interact with the central Glu_9 segment, thereby forming a bent structure that would hinder the formation of the helix type structure.

Table S2. Summary of the properties of the representative structures classified by our cluster analysis (see Figure 3). This table shows the N_{helix} values and the lengths of $Helix^{NB}$ and $Helix^{CB}$ in each representative structure, as shown in Figure 3(b-f).

State	N_{helix}	Length of $Helix^{NB}$	Length of $Helix^{CB}$	Representative structure
State 1 (NC-B)	11	4	7	b
	7	4	3	c
State 2 (C-B)	7	0	7	d
	4	0	4	e
State 3 (N-B)	4	4	0	f

S6. Correlations of the ASA values with the long-range and short-range contacts

To evaluate the correlation of the ASA of the SDR peptide with the numbers of long-range or short-range hydrophobic or electrostatic contacts, the free energy landscapes were calculated with respect to those quantities. Here, the criteria of counting the number of long-range and short-range contacts were identical to those described in the text.

As a result, we found weak correlations of the ASA with the numbers of long-range hydrophobic contacts (correlation coefficient: −0.36) and long-range electrostatic contacts (−0.49) (Figure S6(a) and (c)). By contrast, we found no correlations of the ASA with the numbers of short-range hydrophobic contacts (−0.16) and short-range electrostatic contacts (−0.11) (Figure S6(b) and (d)).

These statistical data mean that the long-range electrostatic and hydrophobic interactions contributed to the compactness of the whole structure of the SDR peptide, while the short-range electrostatic and hydrophobic interactions did not contribute to it (but contributed to the local structures, such as helical structures).

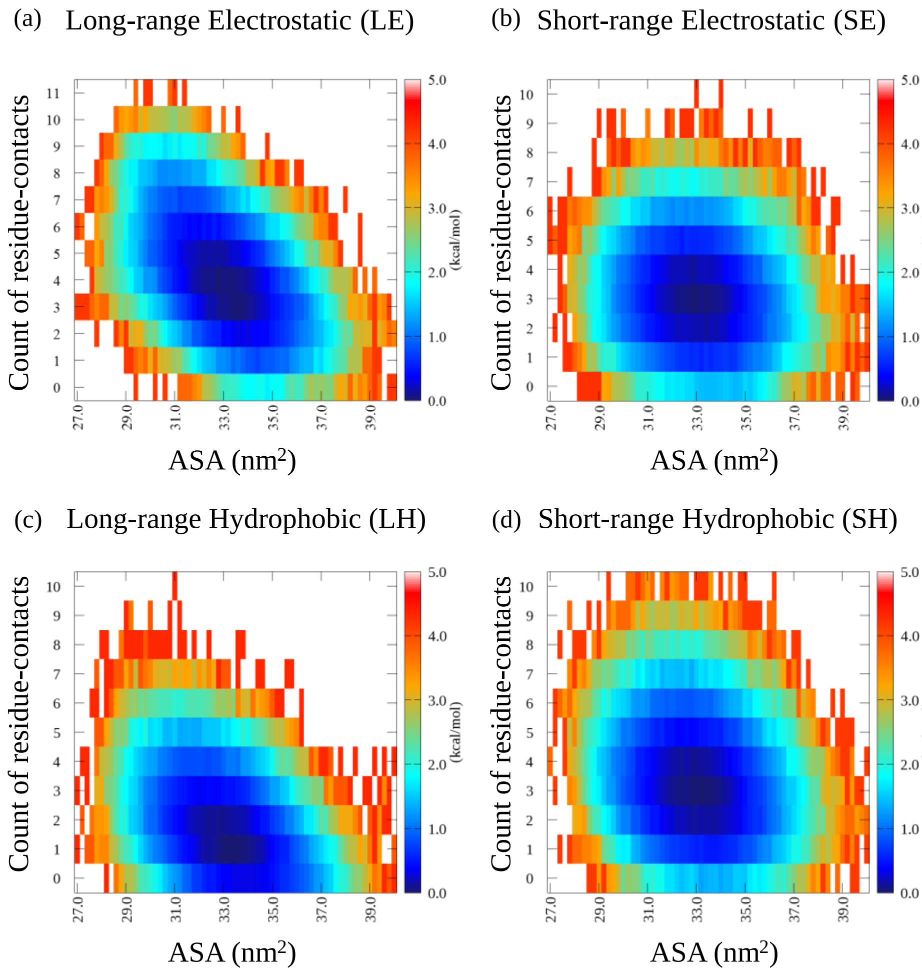

Figure S6. Correlations of the ASA (the horizontal axis) with the following four distinct types of residue-contacts (the vertical axes): (a) long-range electrostatic (LE), (b) short-range electrostatic (SE), (c) long-range hydrophobic (LH), and (d) short-range hydrophobic (SH) contacts. The free energy values were obtained by the conversion of the frequencies of the contacts, in combination with the ASA.

S7. Solvent exclusion by hydrophobic clusters identified in the SDR peptide

In the representative structure of State 1 (Figure 3), hydrophobic clusters were identified (Figures 4 and 5), contributing to the structural stability of the SDR peptide through water exclusion, as shown in Figure S7.

In our additional analysis employing canonical MD simulations, the hybrid hydrophobic and electrostatic contacts were achieved in a concerted manner (data not shown).

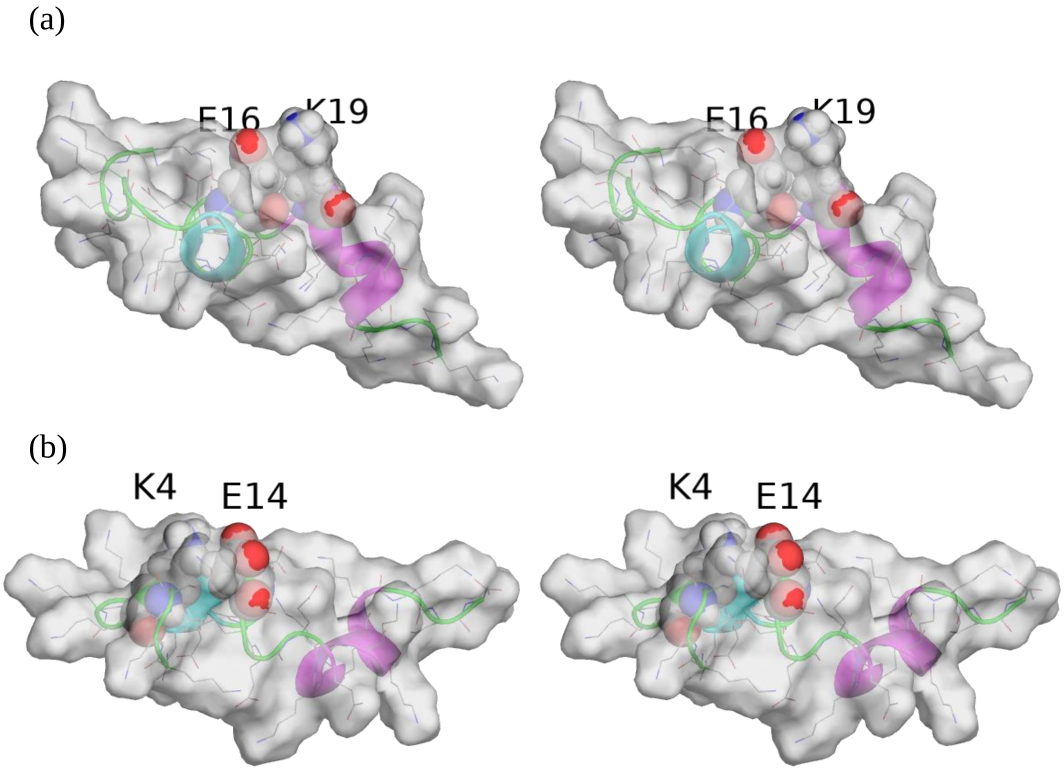

Figure S7. Stereo view of the solvent accessible surface (gray) of the representative structure found in State 1 (Figure 3). For panels (a) and (b), the viewpoints are identical to Figure 5(d) and 5(f), respectively. The hybrid residue contacts that are shown in Figure 5(d) and 5(f) are labeled.

Further Characterization of Hydrophobic Clusters Identified in the SDR peptide

The thermodynamic properties of the SDR peptide (*e.g.*, a mixture of the distinct conformations) are apparently similar to those of some denatured peptides/proteins.[37-39] For example, c-src SH3 exhibited transient helicity under denaturing conditions (pH 2);[39] in particular, for residues 40-46, an apparent helicity of ~45% was observed. This may be induced by the transitions of the charge states of the amino acid residues due to the extremely acidic conditions, as in the cases of poly-glutamic acids (*i.e.*, the charge state transition of poly-glutamic acids at lower pH induces the helical structure, as mentioned in the text).[11] As a result, the helical content of the SDR peptide was "apparently" similar to those of proteins under denaturing conditions.

Conversely, we should separate these two types of helix formation as distinct mechanisms: For some peptides/proteins under denaturing conditions, transient helical structures would be induced through interactions with the denaturant or by the extremely high or low pH. In contrast, the hydrophobic contact clusters identified in our REMD simulation stabilized the whole structure of the SDR peptide, through the formation of the compacted conformations (Figure S7), which is a characteristic feature that is different from those of other peptides/proteins under denaturing conditions.

In this manner, the helical content is insufficient to characterize the peptide. The hydrophobic cluster packing identified in the present study excluded water molecules (Figure S7), and thereby stabilized the whole structure of the peptide. In fact, in the CD spectra of the SDR peptide under the physiological and denaturing conditions, which were regulated through the concentration of the denaturant, significantly different properties were identified (Figure 2). Under the denaturing conditions, the helical trends of the CD spectrum completely disappeared. Thus, the helical content of the SDR peptide was shown to be quite different from the denatured states, which also led to the difference in the conformational ensemble.

S8. Amino acid sequences of helical structures identified as homologues of Lys_4Glu_4 or Glu_4Lys_4 sequences in the Protein Data Bank

In the present study, we identified the α-helical propensity in the boundary between the positively- and negatively-charged hydrophilic amino acid residue segments ($Helix^{NB}$ and $Helix^{CB}$). We further assessed whether this α-helical propensity is a general feature in other proteins, by analyzing the structural data in the Protein Data Bank (PDB). We employed the sequences of KKKKEEEE and EEEEKKKK as probes for the sequence search, with the use of BLASTp[40]. The analysis revealed many helical structures existing in positively- and negatively-charged amino acid residue segments (see Figure S8 for sequences and Figure S9 for three-dimensional structures), thus indicating that the α-helical propensity discovered with the SDR peptide is a general trend in protein structures. A more detailed analysis is described later (see the Supplementary Information, S10).

PDB accession code / Chain ID : Amino acid sequence

```
(a)
2J1D/G: 986 AKQENENMRKKKEEEERRARMEAQLKEQRE 1015
S.S.        HHHHHHHHHTHHHHTHHHHHHHHHHHHHHH

(b)
3WY9/C:  77 EEKKEEEKKEEEEKEEEVSEEEALAGLSAL 106
S.S.        DDDDDDHHHHHHHHHHHHHHHHHHHHHHHH

(c)
5MPS/c: 331 ANPTKYEYLKKKREQEETKQPKIVSIGDLE 360
S.S.        SSSTHHHHHHHHHHHHH DDDDDDDDDDDD

(d)
1G7R/A: 441 YEEWVRGIEEEKKKKWMEAIIKPASIRLIP 470
S.S.        HHHHHHHHHHHHHHHHHHTS   EEEEEEE

(e)
4KVM/A: 591 FEKLSSGEINEEEEKKIYKKLKKDLSKRLE 620
S.S.        HHHHHHT S HHHHHHHHHHHHHHHHHHHH

(f)
5HMO/A: 811 LLAEKRAEEEKRKREEEEKRKREEEERERE 840
S.S.        HHHHHHHHHHHHHHHHHHHHHHHHHHHHHH
```

Figure S8. Amino acid sequences of helical structures identified by BLASTp[40] as homologues of Lys_4Glu_4 or Glu_4Lys_4 sequences in the Protein Data Bank. The identified sequences are shown together with their neighboring sequences, and the homologous parts are colored red. PDB accession codes and chain IDs of the homologues are also shown with the residue numbers. The secondary structure assignments (S.S.) are shown for each amino acid residue (calculated from the experimental coordinates using DSSP); H, B, E, G, I, T, S, and D denote α-helix, residue in isolated β-bridge, extended strand that participates in β-ladder, 3_{10} helix, π helix, hydrogen bonded turn, bend, and regions that have not been experimentally observed, respectively, and a blank stands for a loop or other irregular structure that has been experimentally observed. (a) The C-terminal fragment of human Daam1 (PDB accession code: 2J1D), which includes the sequence KKKEEE.

(b) The C-terminal domain of the archaeal ribosomal stalk protein aP1 complexed with the GDP-bound archaeal elongation factor aEF1alpha (PDB accession code: 3WY9), which includes the sequence EKKEEEEKEEE. (c) The pre-mRNA-splicing factor Slu7 (PDB accession code: 5MPS), which includes the sequence KKKREQEE. (d) The translation initiation factor IF2/EIF5B (PDB accession code: 1G7R), which includes the sequence EEEKKKK. (e) The NatA (Naa10p/Naa15p) amino terminal acetyltransferase complex (PDB accession code: 4KVM), which includes the sequence EEEEKKIYKKLKKD. (f) Myosin X (PDB accession code: 5HMO), which includes the sequence EKRAEEEKRKREEEEKRKREEEERERE.

S9. Three-dimensional structures including the amino acid sequences Lys_4Glu_4 and Glu_4Lys_4 identified in the Protein Data Bank

In the Protein Data Bank, we searched for the three-dimensional (3D) structures with the amino acid sequences that were homologues of Lys_4Glu_4 or Glu_4Lys_4, identified by BLASTp[40] (Figure S9). The analysis revealed that in the identified 3D structures, the Glu and Lys side chains on the helix structures formed hybrid contacts that were similar to those found in the SDR peptide as the stable conformations. Notably, all of the helical structures shown in Figure S9 were exposed to the solvent, whereas in the SDR peptide, the helical structures were involved in the packing of the peptide.

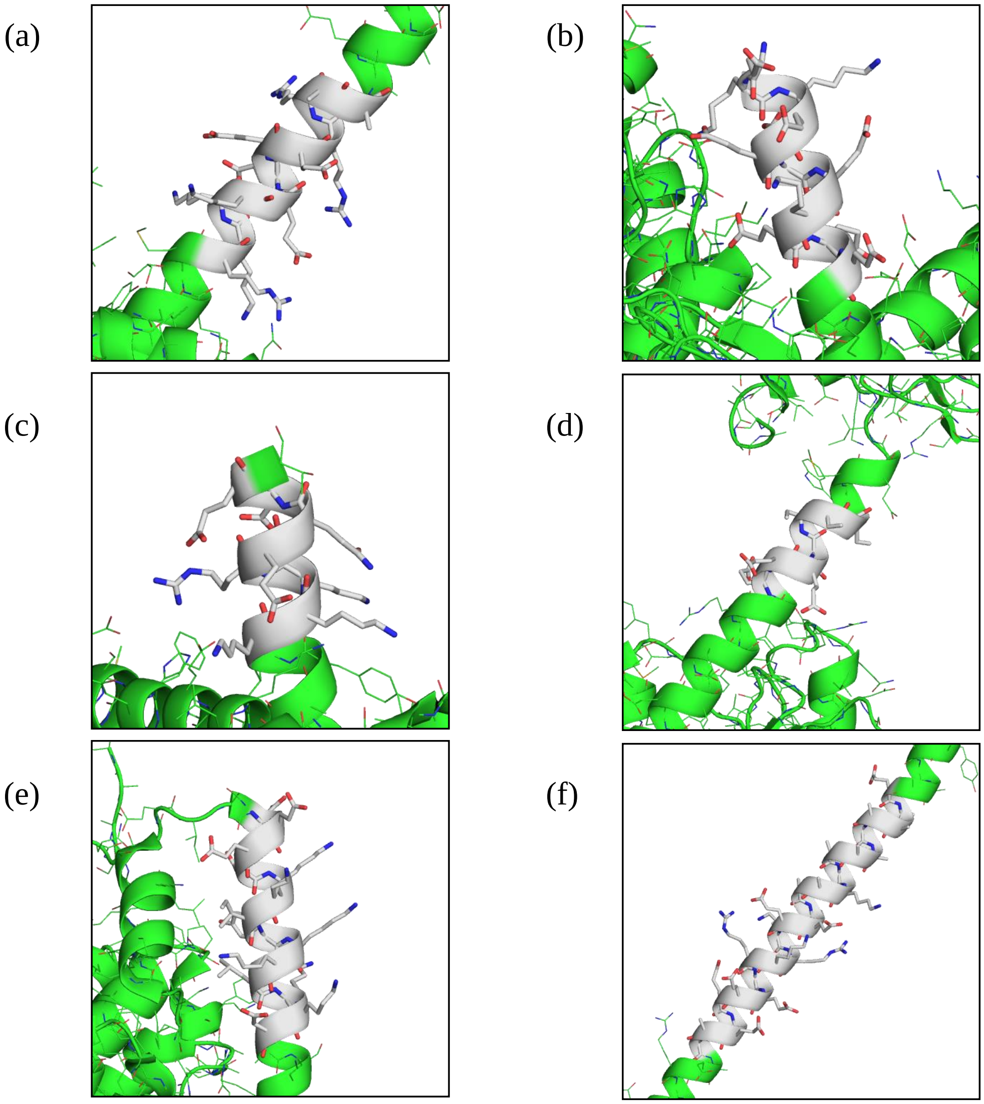

Figure S9. Three-dimensional structures including the amino acid sequences K_4E_4 and E_4K_4, identified in the Protein Data Bank. Cα atoms in the homologous regions are colored gray, and others are green. (a) The C-terminal fragment of human Daam1 (PDB accession code: 2J1D), which includes the sequence KKKEEEE. (b) The C-terminal domain of the archaeal ribosomal stalk protein aP1 complexed with the GDP-bound archaeal elongation factor aEF1alpha (PDB accession code: 3WY9), which includes the sequence EKKEEEEKEEE. (c) The pre-mRNA-splicing factor Slu7 (PDB accession code: 5MPS), which includes the sequence KKKREQEE.

(d) The translation initiation factor IF2/EIF5B (PDB accession code: 1G7R), which includes the sequence EEEKKKK. (e) The NatA (Naa10p/Naa15p) amino terminal acetyltransferase complex (PDB accession code: 4KVM), which includes the sequence EEEEKKIYKKLKKD. (f) Myosin X (PDB accession code: 5HMO), which includes the sequence EKRAEEEKRKREEEEKRKREEEERERE.

S10. Frequency of α-helix occurrences in positively- and negatively-charged hydrophilic amino acid residue segments identified in the Protein Data Bank

We searched for amino acid sequences including a repetitive K_nE_n sequence fragment, where n is the number of amino acid residues, and calculated the ratios of the involvement of an α-helix with respect to the number of identified sequences (Figure S10(a)). As a consequence, we found that the K_3E_3 and E_3K_3 fragments exhibited significant α-helix propensities, although K_nE_n and E_nK_n fragments with $n > 4$ were not found. Moreover, the ratio of α-helix involvement in the E_3K_3 fragment was larger than that in the K_3E_3 fragment. The resultant data are consistent with those of the present study. In Figure 1, the probability of an α-helix occurring in Helix^{CB} was higher than that in Helix^{NB}.

In Figure S10(b), we performed a similar analysis to that shown in Figure S10(a), by using an extended definition of the positively- and negatively-charged amino acid residue segments, which also included Arg and Asp as well as Lys and Glu. As a consequence, in Figure S10(b), we found similar trends to those in Figure S10(a).

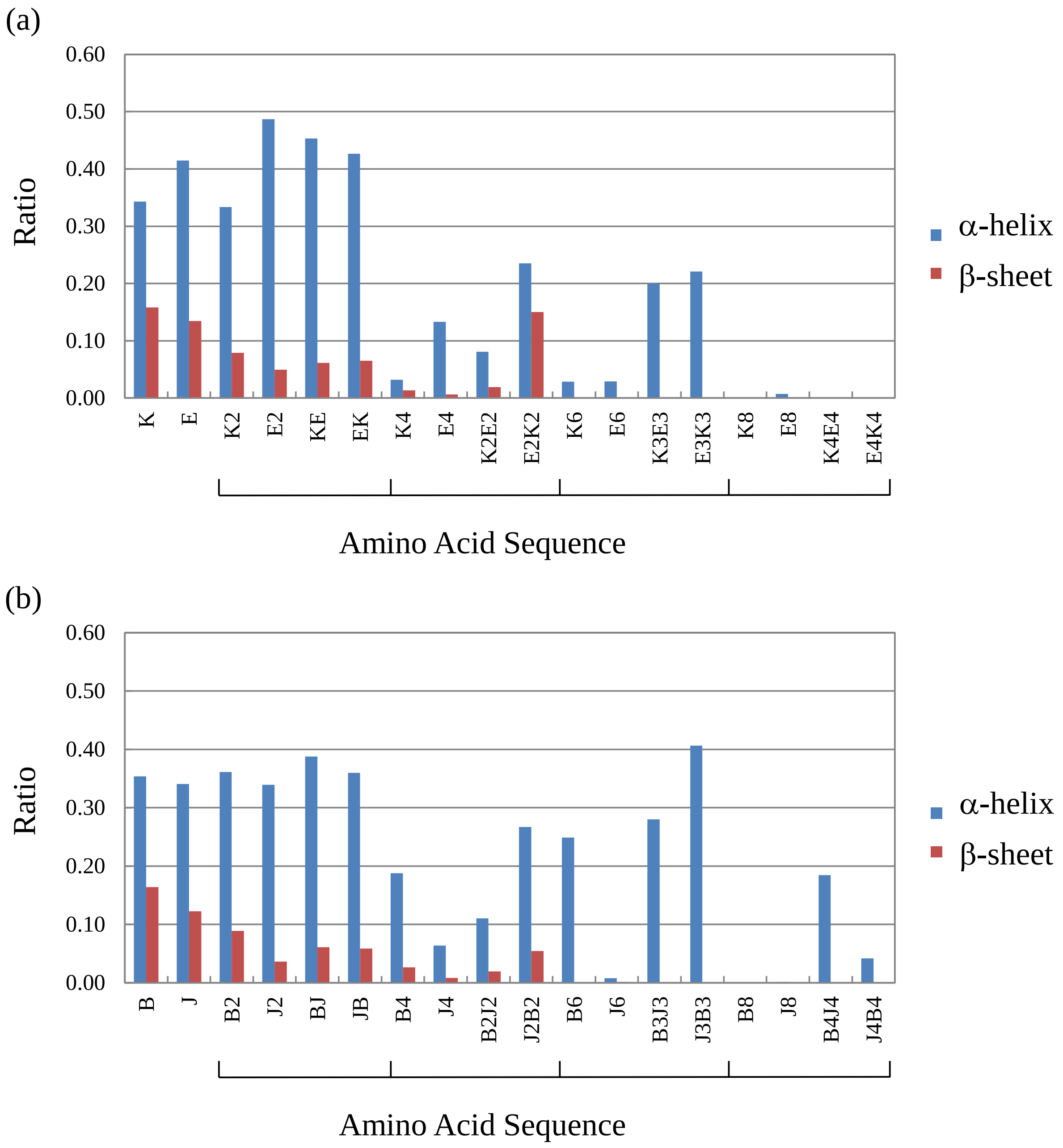

Figure S10. Propensities of α-helix and β-sheet occurrences in amino acid sequences that are similar to positively- and negatively-charged segments. The 133,917 protein entries in the PDB (available in November 2017) were employed for the present statistical analysis. The secondary structure assignments using the DSSP program are available in the Research Collaboratory for

Structural Bioinformatics Protein Data Bank (RCSB PDB, http://rcsb.org)[41]. The combined data of the amino acid sequences and their corresponding secondary structure assignments were employed for the above-mentioned database search. The identified amino acid sequences (the horizontal axis) were exactly identical to the probe sequences, involving α-helices or β-sheets. The ratios of the involvements of α-helices or β-sheets were obtained and compared. (a) Analysis employing K_nE_n sequences as the query, where *n* is the length of positively- and negatively-charged amino acid residues, and K and E denote Lys and Asp residues, respectively. (b) Analysis employing extended amino acid sequences, including B_nJ_n, where B denotes Arg or Lys residues and J denotes Glu or Asp residues, respectively, as the query for the database search.

REFERENCES

1. K. M. Tsutsui, K. Sano, O. Hosoya, T. Miyamoto and K. Tsutsui, *Nucleic Acids Res.*, 2011, **39**, 5067-5081.

2. D. S. Wishart, C. G. Bigam, A. Holm, R. S. Hodges and B. D. Sykes, *J. Biomol. NMR*, 1995, **5**, 67-81.

3. J. Kang, K. Yamasaki, K. Sano, K. Tsutsui, K. M. Tsutsui and M. Tateno, *J. Phys. Soc. Jpn*, 2017, **86**, 014802.

4. P. Cherepanov, G. Maertens, P. Proost, B. Devreese, J. Van Beeumen, Y. Engelborghs, E. De Clercq and Z. Debyser, *J. Biol. Chem.*, 2003, **278**, 372-381.

5. M. Aftabuddin and S. Kundu, *Biophys. J.*, 2007, **93**, 225-231.

6. M. A. Moret, M. C. Santana, G. F. Zebende and P. G. Pascutti, *Phys. Rev. E* 2009, **80**, 041908.

7. S. Sacquin-Mora, *J. R. Soc. Interface*, 2015, **12**, 20150876.

8. J. W. Neidigh, R. M. Fesinmeyer and N. H. Andersen, *Nat. Struct. Mol. Biol.*, 2002, **9**, 425-430.

9. H. Meuzelaar, K. A. Marino, A. Huerta-Viga, M. R. Panman, L. E. Smeenk, A. J. Kettelarij, J. H. van Maarseveen, P. Timmerman, P. G. Bolhuis and S. Woutersen, *J. Phy. Chem. B*, 2013, **117**, 11490-11501.

10. H. Meshkin and F. Zhu, *J. Chem. Theory Comput.*, 2017, **13**, 2086-2097.

11. A. Holtzer and R. B. Hawkins, *J. Am. Chem. Soc.*, 1996, **118**, 4220-4221.

12. L. Mendonça, A. Steinbacher, R. Bouganne and F. Hache, *J. Phys. Chem. B,* 2014, **118**, 5350-5356.

13. J. M. Finke, P. A. Jennings, J. C. Lee, J. N. Onuchic and J. R. Winkler, *Biopolymers*, 2007, **86**, 193-211.

14. J. M. Scholtz, H. Qian, V. H. Robbins and R. L. Baldwin, *Biochemistry*, 1993, **32**, 9668-9676.

15. R. P. Cheng, P. Girinath and R. Ahmad, *Biochemistry*, 2007, **46**, 10528-10537.

16. R. P. Cheng, W.-R. Wang, P. Girinath, P.-A. Yang, R. Ahmad, J.-H. Li, P. Hart, B. Kokona, R. Fairman, C. Kilpatrick and A. Argiros, *Biochemistry*, 2012, **51**, 7157-7172.

17. S. Chakraborty and P. Das, *Sci. Rep.*, 2017, **7**, 9941.

18. C. Yang, S. Jang and Y. Pak, *Nat. Commun.*, 2014, **5**, 5773.

19. Y. Luo, B. Ma, R. Nussinov and G. Wei, *J. Phys. Chem. Lett.*, 2014, **5**, 3026-3031.

20. R. B. Best and J. Mittal, *J. Phys. Chem. B*, 2010, **114**, 14916-14923.

21. K. Sanbonmatsu and A. Garcia, *Proteins*, 2002, **46**, 225-234.

22. C. A. Hanke and H. Gohlke, *J. Chem. Inf. Model.*, 2017, **57**, 2822-2832.

23. N. A. Alves and R. B. Frigori, *J. Phys. Chem. B*, 2018, **122**, 1869-1875.

24. A. Fersht, *Enzyme structure and mechanism*, 1985.

25. R. Friedman, E. Nachliel and M. Gutman, *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, 2005, **1710**, 67-77.

26. D. G. Isom, C. A. Castañeda, B. R. Cannon, P. D. Velu and B. García-Moreno E., *Proc. Natl. Acad. Sci. U.S.A.*, 2010, **107**, 16096-16100.

27. P. Robustelli, S. Piana and D. E. Shaw, *Proc. Natl. Acad. Sci. U.S.A.*, 2018, 201800690.

28. K. Lindorff‐Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror and D. E. Shaw, *Proteins*, 2010, **78**, 1950-1958.

29. S. Piana, K. Lindorff-Larsen and D. E. Shaw, *Biophys. J.*, 2011, **100**, L47-L49.

30. R. B. Best, W. Zheng and J. Mittal, *J. Chem. Theory Comput.*, 2014, **10**, 5113-5124.

31. M. E. Johnson, C. Malardier-Jugroot, R. K. Murarka and T. Head-Gordon, *J. Phys. Chem. B*, 2008, **113**, 4082-4092.

32. A. Nath, M. Sammalkorpi, D. C. DeWitt, A. J. Trexler, S. Elbaum-Garfinkle, C. S. O'Hern and E. Rhoades, *Biophys. J.*, 2012, **103**, 1940-1949.

33. S. Piana, A. G. Donchev, P. Robustelli and D. E. Shaw, *J. Phys. Chem. B*, 2015, **119**, 5113-5123.

34. S. Rauscher, V. Gapsys, M. J. Gajda, M. Zweckstetter, B. L. de Groot and H. Grubmüller, *J. Chem. Theory Comput.*, 2015, **11**, 5513-5524.

35. Q. Qiao, G. R. Bowman and X. Huang, *J. Am. Chem. Soc.*, 2013, **135**, 16092-16101.

36. R. B. Best and G. Hummer, *J. Phys. Chem. B*, 2009, **113**, 9004-9015.

37. D. Shortle, *The FASEB Journal*, 1996, **10**, 27-34.

38. D. Shortle and M. S. Ackerman, *Science*, 2001, **293**, 487-489.

39. H. I. Rösner and F. M. Poulsen, *Biochemistry*, 2010, **49**, 3246-3253.

40. S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman, *Nucleic Acids Res.*, 1997, **25**, 3389-3402.

41. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne, in *International Tables for Crystallography Volume F: Crystallography of biological macromolecules*, Springer, 2006, pp. 675-684.