The Protein Journal (2022) 41:551–562
https://doi.org/10.1007/s10930-022-10076-3

Propensities of Some Amino Acid Pairings in α-Helices Vary with Length

Cevdet Nacar[1]

Accepted: 15 September 2022 / Published online: 28 September 2022

Abstract
The results of secondary structure prediction methods are widely used in applications in biotechnology and bioinformatics. However, the accuracy limit of these methods could be improved up to 92%. One approach to achieve this goal is to harvest information from the primary structure of the peptide. This study aims to contribute to this goal by investigating the variations in propensity of amino acid pairings to α-helices in globular proteins depending on helix length. (n):(n + 4) residue pairings were determined using a comprehensive peptide data set according to backbone hydrogen bond criterion which states that backbone hydrogen bond is the dominant driving force of protein folding. Helix length is limited to 13 to 26 residues. Findings of this study show that propensities of ALA:GLY and GLY:GLU pairings to α-helix in globular protein increase with increasing helix length but of ALA:ALA and ALA:VAL decrease. While the frequencies of ILE:ALA, LEU:ALA, LEU:GLN, LEU:GLU, LEU:LEU, MET:ILE and VAL:LEU pairings remain roughly constant with length, the 25 residue pairings have varying propensities in narrow helix lengths. The remaining pairings have no prominent propensity to α-helices.

Keywords Secondary structure prediction · Residue pairing · Residue propensity · Helix stability

1 Introduction

Secondary structure prediction methods, SSPMs [1–18], a class of in silico methods, attempt to identify the secondary structural elements of the protein (in case of three-state, these are α-helix, β-sheet and random coil) from primary structure (i.e. amino acid sequence of the protein). This approach is based on the assumption proposed by Haber and Anfinsen [19] which says that all the information required for folding is stored in amino acid sequence. The SSPMs have a wide range of application from 3D structure prediction methods and bioinformatics tools [20] to protein engineering and drug design [21]. The five new methods developed every year since 2010 [22], the increasing number of publications per year [22] and nearly 300 published algorithms [23] also show the dynamics of the research field.

The improvement in SSPMs is mainly measured by their prediction accuracies. The existing SSPMs have achieved 84% accuracy [22, 24] mainly by implementing neural networks and deep learning to prediction algorithms [24]. Another factor in this achievement is the growing number of 3D structures deposited in protein databases [25]. The achievable accuracy limit of SSPMs for three-state secondary structure (α-helix, β-sheet and random coil) of protein was proposed as 88% by Rost [25] but, it was updated to 92% by Ho et al. [23]. Despite the great success in prediction accuracy, there are at least 8 points to improve. This goal could only be achieved by gathering the information stored in the primary structure of the peptide. This information is partly based on the propensities of the amino acids to secondary structural elements. Therefore, studies on propensities of residues would provide important information about the peptides in regards of both secondary structure prediction and structural features.

Many theoretical and experimental studies have shown us that some amino acids, as either single or groups (pairing, di-, tri-, tetrapeptides, etc.), have an evident tendency to helix structure [6, 26–40]. Among these, residue pairings which have backbone hydrogen bonds deserve a special interest. Hydrogen bonding (HB) is one of the major factors in protein folding process [30, 31, 34, 36, 41, 42] but backbone hydrogen bond is dominant [32, 43]. Therefore, it can be concluded from all these studies that backbone hydrogen

✉ Cevdet Nacar
cevnacar@marmara.edu.tr

[1] Department of Biophysics, School of Medicine, Marmara University, Istanbul, Turkey

bond between the (n):(n + 4) residue pairing in helices is the major driving force of protein folding. Despite the studies and findings on residue pairing tendencies, almost nothing is known about the variations in these tendencies with helix length. This is the main problem addressed by this study.

This study aims to contribute to both improvement of SSPMs (in regard of accuracy limit of the SSPMs) and understanding of structural characteristics of globular helices (in regard of helix stability) by retrieving valuable information from the primary structure. This information is in form of variations in (n):(n + 4) pairing propensities with helix length. Although this kind of information is critical for propensity-based SSPMs and has potential to improve the accuracy limit, the number of studies on this issue is very limited. Due to small protein data set used in the study [37] or the preference of single amino acids instead of residue pairings [44], these few studies do not fully cover the issue. Considering the intrinsic role of backbone HB in formation and stability of α-helices in globular proteins, only (n):(n + 4) core residue pairings of α-helices including backbone hydrogen bonds were selected for this study. The length of the helices ranges 13 to 26 residues. It has been shown that as helix length increases, propensities of ALA:GLY and GLY:GLU pairings to α-helix in globular protein increase but of ALA:ALA and ALA:VAL decrease. Frequencies of ILE:ALA, LEU:ALA, LEU:GLN, LEU:GLU, LEU:LEU, MET:ILE and VAL:LEU pairings do not vary with helix length. While 25 residue pairings have varying regularities in narrow length range (i.e., propensities of pairings were investigated in a range of longer than 13 residues and shorter than 26 residues rather than full 13-to-26 residue range), the remaining pairings have no prominent propensity to α-helix.

2 Materials and Methods

2.1 Protein Dataset

Protein structure data set was obtained from Nacar [40]. Data set includes 4594 globular peptide chains from Protein Data Bank [45]. Each chain is no shorter than 100 residues and has at least one helix secondary structure. Resolution of each peptide is better than 2.00 Å and sequence identity is smaller than 25%. Total number of helices and of pairings in each length group (13-to-26 residues) are listed in Table 1.

Table 1 Distribution of helices according to their length

Length (residue)	Number of helices	Number of pairings
13	2912	2737
14	2437	4542
15	2611	7310
16	2289	8487
17	1718	7972
18	1352	7468
19	1223	7903
20	1000	7366
21	684	5520
22	587	5348
23	466	4613
24	360	3930
25	244	2859
26	242	3081
Total	18,125	79,136

2.2 Residue Pairings

(n):(n + 4) amino acid pairings of helices were determined from the findings by Nacar [40] according to this criterion: each residue of pairing must have two backbone hydrogen bonds. This criterion is mainly based on the fact that hydrogen bond network is a determining factor for both protein folding process and helix stability. This factor needs to be equalized for all pairings when determining the propensities of the pairings, to prevent pairing from suppressing their true propensities. Therefore, each residue of pairing must have the same number of hydrogen bond. Hydrogen bonds were determined according to HB criteria identified by Baker and Hubbard [46]. If a pairing did not satisfy this criterion, it was excluded from the study even though it remains within the helix boundaries specified in the PDB file. Because the residues at the N- and C-capping regions (first and last four residues of helix, respectively [47]) may affect the stability of the helix [48], only residue pairings in the core of helices were included and pairings those have any amino acid from N- or C-capping sites were discarded. Because of the absence of free –NH group, proline residue can only have a backbone hydrogen bond with (n + 4) residue. So, pairings including proline residue (that is, PRO:PRO, PRO:XXX and XXX:PRO pairings) were also excluded. Therefore, each residue of accepted pairing, that is (n):(n + 4), has two backbone hydrogen bonds: (n) with (n-4) and (n + 4) residues, and (n + 4) with (n) and (n + 8) residues. All accepted pairings are homogeneous in this context. Due to this restriction, helix with a length of 13 residues has only one pairing, helix with a length of 14 residues has two pairings and so forth.

2.3 Limits of Helix Length

As discussed in Sect. 2.2, the smallest helix length that could include at least one pairing which satisfying the HB criterion is 13 residues. So, lower helix length limit was assigned as 13 residues. Since the number of helices longer than 26 residues in data set is very limited and long helices are mostly

found in membrane and fibrous proteins, the upper helix length limit was set at 26 residues. Therefore, helix length was limited to 13-to-26 residues.

2.4 Frequencies of the Residue Pairings

Frequency of each pairing was calculated as the ratio of the total number of each pairing in specified helix length (L) to the total number of all pairings in the same helix length as percentage (1).

$$f_{XXX_1:XXX_2(L)} = \frac{N_{XXX_1:XXX_2(L)}}{\sum N_{pairings(L)}} \times 100, \quad (1)$$

$f_{XXX_1:XXX_2(L)}$ = *Frequency of* XXX_1 : XXX_2 *pairing in helices with length of L residues*,

$N_{XXX_1:XXX_2(L)}$ = *Number of* XXX_1 : XXX_2 *pairing in helices with length of L residues*,

$\sum N_{pairings(L)}$ = *Total number of all pairings in helices with length of L residues*.

The residue location in pairing is preferential, that is XXX_1:XXX_2 and XXX_2:XXX_1 pairings are not identical. Backbone HB between pairs in helices requires –NH group in (n + 4) residue but proline residue lacks of free –NH group. Because of this restriction, frequencies of the XXX:PRO residue pairings were not calculated.

2.5 Frequencies of the Amino Acids

Frequency of each amino acid in helices with a certain length was calculated as the ratio of the total number of each amino acid in helices with the same length to the total number of all amino acids in helices with the same length as percentage. Amino acid frequencies were calculated in two different ways. In the first way (labeled as $\mathbf{f_{nonHB}}$), amino acid set contains all residues remaining within helix boundaries specified in PDB files (disregarding the backbone hydrogen bonding, HB criterion) (2) but, in the second (labeled as $\mathbf{f_{HB}}$), it contains only residues satisfying the backbone HB criterion in these helices (3).

$$f_{nonHB(L)} = \frac{N_{XXX(L)}}{\sum N_{amino\_acids(L)}} \times 100, \quad (2)$$

$f_{nonHB(L)}$ = *Frequency of XXX amino acid in helices with length of L residues*,

$N_{XXX(L)}$ = *Number of XXX amino acid in helices with length of L residues*,

$\sum N_{amino\_acids(L)}$ = *Total number of all amino acids in helices with length of L residues*,

$$f_{HB(L)} = \frac{N_{XXX(L)}}{\sum N_{amino\_acids(L)}} \times 100 \quad (3)$$

$f_{HB(L)}$ = *Frequency of XXX amino acid in helices with length of L residues*,

$N_{XXX(L)}$ = *Number of XXX amino acid in helices with length of L residues*,

$\sum N_{amino\_acids(L)}$ = *Total number of all amino acids in helices with length of L residues*.

2.6 Trend lines in pairing figures

Trend lines in pairing figures (Figs. 1, 2, 3, 4, 5) were drawn by Excel Software from Microsoft Office Professional Plus 2016 package using simple linear regression method based on least-square estimation technique. Parameters of trend lines [slope, y-intercept, coefficient of determination (R^2) and standard error of estimate (SEE)] were also calculated using the same software and represented in Table 2.

3 Results

3.1 Frequencies of the Amino Acids

Amino acid frequencies calculated as $\mathbf{f_{nonHB}}$ and $\mathbf{f_{HB}}$ are represented in Tables 3 and 4, respectively. As seen in Tables 3 and 4, variations in amino acids frequencies whose calculated for helices with a certain length are negligible except Proline. This finding implies that variations in propensities of residue pairings in helices are independent of their residue frequencies.

3.2 Propensities of the Residue Pairings in Helices with Different Length

Frequencies of all 400 residue pairings (except pairings include PRO) in helices with different length, errors in frequencies, and calculations for parameters of trend lines are

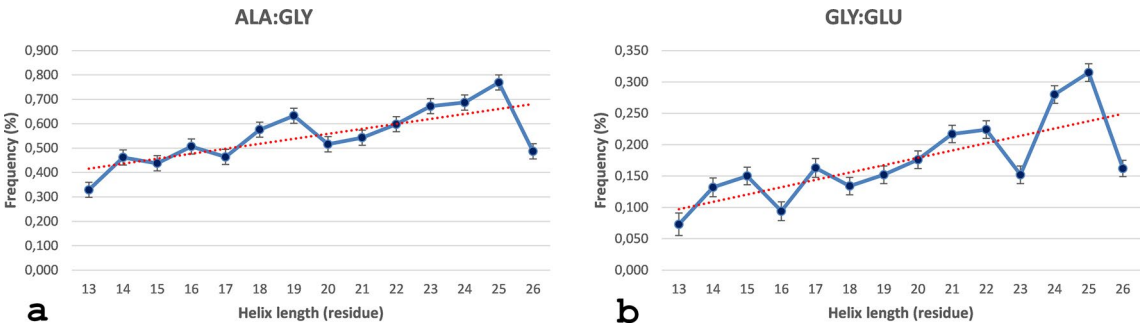

Fig. 1 The propensities of ALA:GLY (**a**) and GLY:GLU (**b**) pairings are increasing with helix length. Trends in increase are represented in red dotted lines drawn by simple linear regression method and standard errors are represented as error bars. See Table 2 for parameters of trend lines (Color figure online)

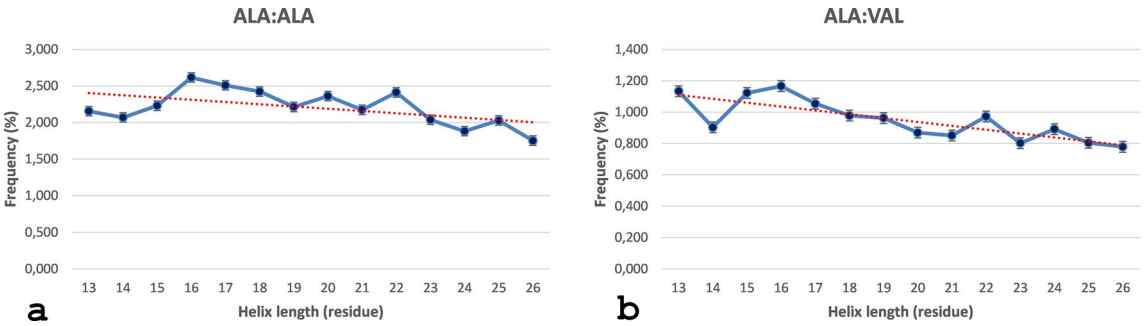

Fig. 2 The propensities of ALA:ALA (**a**) and ALA:VAL (**b**) pairings are decreasing with helix length. Trends in decrease are represented in red dotted lines drawn by simple linear regression method and standard errors are represented as error bars. Some error bars may not be visible because they are smaller than points. See Table 2 for parameters of trend lines (Color figure online)

represented in Supplement_1-Pairing Frequencies, Errors, SEE.xslx. Variations of these frequencies with length (that is, propensities of the pairings) are represented in Supplement_2-All Pairing Propensities.docx as graphs with linear regression line. Standard errors of frequencies of pairings in each length group were represented in figures as error bars. Although frequencies of pairings including proline residue were determined, these results were not evaluated for pairing propensities because they do not satisfy the double backbone hydrogen bond requirement described in Sect. 2.2. Propensities of ALA:GLY (Fig. 1a) and GLY:GLU (Fig. 1b) pairings to α-helix in globular protein increase with increasing helix length while propensities of ALA:ALA (Fig. 2a) and ALA:VAL (Fig. 2b) pairings decrease.

The steep decreases at residues of 26 in both ALA:GLY (Fig. 1a) and GLY:GLU (Fig. 1b) pairings seem like kind of variations those seen in other pairings such as HIS:MET (at residue 24, Fig. 4b) or HIS:ARG (at residues 14 and 23, Fig. 5a). Despite this similarity, these decreases requires a further explanation because of their locations. Their propensities may vary in longer peptides. However, because this study limits the maximum helix length to 26 residues, this discrepancy could be only resolved by further studies including longer peptides.

Propensities of ILE:ALA, LEU:ALA, LEU:GLN, LEU:GLU, LEU:LEU, MET:ILE and VAL:LEU (Fig. 3a–g) pairings were considered as steady despite some relatively small discrepancies in frequencies corresponding to certain lengths. These discrepancies are frequencies corresponding to length of 13, 13, 14, 24, 13, 19, and 13 residues in ILE:ALA (Fig. 3a), LEU:ALA (Fig. 3b), LEU:GLN (Fig. 3c), LEU:GLU (Fig. 3d), LEU:LEU (Fig. 3e), MET:ILE (Fig. 3f) and VAL:LEU (Fig. 3g) pairings, respectively.

There are also regularities in narrow length ranges in 25 pairings (listed in Table 5) corresponding to local propensities in related graphs (see Supplement_3-25 Pairings Propensities.docx). Four of them, two for increase (GLY:VAL, HIS:MET pairings) and two for decrease (ASP:ILE and VAL:PHE pairings), are represented in Fig. 4a–d, respectively. The remaining residue pairings have no prominent propensity or frequency variations.

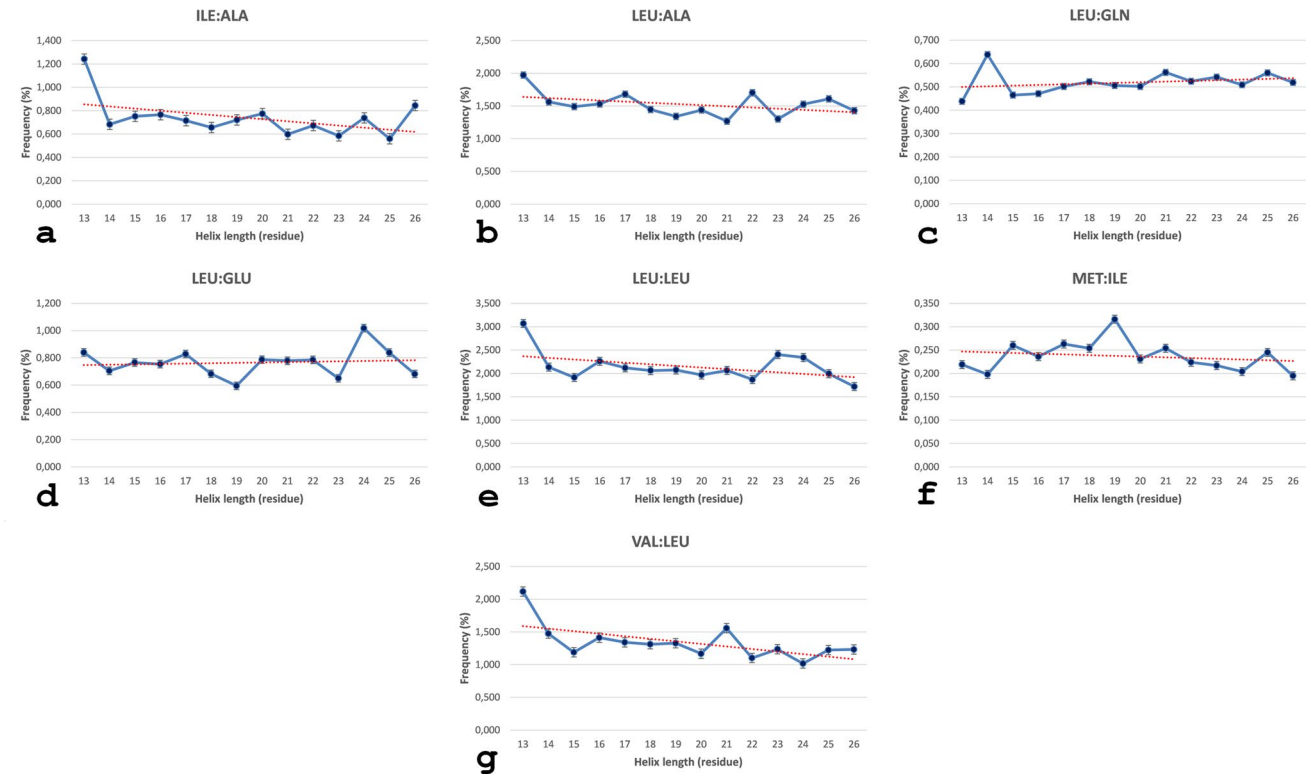

Fig. 3 The propensities of ILE:ALA, LEU:ALA, LEU:GLN, LEU:GLU, LEU:LEU, MET:ILE, and VAL:LEU pairings are remaining roughly constant with helix length despite some exceptions. For instance, the frequencies corresponding to helix length of 13 residues in ILE:ALA (**a**), LEU:LEU (**e**) and VAL:LEU (**g**) pairings are relatively higher than the remaining lengths. Trends are represented in red dotted lines drawn by simple linear regression method and standard errors are represented as error bars. Some error bars may not be visible because they are smaller than points. See Table 2 for parameters of trend lines (Color figure online)

4 Discussion

In this study, how the propensities of (n):(n + 4) amino acid pairings in α-helices in globular proteins varies with helix length was investigated using a comprehensive and qualified protein data set [40]. In statistical studies, the size of the data set directly effects the quality of the findings. Therefore, a data set that is representative of all globular helices was chosen for this study. Residue pairings satisfying only that criterion were accepted: each residue of pairing must have two backbone-based hydrogen bonds; one with (i − 4)th residue and one with the (i + 4)th residue. This criterion was set to make all pairings identical in context of hydrogen bond numbers and it is mainly based on Rose et al.'s "backbone-based theory of protein folding" assumption [43]. This theory states that backbone-based HB is the dominant driving force of protein folding. Since α-helices have many backbone-based hydrogen bonds, their structural features should be highly related to these bonds. Therefore, improvement in understanding of relationship between structural characteristics of helices and backbone HB would be valuable in regards of both SSPMs and protein stability.

The number of studies on relation between residue pairing propensities and helix length is very limited. Some studies have shown that propensities of single amino acids vary with helix length [44, 49] or with location in the helix [33]. In one of the rare studies on pairings, Sundaralingam, et al. [37] have investigated how distribution of the amino acids in protein varied with helix length using only charged residues from 47 globular proteins. These residues were grouped as Ion Pairs (charged residues) and Like Pairs (residue with same charge). The study has not included all possible pairings (only Ion Pairs) and HB criterion. When their findings from Table VI on distributions of (n):(n + 4) ion pairs with helix length were represented in a graph (Fig. 5b), it is seen that observed frequencies decrease with length. According to findings of this study, charged residue pairings have no variation in propensity for α-helix but except HIS:ARG. Propensity of HIS:ARG pairing to helix with length range of 16 to 20 residues is decreasing with length and this finding is consistent with the finding by Sundaralingam et al. [37] shown in their Table VI (Fig. 5a, b). However, this deduction is not entirely valid as Sundaralingam et al. [3737] did not show their findings for each Ion Pair separately in their study.

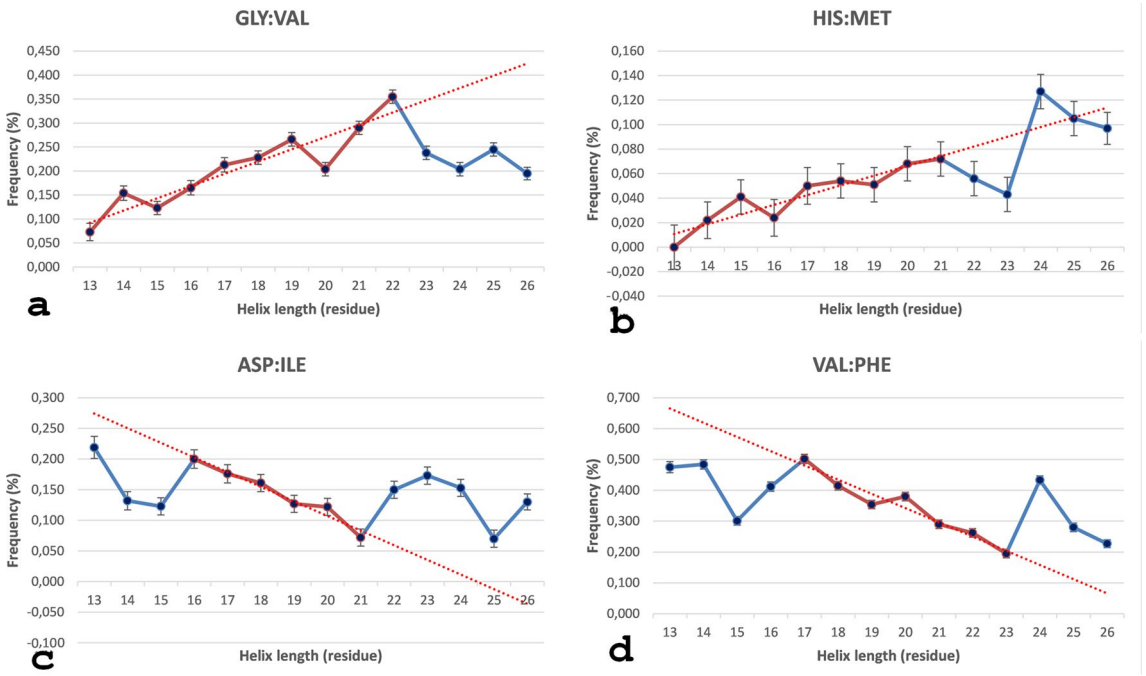

Fig. 4 GLY:VAL (**a**) and HIS:MET (**b**) pairings are shown as two examples for increasing propensities in limited range of helix length. Likewise, ASP:ILE (**c**) and VAL:PHE (**d**) pairings are shown two examples for decreasing propensities. Narrow length ranges of GLY:VAL, HIS:MET, ASP:ILE and VAL:PHE pairings for local propensities are 13–22, 13–21, 16–21 and 17–23 residues, respectively. See Supplement_3-25 Pairings Propensities.docx and Table 5 for remaining pairings. Trends are represented in red dotted lines drawn by simple linear regression method and standard errors are represented as error bars. Some error bars may not be visible because they are smaller than points. See Table 2 for parameters of trend lines (Color figure online)

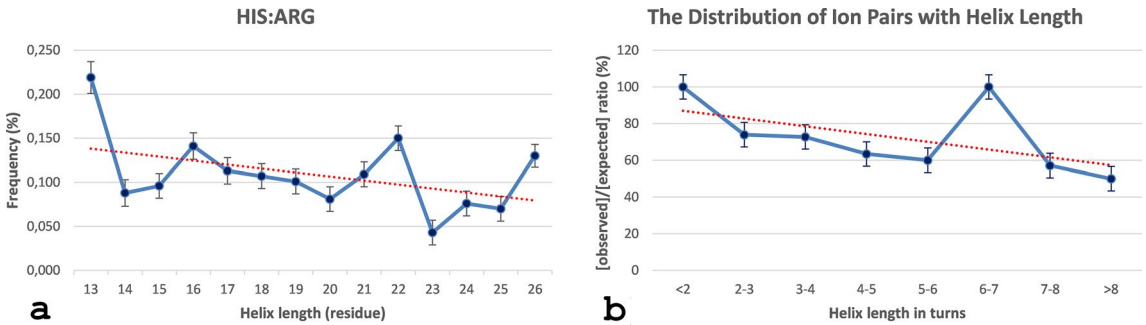

Fig. 5 **a** The finding of "The Distribution of Ion Pairs with Helix Length" is partially consistent with the finding of this study on propensity of HIS:ARG pairing. **b** Was prepared using Table VI from the article by Sundaralingam et al. Because each "Ion Pairs" did not shown separately by Sundaralingam et al., this comparison is not completely valid

In another study done by Wang et al. using 1430 peptides [44], it has been shown that propensities of PRO and TRP amino acids to helix vary with helix length. Also, propensities of residue dyads or adjacent residues (n:n + 1/n − 1:n pairs) to helix structure has been analyzed and shown that propensities of many dyads vary with helix length. Despite the huge protein data set used, the study has not included (n):(n + 4) pairings. Moreover, because the helix length was classified roughly as short, middle and long, variations in propensities can not be assessed precisely. Therefore, findings from study by Wang et al. [44] and from this study are not comparable. Since the helix length in this study was limited to a minimum of 13 residues, studies on short helices were not included in the discussion.

The findings of this study address two issues: secondary structure prediction and helix stability.

Table 2 Parameters of trend lines in pairing figures

Pairing	Slope	y-Intercept	R^2	SEE
ALA:ALA	−0.031	2.803	0.279	0.215
ALA:GLY	0.020	0.151	0.541	0.082
ALA:VAL	−0.025	1.429	0.637	0.083
GLY:GLU	0.012	−0.055	0.544	0.047
ILE:ALA	−0.018	1.090	0.210	0.153
LEU:ALA	−0.018	1.873	0.168	0.175
LEU:GLN	0.003	0.462	0.061	0.049
LEU:GLU	0.003	0.713	0.012	0.108
LEU:LEU	−0.034	2.812	0.197	0.301
MET:ILE	−0.002	0.267	0.042	0.035
VAL:LEU	−0.039	2.095	0.371	0.220
Pairings in narrow length ranges				
ASP:ILE (16–21)*	−0.024	0.585	0.960	0.000
GLY:VAL (13–22)*	0.026	−0.240	0.863	0.035
HIS:MET (13–21)*	0.008	−0.092	0.868	0.000
VAL:PHE (17–23)*	−0.046	1.265	0.941	0.020

*The numbers in paranthesis represent the length range

4.1 Secondary Structure Prediction

An accurate secondary structure prediction is an important step in predicting the tertiary structure of the protein using ab initio prediction methods. One of the most important problems of SSPMs is that the helix boundaries cannot be determined precisely. The helix residues are classified as N-, C-capping and core residues. Since the N- and C-capping residues are the first and last four residues of the helix, respectively, boundaries of the α-helix would be determined precisely if the core residues can be predicted exactly. Because this study includes only core residues, the findings of this study would be valuable in predicting the core region of the helix. Amino acid pairings those have varying propensities with helix length could be used as helix core markers depending on the helix length. Therefore, these residue pairings determined as a consequence of information obtained from primary structure of peptide would improve the accuracy limits of SSPMs. Associated with this progress, ab initio predicting methods would also improve. Since minimum helix length was limited to 13 residues, findings from this study would be useless in predicting the short helices.

4.2 Helix Stability

Although thermodynamics of α-helix formation has been well known for many years [50, 51], it is not clear how the factors that determine the helix stability [30, 34, 36, 39, 52, 53] change with helix length. It is thought that helix stability increases with length and this is mainly related to the hydrogen bond network [54–59]. In this context, it could be proposed that the propensity of (n):(n+4) amino acid pairings to vary with length could be related to the helix stability. Therefore, variations in the propensities of the amino acid pairings presented in the Sect. 3 can be associated with helix stability. So, it could be concluded that the propensity of ALA:GLY and GLY:GLU pairings to increase with length would increase the helix stability. Likewise, ALA:ALA and ALA:VAL pairings would also affect the helix stability negatively. Seven pairings whose tendencies remain constant could be considered as neutral in this sense. Also, the other 25 residue pairings with a certain propensity over a more restricted length range could be evaluated similarly. Although there are studies in the literature with single amino acids, especially with polyalanine peptides, there are no studies those overlap with the findings obtained in this study. It should be noted also that there were no short helices in this study and therefore findings of this study on structural features do not cover all α-helices.

An improved understanding of helix stability would be very useful especially in protein engineering, de novo protein design and protein folding. Some specific amino acid pairings may be preferred or avoided in order to obtain proper degree of stability. Considering that the information on variations in propensities of amino acids with helix length is very limited, findings of this study could make important contributions to the field in this context.

Table 3 The calculated amino acid frequencies for each helix with length of 13-to-26 residues (hydrogen bond criterion not satisfied)

Helix length (residue)	Amino acids frequencies (%) (f_{nonHB})																			
	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
13	11.18	5.95	3.75	5.58	1.25	4.36	8.20	4.88	2.21	5.88	11.52	6.18	2.20	4.16	2.18	5.34	4.57	1.39	3.28	5.96
14	11.87	5.87	3.55	5.05	1.19	4.59	7.84	5.35	2.24	5.65	11.57	5.84	2.35	3.89	1.93	5.96	4.87	1.59	3.22	5.58
15	11.52	6.31	3.76	5.64	1.21	4.61	8.36	4.36	2.26	5.55	11.49	6.14	2.42	3.67	1.99	5.43	4.88	1.37	3.18	5.86
16	11.58	6.08	3.75	5.57	1.28	4.61	8.26	4.62	2.27	5.59	11.14	6.32	2.16	3.83	1.85	5.46	4.85	1.42	3.42	5.94
17	11.01	6.16	3.79	5.42	1.15	4.56	7.99	4.86	1.99	6.01	11.03	6.85	2.26	4.13	1.67	5.35	4.51	1.47	3.52	6.28
18	11.39	5.98	4.03	5.73	1.31	4.93	8.18	4.57	2.21	5.53	11.28	5.92	2.37	3.75	1.76	5.50	4.67	1.42	3.41	6.04
19	11.78	6.68	4.66	5.29	1.12	5.27	7.85	4.31	2.72	5.33	11.30	5.45	2.54	3.75	1.85	5.34	4.48	1.30	3.28	5.69
20	11.57	6.74	3.94	5.12	1.26	4.84	7.86	4.44	2.40	5.42	10.90	5.98	2.40	3.89	1.84	5.54	4.58	1.56	3.72	6.01
21	10.97	6.09	4.05	5.75	1.21	4.97	8.01	4.44	2.52	5.75	11.14	6.18	2.05	3.76	1.54	5.55	4.74	1.53	3.46	6.29
22	12.06	6.25	3.74	5.04	1.41	5.00	8.14	4.50	2.33	5.52	11.03	5.78	2.30	4.01	1.62	5.59	4.61	1.41	3.52	6.15
23	11.63	6.68	3.97	5.57	1.26	4.87	8.33	4.38	2.24	5.55	11.70	6.31	2.37	3.45	1.61	4.72	4.62	1.44	3.49	5.82
24	10.78	6.62	3.68	4.85	1.30	5.03	8.40	4.55	2.52	5.37	11.57	6.63	2.48	3.82	1.46	5.15	4.76	1.32	3.67	6.02
25	11.13	5.69	3.79	4.99	1.25	4.82	8.64	4.77	2.45	5.28	11.44	6.76	2.49	3.92	1.33	4.87	5.07	1.74	3.24	6.32
26	11.53	6.55	3.11	5.23	1.25	5.07	9.01	3.93	2.43	5.50	11.04	6.74	2.36	3.98	1.33	5.09	4.92	1.41	3.19	6.33
Mean*	11.43	6.26	3.83	5.35	1.25	4.82	8.22	4.57	2.34	5.57	11.30	6.22	2.34	3.86	1.71	5.35	4.72	1.46	3.40	6.02
SD**	0.36	0.33	0.32	0.29	0.07	0.24	0.31	0.32	0.17	0.20	0.24	0.40	0.13	0.18	0.24	0.30	0.17	0.11	0.17	0.23

*Mean value of the residue frequencies for length groups

**Standard Deviation value of the residue frequencies for length groups

Table 4 The calculated amino acid frequencies for each helix with length of 13-to-26 residues (hydrogen bond criterion satisfied)

Helix length (residue)	Amino acids frequencies (%) (f_{HB})																			
	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
13	11.59	5.91	3.61	5.47	1.23	4.36	8.32	4.12	2.05	6.12	11.88	6.06	2.27	4.26	2.08	5.21	4.50	1.42	3.39	6.14
14	11.92	5.83	3.52	5.14	1.19	4.66	7.82	4.57	2.09	5.58	12.22	5.90	2.55	4.07	1.89	5.45	4.78	1.46	3.41	5.95
15	11.79	6.39	3.55	5.56	1.20	4.62	8.37	3.87	2.04	5.75	11.84	6.08	2.53	3.76	1.93	5.24	4.87	1.43	3.19	6.00
16	11.91	6.11	3.60	5.45	1.26	4.63	8.35	3.89	2.15	5.81	11.43	6.30	2.27	3.90	1.73	5.40	4.82	1.45	3.48	6.05
17	11.47	6.22	3.55	5.32	1.17	4.75	8.40	4.18	1.91	5.94	11.66	6.25	2.46	4.16	1.64	5.15	4.70	1.31	3.51	6.24
18	11.73	6.10	3.88	5.57	1.26	4.99	8.20	3.97	2.06	5.71	11.62	5.82	2.46	3.84	1.70	5.36	4.60	1.47	3.48	6.15
19	11.82	6.55	3.74	5.27	1.18	4.95	7.98	4.04	2.43	5.54	11.68	5.74	2.45	3.98	1.52	5.21	4.76	1.43	3.57	6.15
20	11.94	6.81	3.78	5.13	1.28	4.87	7.88	3.81	2.23	5.57	11.22	5.93	2.48	3.94	1.66	5.45	4.52	1.61	3.73	6.17
21	11.25	6.19	3.86	5.69	1.21	5.06	8.02	3.83	2.24	6.00	11.55	6.16	2.13	3.81	1.44	5.43	4.63	1.54	3.51	6.46
22	12.31	6.32	3.55	4.95	1.39	5.04	8.16	4.18	2.14	5.67	11.35	5.69	2.42	4.10	1.55	5.47	4.59	1.44	3.48	6.22
23	11.97	6.75	3.79	5.40	1.21	4.89	8.21	3.87	2.10	5.75	11.92	6.35	2.47	3.53	1.57	4.65	4.64	1.46	3.47	6.01
24	11.08	6.67	3.54	4.79	1.22	5.03	8.42	4.12	2.37	5.61	11.95	6.49	2.58	3.83	1.29	5.04	4.70	1.35	3.74	6.16
25	11.26	5.78	3.70	5.06	1.27	4.88	8.65	4.21	2.33	5.30	11.72	6.73	2.52	4.00	1.29	4.76	5.04	1.83	3.31	6.36
26	11.76	6.63	2.98	5.30	1.19	5.08	9.04	3.65	2.28	5.46	11.25	6.73	2.28	4.08	1.23	5.05	4.89	1.40	3.29	6.41
Mean*	11.70	6.30	3.62	5.29	1.23	4.84	8.27	4.02	2.17	5.70	11.66	6.16	2.42	3.95	1.61	5.21	4.72	1.47	3.47	6.18
SD**	0.32	0.33	0.21	0.25	0.06	0.20	0.31	0.22	0.14	0.21	0.28	0.32	0.13	0.18	0.24	0.25	0.15	0.12	0.15	0.15

*Mean value of the residue frequencies for length groups

**Standard Deviation value of the residue frequencies for length groups

Table 5 25 Residue pairings those have local propensities in narrow length range

AA pair	Length	Propensity	AA pair	Length	Propensity
ALA:ASN	15 to 22	↓	HIS:MET	13 to 21	↑
ARG:ALA	13 to 20	↔	HIS:TRP	16 to 23	↔
ASN:SER	15 to 19	↔	ILE:LEU	16 to 23	↓
ASP:ILE	16 to 21	↓	LEU:ASP	16 to 21	↔
CYS:TRP	14 to 21	↔	LEU:THR	18 to 22	↔
GLU:GLN	16 to 22	↔	LYS:VAL	14 to 19	↑
GLU:PHE	16 to 22	↔	MET:ALA	18 to 26	↓
GLU:TYR	13 to 20	↑	PHE:ALA	14 to 21	↔
GLY:MET	14 to 21	↔	THR:VAL	15 to 22	↔
GLY:VAL	13 to 22	↑	TYR:THR	14 to 20	↔
HIS:ALA	15 to 24	↔	VAL:PHE	17 to 23	↓
HIS:ARG	16 to 20	↓	VAL:VAL	14 to 21	↓
HIS:GLU	13 to 21	↔			

5 Conclusion

Even though there are many studies on propensities of amino acids to α-helix, the number of the studies on variation of propensities with length is very limited, especially the ones including (n):(n + 4) pairings. Besides that, findings from these rare studies are not conclusive due to their drawbacks such as insufficient data set or poor pairing criteria. In this study, the variations in propensities of residue pairings with helix length were investigated using a comprehensive data set and a rigorous biophysical criterion based on backbone HB. Findings from this study have shown that as helix length increases, propensities of ALA:GLY and GLY:GLU pairings increase but of ALA:ALA and ALA:VAL decrease. Frequencies of ILE:ALA, LEU:ALA, LEU:GLN, LEU:GLU, LEU:LEU, MET:ILE and VAL:LEU pairings do not vary with helix length. Besides those, 25 residue pairings have varying regularities in narrow length range, the remaining pairings have no prominent propensity to α-helix. These pairings, except the last group, could be used as additional parameters to specifically predict the core region of the α-helix. Therefore, these findings may move forward the SSPMs in regard of accuracy limit. The other contribution of these findings to the field could be in helix stability. Since length is one of the factors of helix stability, parameters related to the length would be useful in evaluating the stability. This issue is especially important in de novo protein design or protein engineering. However, findings of this study do not cover the shorter peptides because of restriction in helix length set at 13-to-26 residues.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10930-022-10076-3.

Funding The authors did not receive support from any organization for the submitted work.

Data Availability All data generated or analysed during this study are included in this published article (and its Supplementary Information Files).

Declarations

Conflict of interest Author declares that he has no conflicts of interest.

Ethical Approval This article does not contain any studies with human or animal participants by the author.

References

1. Chou PY, Fasman GD (1974) Prediction of protein conformation. Biochemistry 13(2):222–245
2. Deleage G, Roux B (1987) An algorithm for protein secondary structure prediction based on class prediction. Protein Eng 1(4):289–294
3. Frishman D, Argos P (1995) Knowledge-based protein secondary structure assignment. Proteins 23(4):566–579
4. Garnier J, Gibrat JF, Robson B (1996) GOR method for predicting protein secondary structure from amino acid sequence. Methods Enzymol 266:540–553
5. Geourjon C, Deleage G (1995) SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. Comput Appl Biosci 11(6):681–684
6. Gibrat JF, Garnier J, Robson B (1987) Further developments of protein secondary structure prediction using information theory. New parameters and consideration of residue pairs. J Mol Biol 198(3):425–443
7. Guermeur Y, Geourjon C, Gallinari P, Deleage G (1999) Improved performance in protein secondary structure prediction by inhomogeneous score combination. Bioinformatics 15(5):413–421
8. King RD, Sternberg MJ (1996) Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. Protein Sci 5(11):2298–2310
9. Levin JM, Robson B, Garnier J (1986) An algorithm for secondary structure determination in proteins based on sequence similarity. FEBS Lett 205(2):303–308
10. Rost B, Sander C (1994) Combining evolutionary information and neural networks to predict protein secondary structure. Proteins 19(1):55–72
11. Garnier J, Osguthorpe DJ, Robson B (1978) Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. J Mol Biol 120(1):97–120
12. Drozdetskiy A, Cole C, Procter J, Barton GJ (2015) JPred4: a protein secondary structure prediction server. Nucleic Acids Res 43(W1):W389–W394
13. Bernhofer M, Dallago C, Karl T, Satagopam V, Heinzinger M, Littmann M et al (2021) PredictProtein—predicting

protein structure and function for 29 years. Nucleic Acids Res 49(W1):W535–W540
14. Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 292(2):195–202
15. Zvelebil MJ, Barton GJ, Taylor WR, Sternberg MJ (1987) Prediction of protein secondary structure and active sites using the alignment of homologous sequences. J Mol Biol 195(4):957–961
16. Salamov AA, Solovyev VV (1995) Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignments. J Mol Biol 247(1):11–15
17. Wako H, Blundell TL (1994) Use of amino acid environment-dependent substitution tables and conformational propensities in structure prediction from aligned sequences of homologous proteins. II. Secondary structures. J Mol Biol 238(5):693–708
18. Mehta PK, Heringa J, Argos P (1995) A simple and fast approach to prediction of protein secondary structure from multiply aligned sequences with accuracy above 70%. Protein Sci 4(12):2517–2525
19. Haber E, Anfinsen CB (1961) Regeneration of enzyme activity by air oxidation of reduced subtilisin-modified ribonuclease. J Biol Chem 236:422–424
20. Pirovano W, Heringa J (2010) Protein secondary structure prediction. Methods Mol Biol 609:327–348
21. Li Z, Yang Y, Zhan J, Dai L, Zhou Y (2013) Energy functions in de novo protein design: current challenges and future prospects. Annu Rev Biophys 42:315–335
22. Yang Y, Gao J, Wang J, Heffernan R, Hanson J, Paliwal K et al (2018) Sixty-five years of the long march in protein secondary structure prediction: the final stretch? Brief Bioinform 19(3):482–494
23. Ho CT, Huang YW, Chen TR, Lo CH, Lo WC (2021) Discovering the ultimate limits of protein secondary structure prediction. Biomolecules 11(11):1627
24. Wardah W, Khan MGM, Sharma A, Rashid MA (2019) Protein secondary structure prediction using neural networks and deep learning: a review. Comput Biol Chem 81:1–8
25. Rost B (2001) Review: protein secondary structure prediction continues to rise. J Struct Biol 134(2–3):204–218
26. de Sousa MM, Munteanu CR, Pazos A, Fonseca NA, Camacho R, Magalhaes AL (2011) Amino acid pair- and triplet-wise groupings in the interior of alpha-helical segments in proteins. J Theor Biol 271(1):136–144
27. Frishman D, Argos P (1996) Incorporation of non-local interactions in protein secondary structure prediction from the amino acid sequence. Protein Eng 9(2):133–142
28. Fonseca NA, Camacho R, Magalhaes AL (2008) Amino acid pairing at the N- and C-termini of helical segments in proteins. Proteins 70(1):188–196
29. Acevedo OE, Lareo LR (2005) Amino acid propensities revisited. OMICS 9(4):391–399
30. Chakrabartty A, Baldwin RL (1995) Stability of alpha-helices. Adv Protein Chem 46:141–176
31. Chakrabartty A, Kortemme T, Baldwin RL (1994) Helix propensities of the amino acids measured in alanine-based peptides without helix-stabilizing side-chain interactions. Protein Sci 3(5):843–852
32. Creamer TP, Rose GD (1992) Side-chain entropy opposes alpha-helix formation but rationalizes experimentally determined helix-forming propensities. Proc Natl Acad Sci USA 89(13):5937–5941
33. Engel DE, DeGrado WF (2004) Amino acid propensities are position-dependent throughout the length of alpha-helices. J Mol Biol 337(5):1195–1205
34. Horovitz A, Matthews JM, Fersht AR (1992) Alpha-helix stability in proteins. II. Factors that influence stability at an internal position. J Mol Biol 227(2):560–568
35. Pace CN, Scholtz JM (1998) A helix propensity scale based on experimental studies of peptides and proteins. Biophys J 75(1):422–427
36. Serrano L, Sancho J, Hirshberg M, Fersht AR (1992) Alpha-helix stability in proteins. I. Empirical correlations concerning substitution of side-chains at the N and C-caps and the replacement of alanine by glycine or serine at solvent-exposed surfaces. J Mol Biol 227(2):544–559
37. Sundaralingam M, Sekharudu YC, Yathindra N, Ravichandran V (1987) Ion-pairs in alpha-helices. Proteins Struct Funct Genet 2(1):64–71
38. Best RB, de Sancho D, Mittal J (2012) Residue-specific alpha-helix propensities from molecular simulation. Biophys J 102(6):1462–1467
39. Rohl CA, Fiori W, Baldwin RL (1999) Alanine is helix-stabilizing in both template-nucleated and standard peptide helices. Proc Natl Acad Sci USA 96(7):3682–3687
40. Nacar C (2020) Propensities of amino acid pairings in secondary structure of globular proteins. Protein J 39(1):21–32
41. Bruch MD, Dhingra MM, Gierasch LM (1991) Side chain-backbone hydrogen bonding contributes to helix stability in peptides derived from an alpha-helical region of carboxypeptidase A. Proteins 10(2):130–139
42. Presta LG, Rose GD (1988) Helix signals in proteins. Science 240(4859):1632–1641
43. Rose GD, Fleming PJ, Banavar JR, Maritan A (2006) A backbone-based theory of protein folding. Proc Natl Acad Sci USA 103(45):16623–16633
44. Wang J, Feng JA (2003) Exploring the sequence patterns in the alpha-helices of proteins. Protein Eng 16(11):799–807
45. Berman H, Henrick K, Nakamura H (2003) Announcing the worldwide Protein Data Bank. Nat Struct Biol 10(12):980
46. Baker EN, Hubbard RE (1984) Hydrogen bonding in globular proteins. Prog Biophys Mol Biol 44(2):97–179
47. Richardson JS, Richardson DC (1988) Amino acid preferences for specific locations at the ends of alpha helices. Science 240(4859):1648–1652
48. Serrano L, Fersht AR (1989) Capping and alpha-helix stability. Nature 342(6247):296–299
49. Kumar S, Bansal M (1996) Structural and sequence characteristics of long alpha helices in globular proteins. Biophys J 71(3):1574–1586
50. Makhatadze GI (2005) Thermodynamics of alpha-helix formation. Adv Protein Chem 72:199–226
51. Murphy KP, Freire E (1992) Thermodynamics of structural stability and cooperative folding behavior in proteins. Adv Protein Chem 43:313–361
52. Aurora R, Creamer TP, Srinivasan R, Rose GD (1997) Local interactions in protein folding: lessons from the alpha-helix. J Biol Chem 272(3):1413–1416
53. Munoz V, Serrano L (1995) Helix design, prediction and stability. Curr Opin Biotechnol 6(4):382–386
54. Baldwin RL (2003) In search of the energetic role of peptide hydrogen bonds. J Biol Chem 278(20):17581–17588
55. Guo H, Karplus M (1994) Solvent influence on the stability of the peptide hydrogen-bond—a supramolecular cooperative effect. J Phys Chem US 98(29):7104–7105
56. Ireta J, Neugebauer J, Scheffler M, Rojo A, Galvan M (2003) Density functional theory study of the cooperativity of hydrogen bonds in finite and infinite alpha-helices. J Phys Chem B 107(6):1432–1437

57. Rossi M, Scheffler M, Blum V (2013) Impact of vibrational entropy on the stability of unsolvated peptide helices with increasing length. J Phys Chem B 117(18):5574–5584
58. Tkatchenko A, Rossi M, Blum V, Ireta J, Scheffler M (2011) Unraveling the stability of polypeptide helices: critical role of van der Waals Interactions. Phys Rev Lett 106(11):118102
59. Wieczorek R, Dannenberg JJ (2004) Comparison of fully optimized alpha- and 3(10)-helices with extended beta-strands. An ONIOM density functional theory study. J Am Chem Soc 126(43):14198–14205