Frontiers in Bioscience Landmark

Front. Biosci. (Landmark Ed) **2023**; 28(6): 127
https://doi.org/10.31083/j.fbl2806127

Original Research

Proline *cis/trans* Isomerization in Intrinsically Disordered Proteins and Peptides

Fanni Sebák[1], János Szolomájer[2], Nándor Papp[1,3], Gábor K. Tóth[2], Andrea Bodor[1,*]

[1]Analytical and BioNMR Laboratory, Institute of Chemistry, Eötvös Loránd University, 1117 Budapest, Hungary
[2]Department of Medical Chemistry, University of Szeged, 6720 Szeged, Hungary
[3]Hevesy György PhD School of Chemistry, Eötvös Loránd University, 1117 Budapest, Hungary
*Correspondence: andrea.bodor@ttk.elte.hu (Andrea Bodor)

Academic Editor: Vladimir N. Uversky
Submitted: 17 April 2023 Revised: 15 May 2023 Accepted: 26 May 2023 Published: 29 June 2023

Abstract

Background: Intrinsically disordered proteins and protein regions (IDPs/IDRs) are important in diverse biological processes. Lacking a stable secondary structure, they display an ensemble of conformations. One factor contributing to this conformational heterogeneity is the proline *cis/trans* isomerization. The knowledge and value of a given *cis/trans* proline ratio are paramount, as the different conformational states can be responsible for different biological functions. Nuclear Magnetic Resonance (NMR) spectroscopy is the only method to characterize the two co-existing isomers on an atomic level, and only a few works report on these data. **Methods**: After collecting the available experimental literature findings, we conducted a statistical analysis regarding the influence of the neighboring amino acid types ($i \pm 4$ regions) on forming a *cis*-Pro isomer. Based on this, several regularities were formulated. NMR spectroscopy was then used to define the *cis*-Pro content on model peptides and desired point mutations. **Results**: Analysis of NMR spectra prove the dependence of the *cis*-Pro content on the type of the neighboring amino acid—with special attention on aromatic and positively charged sidechains. **Conclusions**: Our results may benefit the design of protein regions with a given *cis*-Pro content, and contribute to a better understanding of the roles and functions of IDPs.

Keywords: intrinsically disordered proteins; NMR spectroscopy; proline; *cis/trans* isomerization; statistical analysis

1. Introduction

Despite the lack of a stable secondary structure, intrinsically disordered proteins/protein regions (IDPs/IDRs) play fundamental roles in many biological processes. One factor contributing to the hindrance of secondary structure formation is the high content of proline residues. Proline is the only naturally occurring amino acid that can exist in two conformations (Fig. 1), as in this case, the free energy difference between the *trans* and *cis* conformers is lower than in all other non-prolyl bonds (with typical values of 20 kcal/mol) [1]. In proteins, the peptide bonds are predominantly in the *trans*-state (>99.5%) considering X-non-Pro connections, and it is only around 95% in the case of the X-Pro bond [1,2]. In folded proteins—due to the relatively high steric hindrances between X-Cα and Pro-Cδ atomic environments—the *cis*-isomer is less frequent [3]. However, spontaneous isomerization can occur in highly flexible proteins such as IDPs, and as a consequence both conformers are present in the solution. The situation may become more complicated as IDPs are generally abundant in proline residues [4].

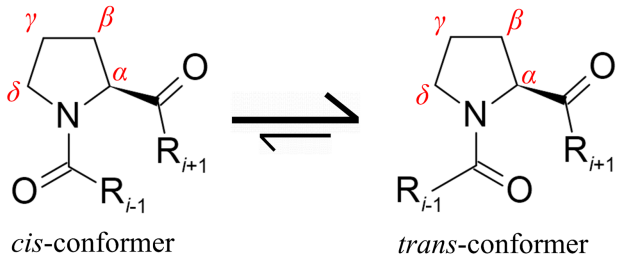

Fig. 1. Schematic representation of proline *cis-trans* isomerization. In proteins, the trans isomer is the predominant form. However, 4–30% of *cis*-isomers in IDPs can occur depending on the Pro-neighboring sequence. Proline atom numbering is shown in red.

The *cis* and *trans*-Pro isomers play key roles in protein–protein interactions as several proteins contain specific polyproline binding domains, such as SRC Homology 3 Domain (SH3), tryptophan-tryptophan domain (WW), ENA-VASP Homology Domain 1 (EVH1), glycine-tyrosine-phenylalanine domain (GYF), and ubiquitin enzyme 2 variant (UEV) [5]. Furthermore, the biological function is associated with the existence of either *cis* or *trans* proline conformation and is linked to cancer [6–8], neurodegenerative diseases [9,10] as well as physiological processes such as the circadian rhythm regulation [11,12].

Therefore, it is important to determine the conformation of the proline residues. However, it is not straightforward to detect and characterize the isomeric ratio. As IDPs are highly mobile systems, atomic resolution studies using X-ray crystallography or cryo-EM cannot be used

[13]. Therefore, Nuclear Magnetic Resonance (NMR) spectroscopy is the only method of investigation capable of discerning between the two conformations. In addition, it is possible to detect several conformations co-existing in the solution. NMR spectra peak multiplications indicate this phenomenon. In the *trans* and *cis*-Pro isomers, the chemical environment is different, and the exchange is slow (10^{-3}–10^{-2} s^{-1}) [14]. Thus, two separate peaks are detected for the Pro as well as for the neighboring residues.

Using 2D ^{1}H,^{1}H-NOESY measurements, the two proline isomers can be distinguished based on the intensity of the Hα-Hα NOE peak between the Pro and the preceding residue [15]. However, the Pro ^{13}Cβ-^{13}Cγ chemical shift difference is a more reliable indicator of the Pro isomer form. This method is commonly used for ^{13}C/^{15}N labeled proteins samples, where specific 3D experiments can detect the Pro sidechain Cβ, Cγ peaks by: hCCCONH [16] in ^{1}H^{N}, Pro-(H)CBCGCAHA in ^{1}H$^{\alpha}$ [17] and 3D (H)CCCON in ^{13}C-detected approaches [18]. In the case of samples with natural isotopic abundance, the approach based on the Pro sidechain ^{13}Cβ, ^{13}Cγ chemical shift determination is rarely used. However, the 2D ^{1}H,^{13}C-HSQC spectrum with appropriate signal-to-noise ratio can be used, even for the low concentration minor form.

Possible *cis*-Pro peak assignment methods for proteins are Pro-Ala mutations [19] or site-specific labeling [20]. Furthermore, the application of proline analogs has gained popularity [12]. The fluorinated amino acids are widely utilized analogs for assessment of *cis/trans*-Pro presence since ^{19}F NMR is more advantageous due to less signal overlap [21].

As a consequence of these experimental difficulties, there are relatively few publications regarding the characterization of the *cis*-Pro isomers in IDPs. Previous studies showed that the *cis-trans* proline isomer ratio depends on the sequence of the Pro neighboring regions in IDPs [22,23]. In our earlier study, we performed a statistical analysis using the available experimental data to determine the effect of the amino acid type of Pro neighboring residues on forming a *cis*-Pro conformer [17]. This analysis was based on 10 IDPs containing 101 Pro neighboring regions ($i \pm 3$ regions, i representing the Pro). Three main groups were considered according to the *cis*-Pro amount: >5%, >10%, and <5% *cis*-Pro containing amino acid sequences.

It was shown at $p = 0.1$ significance level that high (>10%) *cis*-Pro content is favored if aromatic residues (Phe, Tyr, Trp) are present in the $i \pm 1$ positions or negatively charged residues (Asp, Glu) are located i-2, i-1 and $i + 3$ positions. Positively charged residues in i-3, i-1 positions can indicate decreased *cis*-Pro content (<5%).

As a continuation of this study, here we propose a more extensive investigation of the Pro occurrence in IDPs from DisProt database (Database of Protein Disorder, https://disprot.org/) and further characterize the amino acid composition of the Pro and Pro-Pro neighborhood [24,25]. As DisProt does not contain information on the amount of *cis*-Pro isomers, we updated and expanded our previous dataset [17] to the $i \pm 4$ proline neighbors to test whether long-range interactions affect *cis*-Pro formation. While the regularities obtained from the statistical analysis seemed valid for our studied p53$^{1-60}$ region, experimental proof by investigations on well-designed mutations has not been performed yet. Therefore, in this study, we fill this research gap, and using short, 12–15 residue long peptide sequences, we use NMR to analyze the variation of *cis*-proline amounts with well-designed amino acid mutations. For this purpose, we used peptides from the lysine-rich K-segments of Early Response to Dehydration 14 (ERD14) dehydrin from *Arabidopsis thaliana* and the C-terminal region of metastasis-associated human S100A4. We showed earlier that these peptides are capable of cell penetration and might be suitable candidates for drug delivery [26].

2. Materials and Methods

2.1 Statistical Analysis

Amino acid composition of intrinsically disordered proteins was collected from DisProt database (release version 2022_12) and was analyzed using in-house built Python scripts and Microsoft Excel. The DisProt database amino acid composition was determined for the whole dataset (10,544 records) and a filtered dataset (4158 records), from which duplicates based on the sequence and region ID were removed (**Supplementary Table 1**).

The *cis*-Pro neighboring sequence preference with proof by NMR measurements was studied for several IDPs [12,17,19,20,22,27–36] (**Supplementary Table 2**). As the dataset has increased since our previous work, we have updated the results [17]. The current dataset contains 15 IDPs with 167 central i proline environments, and the proline $i \pm 4$ environments were investigated. Residues were divided into 7 groups based on the traditional classification of the amino acids: (1) Gly; (2) aliphatic side chain (Ala, Val, Leu, Ile, Met); (3) polar side chain (Ser, Thr, Cys, Asn, Gln); (4) positively charged side chain (Arg, Lys, His); (5) negatively charged side chain (Asp, Glu); (6) aromatic side chain (Phe, Tyr, Trp); and (7) Pro. Two-sided binomial tests at a significance level of 0.05 and 0.1 were conducted for the statistical analysis.

2.2 Peptide Synthesis and Purification

The designed peptides (Table 1) were synthesized using solid-phase peptide synthesis, applying the Fmoc/tBu strategy using a CEM microwave-assisted fully automated peptide synthesizer. The syntheses were carried out at a 0.25 mmol synthesis scale using a TentaGel S Ram resin (Rapp Polymere GmbH, Tübingen, Germany) with a loading of 0.23 mmol/g amino function. The crude peptides were detached from the solid support using TFA (90%) in the presence of water (5%), 1, 4-dithiothreitol (DTT, 2.5%), and TIS (2.5%). The crude peptides were purified

Table 1. Properties of the studied peptides.

Peptide	Sequence	Length	Number of Pro	$M_{calc(mono)}$	M_{meas}
I.A	DRGLF**P**FLGKKK	12	1	1404.82	1404.57
I.B	DRGLR**P**FLGKKK	12	1	1413.86	1414.01
II.A	FFEGF**P**DKQ**P**RKK	13	2	1622.86	1623.45
II.B	FFEGFADKQ**P**RKK	13	1	1596.84	1597.09
II.C	FFEGF**P**DKQARKK	13	1	1596.84	1596.96
II.D	FFEGFADKQARKK	13	0	1570.83	1570.78
III.	EKKGF**P**EKLKEKL**P**G	15	2	1727.00	1727.04

Prolines are shown in bold, and mutations for Peptide I.A are highlighted with red and for Peptide II.A with blue.

using a C18 RP-HPLC on a PerfectSil 100 ODS3 5 µm column (250 × 10 mm). All compounds were >95% pure by HPLC analysis. The cleaved peptides were analyzed by RP-HPLC/MS (A: 0.1% TFA; B: 80% AcN/water) on a PerfectSil 100 ODS-3 5 µm column (250 × 4.6 mm), with a flow of 1 mL/min. The mass accuracy of the products was determined by ESI-MS using a Waters SQ Detector2 (Waters Corporation, Milford, MA, USA). The mass spectra were recorded in positive ion mode in the 200.0–3000.0 m/z range (**Supplementary Figs. 1,2**).

2.3 Nuclear Magnetic Resonance Experiments and Data Analysis

Typical NMR samples contained 1 mM peptide in 10% D_2O and 0.05 mM 3-(trimethylsilyl)-1-propanesulfonic acid sodium salt (DSS), and the pH was adjusted to 3.0. All NMR spectra were recorded on a Bruker Avance III 700 spectrometer (Bruker GmbH, Ettlingen, Germany) (700.05 MHz for ^{1}H; 70.94 MHz for ^{15}N; 176.03 MHz for ^{13}C) using a Prodigy TCI H&F-C/N-D, 5 mm z-gradient probe head. The measurements were conducted at 298 K. ^{1}H chemical shifts were referenced to the internal DSS standard, whereas ^{15}N and ^{13}C chemical shifts were referenced indirectly via the gyromagnetic ratios.

Resonance assignment and sequential connectivities were determined from classical 2D homonuclear ^{1}H,^{1}H-TOCSY (mixing time: 80 ms) and ^{1}H,^{1}H-NOESY (mixing time: 250 ms) measurements. The typical spectral resolution was 2048 × 512, and the measurements were acquired with 8 and 16 transients, respectively. ^{1}H,^{15}N SOFAST-HMQC, and ^{1}H,^{13}C-HSQC measurements were performed on peptides with a naturally abundant isotope content. The ^{1}H,^{15}N SOFAST-HMQC spectra were acquired with 2048 × 128 resolution, and the number of scans varied between 160–480 to detect the lowly populated minor forms. ^{1}H,^{13}C-HSQC were acquired using a 2048 × 256 resolution with 32 transients. All spectra were processed using TopSpin 3.6.0. (Bruker GmbH, Ettlingen, Germany) Peak assignment was completed using Sparky (University of California, San Francisco, CA, USA) [37].

3. Results and Discussion

3.1 Proline Neighboring Residues in DisProt

Proline, with 7.41% occurrence, is the 5th most frequent amino acid in IDPs according to DisProt (release version 2022_12) (Fig. 2, **Supplementary Table 1**). In order to confirm the amino acid preference, the proline preceding and succeeding neighboring residue type was collected for the Pro residues and Pro-Pro motifs (Fig. 2, **Supplementary Tables 3,4**). Statistical analysis of these data shows that the distribution of amino acid type in the Pro ± 1 positions differs significantly from the DisProt amino acid distribution (Table 2). This signifies that the number of aliphatic and Pro residues significantly increases in these positions, whereas charged and aromatic residues are significantly less frequent. The number of Gly and polar residues are significantly different in the Pro preceding position, but no significant differences in occurrence can be found for $i + 1$ position. The composition of X-Pro-Pro and Pro-Pro-X motifs differ as well: in both cases, the number of aliphatic residues and Pro is increased, and interestingly the number of negatively charged residues shows significant deviation from the DisProt database (**Supplementary Table 5**).

It is important to note that prolines are often situated in proline-rich regions with repetitive Pro containing motifs that often form polyproline II-type helices. While the number of polyproline (containing consecutive proline residues) sequences longer than 20 residues in the UniProt database (https://www.uniprot.org/) is more than 6000, these motifs are underrepresented in the DisProt database, as here the longest polyproline sequence is only 13 residues long (DisProt ID: DP02591r001).

3.2 Proline Neighboring Residues in Intrinsically Disordered Proteins with Known cis-Pro Content

In order to determine the sequence dependence of the *cis*-Pro amount (calculated as [*cis*]/([*cis*] + [*trans*])), we updated our previously published dataset of IDPs and expanded our previous dataset to the Pro neighboring i ± 4 range [17]. The updated dataset now contains 15 IDPs (**Supplementary Table 2**) with 167 central proline residues. The *cis*-Pro content depends on the peptide se-

Table 2. Comparison of the amino acid type occurrence in Pro ± 1 position in DisProt.

X–P					
Residue type	Occurrence in DisProt	Expected value range		Number of occurrences	Comparison to DisProt
		Min	Max		
Gly	7.51%	1172	1258	913	Significantly less
Aliphatic	24.85%	3950	4091	4407	Significantly more
Polar	25.50%	4056	4198	4401	Significantly more
Positive	14.29%	2256	2370	2231	Significantly less
Negative	15.26%	2411	2528	1772	Significantly less
Aromatic	5.17%	800	873	702	Significantly less
Pro	7.41%	1157	1242	1754	Significantly more
Total number of residues:				16,180	
P–X					
Residue type	Occurrence in DisProt	Expected value range		Number of occurrences	Comparison to DisProt
		Min	Max		
Gly	7.51%	1175	1261	1252	No difference
Aliphatic	24.85%	3960	4101	4229	Significantly more
Polar	25.50%	4065	4208	4144	No difference
Positive	14.29%	2261	2375	2073	Significantly less
Negative	15.26%	2417	2534	1974	Significantly less
Aromatic	5.17%	802	875	749	Significantly less
Pro	7.41%	1160	1245	1798	Significantly more
Total number of residues:				16,219	

Based on a two-sided binomial test with a 0.1 significance level.

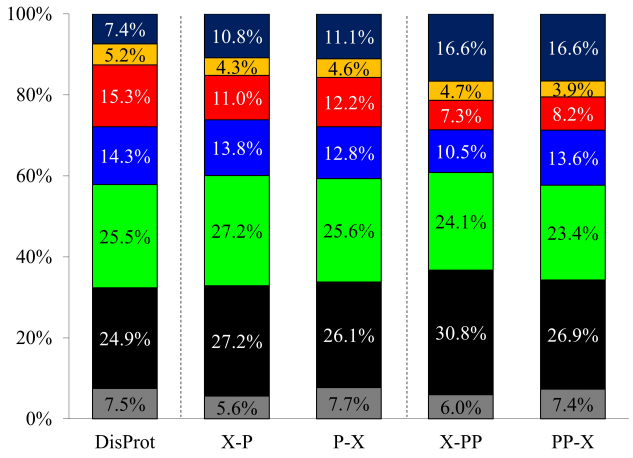

Fig. 2. Amino acid type occurrences in DisProt database and the Pro and Pro-Pro neighboring residues (denoted as X-P, P-X, X-PP, and PP-X, respectively). Gly (gray), aliphatic (black), polar (green), positively charged (blue), negatively charged (red), aromatic (yellow), and Pro (dark blue).

quence length. Therefore peptides (less than 20 residues) are not included in the dataset [34]. Note that polyproline sequences with several consecutive prolines are also excluded due to ambiguous peak assignment.

The $i \pm 4$ Pro neighboring sequences contain a total of 1312 amino acids (not considering Pro in the i position), and the amino acid type occurrence slightly differs from DisProt (Fig. 3, **Supplementary Table 6**); thus, we use the amino acid composition of our overall dataset as a reference.

The Pro neighboring sequences were divided into two groups according to the *cis*-Pro isomer content (Fig. 3A). For *cis*-Pro content $<10\%$, the amino acid type occurrence is similar to the overall dataset. However, significant differences are found for the sequences with $>10\%$ *cis*-Pro content: the positively charged residues are significantly less. In contrast, prolines occur significantly more than expected (Fig. 3A, **Supplementary Table 7**).

In the individual positions for the complete dataset (Fig. 3B), some residue types deviate significantly in the occurrence. The largest differences can be found for Pro and polar residues, where more than a 12% deviation between the highest and the least populated positions can be observed. In addition, two-sided binomial tests at a significance level of 0.05 were conducted to investigate which amino acid types (Fig. 3B) alter in the individual positions. We found that the distant positions (i-3, i-4) do not deviate significantly from the reference. Negatively charged residues are significantly more frequent in the i-2 position. There are significantly more polar residues in i-1 and aliphatic amino acids in $i + 1$ position. Prolines occur significantly more in $i + 4$ position and less in $i \pm 1$ positions. Aromatic amino acids are more frequent in $i + 2$, and less in $i + 3$ positions.

In sequences with more than 10% *cis*-Pro content-considering a 0.05 significance level—the positively

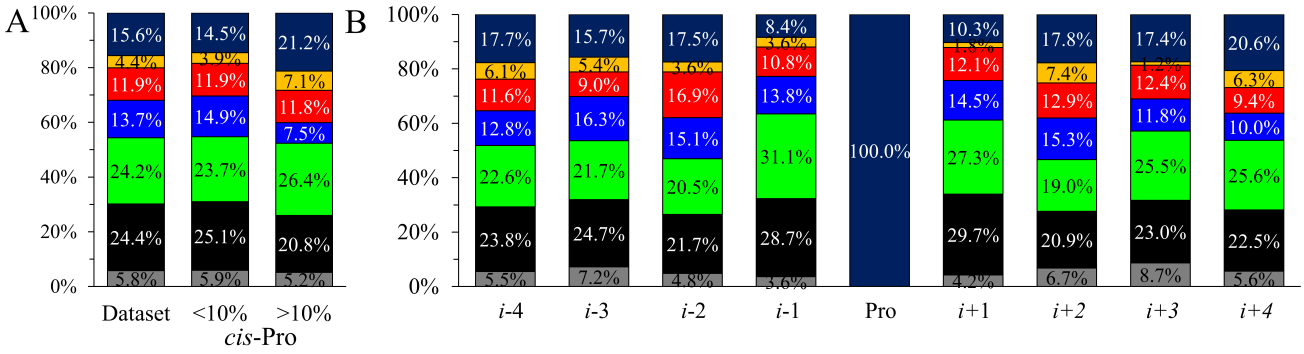

Fig. 3. Comparison of amino acid type occurrences. (A) In the current database and the corresponding two subgroups according to the *cis*-Pro ratio. (B) Amino acid type occurrences in each position in the $i \pm 4$ regions in the current dataset. Gly (gray), aliphatic (black), polar (green), positively charged (blue), negatively charged (red), aromatic (yellow), and Pro (dark blue).

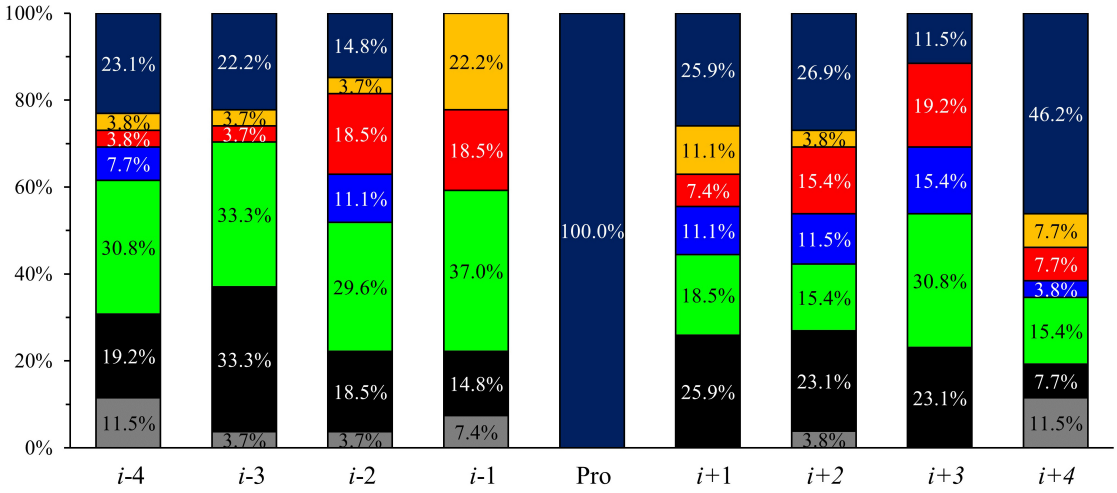

Fig. 4. Acid-type occurrences in more than 10% *cis*-Pro containing sequences in $i \pm 4$ range. Amino acid type occurrences in each position in the $i \pm 4$ regions in the current dataset. Gly (gray), aliphatic (black), polar (green), positively charged (blue), negatively charged (red), aromatic (yellow), and Pro (dark blue).

charged residues are significantly less frequent in *i*-3 and *i*-1 positions (Fig. 4). In the *i*-1 position, aromatic and polar amino acids occur significantly more often, whereas the number of Pro is reduced. Aromatic residues are common as well in the $i + 1$ position. Note, Pro occurrence is significantly higher in $i + 1$, $i + 2$, and even in $i + 4$ positions, and the number of Gly is also increased at $i \pm 4$.

Compared to our previous work, the general rules hold [17]: The number of aromatic amino acids adjacent to Pro ($i \pm 1$) is increased. In Pro-Pro motifs, the *cis*-Pro amount is higher for *cis*-Pro-*trans*-Pro than for *trans*-Pro-*cis*-Pro. Positive charges at *i*-3 and *i*-1 indicate a decreased *cis*-Pro content.

The increased number of Asp and Glu in *i*-2, *i*-1, and $i + 3$ positions does not hold at 0.05 significance level, only at $p = 0.1$.

3.3 Effect of Mutations on the cis-Pro Content

To validate our findings, model peptides with designed mutations were synthesized (Table 1). All peptides were 12–15 residues long and were enriched in Lys and Arg residues to test the effect of the positive charge. Since aromatic residues in the $i \pm 1$ position of the Pro has the strongest effect, Peptide I.A contains two Phe in the Pro vicinity. To test the effect of a positive charge, instead of the aromatic sidechain on the *cis*-Pro content, a Phe5Arg mutation (Peptide I.B) was designed. Peptide II.A contains two prolines: in the case of Pro6, two Phe residues are located in the Pro preceding region: in the *i*-1 and the more distant *i*-4 position. In this case, the $i + 1$ residue is a negatively charged Asp. In the Pro10 neighboring region, there are no aromatic rings, the *i*-4–*i*-1 region contains residues that do not significant affect the *cis*-Pro ratio, and the $i + 1$–$i + 4$ positions are enriched in positively charged residues. In peptides II.B-II.D Pro-Ala mutations were introduced. Peptide III contains two prolines: the $i \pm 2$ regions of Pro6 are similar to Pro6 in Peptide II.A (Gly-Phe-Pro-Asp/Glu-Lys). The main differences are in positions *i*-4 and *i*-3, where Lys is located. Pro14 is close to the C-terminus, and the preceding *i*-4–*i*-1 region contains positive charges in positions

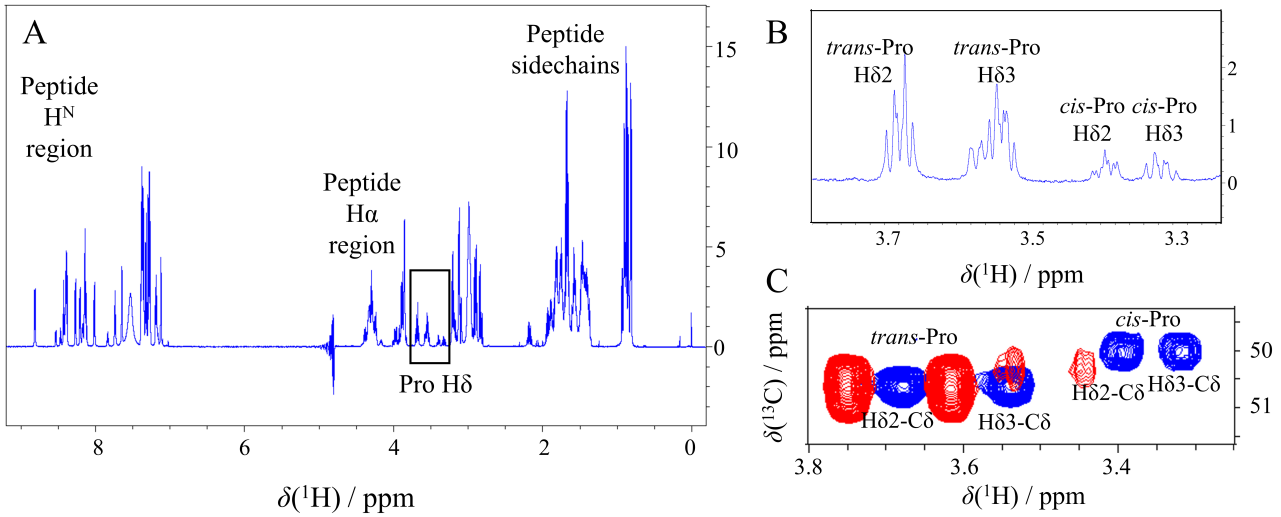

Fig. 5. *Cis*-Pro detection on 1D ^{1}H and 2D ^{1}H,^{13}C-HSQC spectra of Peptide I.A. (A) 1D ^{1}H spectrum of Peptide I.A. Proline Hα (4.2–4.6 ppm) and sidechain Hβ and Hγ (~1.8–2.2 ppm) peak assignment is not possible due to signal overlaps. Pro Hδ region is highlighted using a black rectangle. (B) Peak multiplication in the Pro Hδ region of Peptide I.A: the population of *cis*-conformer is ~26%. (C) Pro Hδ-Cδ region on the ^{1}H,^{13}C-HSQC spectrum of Peptide I.A (blue) and I.B (red). The Pro preceding Phe5 in Peptide I.A was mutated to an Arg residue, causing a significant decrease in the *cis*-Pro content.

Table 3. Peptides and the Pro neighboring regions used in NMR studies with the *cis*-Pro amount.

Peptide	Proline	*i*-4	*i*-3	*i*-2	*i*-1	*i*	*i* + 1	*i* + 2	*i* + 3	*i* + 4	*cis*-Pro%
I.A	P6	R	G	L	*F*	**P**	*F*	L	G	K	~26
II.A	P6	*F*	E	G	*F*	**P**	D	K	Q	**P**	~20
II.C	P6	*F*	E	G	*F*	**P**	D	K	Q	A	~20
III.	P6	K	K	G	*F*	**P**	E	K	L	K	~16
III.	P14	K	E	K	L	**P**	G				~10
II.A	P10	**P**	D	K	Q	**P**	R	K	K		~9
II.B	P10	A	D	K	Q	**P**	R	K	K		~9
I.B	P6	R	G	L	R	**P**	*F*	L	G	K	~9

Prolines are shown in bold; phenylalanines are shown in italics.

that should not influence the *cis*-Pro content.

NMR measurements were performed to test these assumptions. Peak assignment was performed using 2D homo- and heteronuclear measurements on peptide samples with natural isotope abundance. Since the concentration of the minor form is low, approximately 50–200 µM for 1 mM protein, and ^{13}C isotope abundance is ~1%, measurement times are lengthy. Pro isomerization results in a peak multiplication for the Pro neighboring residues (Fig. 5, **Supplementary Fig. 3**). For quantitative *cis*-Pro determination, the Pro Hδ2-Cδ and Hδ3-Cδ cross peaks were integrated on the ^{1}H,^{13}C-HSQC spectrum, since these peaks could be unambiguously assigned as signal overlap rarely occurs in this region (Table 3).

In Peptide I.A Pro6 has phenylalanine in $i \pm 1$ position; therefore, the *cis*-Pro content is increased by 26%. If the disadvantageous Phe to Arg mutation occurs at *i*-1 position (Peptide I.B), the *cis*-Pro ratio decreases to 9% (Fig. 5C).

Peptide II.A contains two prolines where two sets of minor peaks of different signal intensities are detected. Here, the major Pro6 and Pro10 Hδ2-Cδ and Hδ3-Cδ cross peaks overlap on the ^{1}H,^{13}C-HSQC spectrum. Therefore, Pro-Ala mutants were synthesized for unambiguous peak assignment. The Pro6Ala and Pro10Ala mutations cause the absence of the corresponding *cis*-Pro isomer, indicating that the high-intensity *cis*-Pro peak (20%) belongs to Pro6, while the *cis*-P10 content is significantly lower (9%). This agrees with our previous observations: the phenylalanine in *i*-1 position indicates an increased minor content, whereas sequential enrichment in positively charged residues hinders the *cis*-Pro formation.

Peptide III. contains two proline residues. Pro6 succeeds an aromatic phenylalanine which is advantageous for a high *cis/trans* ratio (**Supplementary Fig. 3B**). However, the several positively charged lysines (*i*-4, *i*-3, *i* + 2, *i* + 4) reduce this effect, resulting in a 16% *cis*-Pro amount. The Pro14 neighboring sequence lacks aromatic residues, and the unfavorable positive charge sidechain is located in the neutral *i*-2 position producing a 10% minor isomer.

4. Conclusions

Considering the DisProt database, this study investigated the residue types and their distribution in the Pro neighborhood (*i*-4 to *i* + 4 region). Pro ± 1 positions have significantly higher Pro content, despite polyproline sequences being underrepresented in DisProt. Furthermore, a dataset of IDPs was collected to investigate the effect of the residue types in the *i* ± 4 regions on the *cis*-Pro content. We found that our earlier formulated observations are also valid for the extended dataset. Moreover, the *i* + 4 and *i*-4 positions are significantly enriched in prolines, and the glycine occurrence is higher.

In order to bring experimental proof to our observations based on the statistical analysis, synthetic peptides were designed. The *cis*-Pro content was determined by NMR spectroscopy. We found that the sidechain of the amino acids placed in the *i* ± 1 positions has the largest effect on the *cis*-Pro formation. Aromatic residues have a larger effect on the *cis*-Pro content if they are situated in the proline preceding rather than the succeeding position. Positively charged residues shift the equilibrium to a larger *trans*-Pro content. However, they have no/moderate effect in other positions. We note that even in the case of short, mobile peptides, the initial *cis*-Pro amount is higher than the values found in a protein; the regularities still should hold.

In conclusion, we prove that rationally designed mutations give rise to a desired increase or decrease of *cis*-Pro content. Our results can greatly benefit biotechnological purposes in the design of preferred proline conformations for functional tests.

Abbreviations

ERD14, Early Response to Dehydration 14; EVH1, ENA-VASP Homology Domain 1; GYF, glycine-tyrosine-phenylalanine domain; IDP, intrinsically disordered protein; IDR, intrinsically disordered protein region; SH3, SRC Homology 3 Domain; UEV, ubiquitin enzyme 2 variant; WW, tryptophan-tryptophan domain.

Availability of Data and Materials

The data presented in this study are available on request from the corresponding author.

Author Contributions

AB designed the research study. FS, JS, NP, GT and AB performed the research. FS and AB analyzed the data. FS, GT and AB wrote the manuscript. All authors contributed to editorial changes in the manuscript. All authors read and approved the final manuscript. All authors have participated sufficiently in the work and agreed to be accountable for all aspects of the work.

Ethics Approval and Consent to Participate

Not applicable.

Acknowledgment

The authors thank Dániel Kovács for fruitful discussions regarding statistical analysis and Tünde Horváth for the help with the python scripts.

Funding

This research was funded by National Research, Development and Innovation Office, Hungary, grant numbers NKFI K124900; K137940, and the ELTE Thematic Excellence Programme 2020, National Challenges Subprogramme - TKP2020-NKA-06.

Conflict of Interest

The authors declare no conflict of interest.

Supplementary Material

Supplementary material associated with this article can be found, in the online version, at https://doi.org/10.31083/j.fbl2806127.

References

[1] Wedemeyer WJ, Welker E, Scheraga HA. Proline cis-trans isomerization and protein folding. Biochemistry. 2002; 41: 14637–14644.

[2] Weiss MS, Jabs A, Hilgenfeld R. Peptide bonds revisited. Nature Structural Biology. 1998; 5: 676.

[3] Steinberg IZ, Harrington WF, Berger A, Sela M, Katchalski E. The configurational changes of poly-L-proline in solution. Journal of the American Chemical Society. 1960; 82: 5263–5279.

[4] Theillet FX, Kalmar L, Tompa P, Han KH, Selenko P, Dunker AK, *et al*. The alphabet of intrinsic disorder: I. Act like a Pro: On the abundance and roles of proline residues in intrinsically disordered proteins. Intrinsically Disordered Proteins. 2013; 1: e24360.

[5] Freund C, Schmalz HG, Sticht J, Kühne R. Proline-rich sequence recognition domains (PRD): ligands, function and inhibition. Handbook of Experimental Pharmacology. 2008; 186: 407–429.

[6] Follis AV, Llambi F, Merritt P, Chipuk JE, Green DR, Kriwacki RW. Pin1-induced proline isomerization in cytosolic p53 mediates BAX activation and apoptosis. Molecular Cell. 2015; 59: 677–684.

[7] Brichkina A, Nguyen NT, Baskar R, Wee S, Gunaratne J, Robinson RC, *et al*. Proline isomerisation as a novel regulatory mechanism for p38MAPK activation and functions. Cell Death and Differentiation. 2016; 23: 1592–1601.

[8] Makinwa Y, Musich PR, Zou Y. Phosphorylation-dependent Pin1 isomerization of ATR: its role in Regulating ATR's anti-apoptotic function at mitochondria, and the implications in cancer. Frontiers in Cell and Developmental Biology. 2020; 8: 281.

[9] Nakamura K, Greenwood A, Binder L, Bigio EH, Denial S, Nicholson L, *et al*. Proline isomer-specific antibodies reveal the early pathogenic tau conformation in Alzheimer's disease. Cell. 2012; 149: 232–244.

[10] Meuvis J, Gerard M, Desender L, Baekelandt V, Engelborghs Y. The conformation and the aggregation kinetics of α-synuclein depend on the proline residues in its C-terminal region. Biochemistry. 2010; 49: 9345–9352.

[11] Lu KP, Finn G, Lee TH, Nicholson LK. Prolyl cis-trans isomerization as a molecular timer. Nature Chemical Biology. 2007; 3: 619–629.

[12] Gustafson CL, Parsley NC, Asimgil H, Lee HW, Ahlbach C, Michael AK, *et al*. A slow conformational switch in the BMAL1 transactivation domain modulates circadian rhythms. Molecular Cell. 2017; 66: 447–457.e7.
[13] Nwanochie E, Uversky VN. Structure determination by single-particle cryo-electron microscopy: only the sky (and intrinsic disorder) is the limit. International Journal of Molecular Sciences. 2019; 20: 4186.
[14] Reimer U, Scherer G, Drewello M, Kruber S, Schutkowski M, Fischer G. Side-chain effects on peptidyl-prolyl cis/trans isomerisation. Journal of Molecular Biology. 1998; 279: 449–460.
[15] Wüthrich K. NMR of Proteins and Nucleic Acids. Wiley: New York. 1986.
[16] Montelione GT, Lyons BA, Emerson SD, Tashiro M. An efficient triple resonance experiment using carbon-13 isotropic mixing for determining sequence-specific resonance assignments of isotopically-enriched proteins. Journal of the American Chemical Society. 1992; 114: 10974–10975.
[17] Sebák F, Ecsédi P, Bermel W, Luy B, Nyitray L, Bodor A. Selective 1 Hα NMR methods reveal functionally relevant proline cis/trans isomers in intrinsically disordered proteins: characterization of minor forms, effects of phosphorylation, and occurrence in proteome. Angewandte Chemie - International Edition. 2022; 61: e202108361.
[18] Bermel W, Bertini I, Felli IC, Kümmerle R, Pierattelli R. Novel 13C direct detection experiments, including extension to the third dimension, to perform the complete assignment of proteins. Journal of Magnetic Resonance. 2006; 178: 56–64.
[19] Dujardin M, Madan V, Gandhi NS, Cantrelle FX, Launay H, Huvent I, *et al*. Cyclophilin A allows the allosteric regulation of a structural motif in the disordered domain 2 of NS5A and thereby fine-tunes HCV RNA replication. The Journal of Biological Chemistry. 2019; 294: 13171–13185.
[20] Urbanek A, Popovic M, Elena-Real CA, Morató A, Estaña A, Fournet A, *et al*. Evidence of the reduced abundance of proline cis conformation in protein poly proline tracts. Journal of the American Chemical Society. 2020; 142: 7976–7986.
[21] Killoran PM, Hanson GSM, Verhoork SJM, Smith M, Del Gobbo D, Lian LY, *et al*. Probing peptidylprolyl bond cis/trans status using distal 19 F NMR reporters. Chemistry. 2023; 29: e202203017.
[22] Mateos B, Conrad-Billroth C, Schiavina M, Beier A, Kontaxis G, Konrat R, *et al*. The ambivalent role of proline residues in an intrinsically disordered protein: from disorder promoters to compaction facilitators. Journal of Molecular Biology. 2020; 432: 3093–3111.
[23] Grathwohl C, Wüthrich K. Nmr studies of the rates of proline cis–trans isomerization in oligopeptides. Biopolymers. 1981; 20: 2623–2633.
[24] Quaglia F, Mészáros B, Salladini E, Hatos A, Pancsa R, Chemes LB, *et al*. DisProt in 2022: improved quality and accessibility of protein intrinsic disorder annotation. Nucleic Acids Research. 2022; 50: D480–D487.
[25] Piovesan D, Tabaro F, Mičetić I, Necci M, Quaglia F, Oldfield CJ, *et al*. DisProt 7.0: a major update of the database of disordered proteins. Nucleic Acids Research. 2017; 45: D219–D227.
[26] Sebák F, Horváth LB, Kovács D, Szolomájer J, Tóth GK, Babiczky Á, *et al*. Novel lysine-rich delivery peptides of plant origin erd and human s100: the effect of carboxyfluorescein conjugation, influence of aromatic and proline residues, cellular internalization, and penetration ability. ACS Omega. 2021; 6: 34470–34484.
[27] Gógl G, Biri-Kovács B, Póti ÁL, Vadászi H, Szeder B, Bodor A, *et al*. Dynamic control of RSK complexes by phosphoswitch-based regulation. FEBS Journal. 2018; 285: 46–71.
[28] Aitio O, Hellman M, Skehan B, Kesti T, Leong JM, Saksela K, *et al*. Enterohaemorrhagic Escherichia coli exploits a tryptophan switch to hijack host f-actin assembly. Structure. 2012; 20: 1692–1703.
[29] Paukovich N, Henen MA, Hussain A, Issaian A, Sikela JM, Hansen KC, *et al*. Solution NMR backbone assignments of disordered Olduvai protein domain CON1 employing Hα-detected experiments. Biomolecular NMR Assignments. 2022; 16: 113–119.
[30] Chaves-Arquero B, Pérez-Cañadillas JM, Jiménez MA. Effect of phosphorylation on the structural behaviour of peptides derived from the intrinsically disordered C-terminal domain of histone H1.0. Chemistry. 2020; 26: 5970–5981.
[31] Szabó CL, Szabó B, Sebák F, Bermel W, Tantos A, Bodor A. The disordered EZH2 loop: atomic level characterization by 1HN- and 1Hα-detected NMR approaches, interaction with the long noncoding HOTAIR RNA. International Journal of Molecular Sciences. 2022; 23: 6150.
[32] Ludzia P, Akiyoshi B, Redfield C. 1H, 13C and 15N resonance assignments for the microtubule-binding domain of the kinetoplastid kinetochore protein KKT4 from Trypanosoma brucei. Biomolecular NMR Assignments. 2020; 14: 309–315.
[33] Ahuja P, Cantrelle FX, Huvent I, Hanoulle X, Lopez J, Smet C, *et al*. Proline conformation in a functional tau fragment. Journal of Molecular Biology. 2016; 428: 79–91.
[34] Alderson TR, Lee JH, Charlier C, Ying J, Bax A. Propensity for cis-proline formation in unfolded proteins. ChemBioChem. 2018; 19: 37–42.
[35] Wang Y, Pinet L, Assrir N, Elantak L, Guerlesquin F, Badache A, *et al*. 1H, 13C and 15N assignments of the C-terminal intrinsically disordered cytosolic fragment of the receptor tyrosine kinase ErbB2. Biomolecular NMR Assignments. 2018; 12: 23–26.
[36] Pinet L, Wang YH, Deville C, Lescop E, Guerlesquin F, Badache A, *et al*. Structural and dynamic characterization of the C-terminal tail of ErbB2: Disordered but not random. Biophysical Journal. 2021; 120: 1869–1882.
[37] Goddard TD, Kneller DG. SPARKY 3, University of California, San Francisco.2000. Available at: https://www.cgl.ucsf.edu/home/sparky/ (Accessed: 15 June 2023).