Evidence of the reduced abundance of proline cis conformation in protein poly-proline tracts

Annika N Urbanek, Matija Popovic, Carlos A Elena-Real, Anna Morató, Alejandro N Estaña, Aurélie Fournet, Frédéric Allemand, Ana M Gil, Carlos Cativiela, Juan Cortés, et al.

Evidence of the reduced abundance of proline *cis* conformation in protein poly-proline tracts

Annika Urbanek[1,#], Matija Popovic[1,#], Carlos A. Elena-Real[1,#], Anna Morató[1], Alejandro Estaña[1,2], Aurélie Fournet[1], Frédéric Allemand[1], Ana M. Gil[3], Carlos Cativiela[3], Juan Cortés[2], Ana I. Jiménez[3], Nathalie Sibille[1], Pau Bernadó[1,*]

[1] Centre de Biochimie Structurale (CBS), INSERM, CNRS, Université de Montpellier. 29, rue de Navacelles, 34090 Montpellier, France.

[2] LAAS-CNRS, Université de Toulouse, CNRS, 7 Avenue du Colonel Roche, 31400 Toulouse, France.

[3] Departamento de Química Orgánica, Instituto de Síntesis Química y Catálisis Homogénea (ISQCH), CSIC–Universidad de Zaragoza, 50009 Zaragoza, Spain.

# These authors contributed equally to this work

Corresponding Author: Pau Bernadó (pau.bernado@cbs.cnrs.fr)

Abstract

Proline is found in a *cis* conformation in proteins more often than other proteinogenic amino acids, where it influences structure and modulates function, being the focus of several high-resolution structural studies. However, until now, technical and methodological limitations have hampered the site-specific investigation of the conformational preferences of prolines present in poly-proline (poly-P) homo-repeats in their protein context. Here, we apply site-specific isotopic labeling to obtain high-resolution NMR data on the *cis*/*trans* equilibrium of prolines within the poly-P repeats of huntingtin exon 1, the causative agent of Huntington's disease. Screening prolines in different positions in long (poly-P_{11}) and short (poly-P_3) poly-P tracts, we found that while the first proline of poly-P tracts adopts similar levels of *cis* conformation as isolated prolines, a length-dependent reduced abundance of *cis* conformers is observed for terminal prolines. Interestingly, the *cis* isomer could not be detected in inner prolines, in line with percentages derived form a large database of proline-centered tripeptides extracted from crystallographic structures. These results suggest a strong cooperative effect within poly-Ps that enhances their stiffness by diminishing the stability of the *cis* conformation. This rigidity is key to rationalize the protection towards aggregation that the poly-P tract confers to huntingtin. Furthermore, the study provides new avenues to probe the structural properties of poly-P tracts in protein design as scaffolds or nanoscale rulers.

Introduction

Proline is unique among the 20 genetically coded amino acids[1]. The side chain of proline is covalently linked to the backbone nitrogen atom, forming a pyrrolidine ring and restricting the ϕ torsion angle. Its cyclic structure also influences the *cis*/*trans* equilibrium of the *Xaa*-Pro peptide bonds (Figure 1), resulting in a lower free-energy difference between the *trans* and *cis* conformations. The interconversion of the two isomers is intrinsically slow (timescale of seconds) and it has been shown to play an important role in controlling the duration of cellular processes, which can be accelerated by peptidyl-prolyl isomerases[2,3]. Isomerization can also modulate protein folding[4,5], amyloid formation[6] and protein recognition events[7].

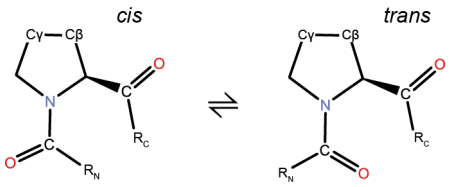


Figure 1. Proline *cis* and *trans* conformations. The positions of Cβ and Cγ are indicated.

Due to their importance, the *cis* and *trans* isomers of proline have been the focus of many structural studies[8–10]. The analysis of high-resolution X-ray structures revealed that in globular proteins ~ 5% of the peptide bonds involving prolines adopt the *cis* conformation, while most other amino acids form peptide bonds in the *trans* conformation (more than 99.5%)[11]. In solution, different nuclear magnetic resonance (NMR) studies of model peptides have reported large variations in the *cis*-proline content (up to 68%), which were dependent on the sequence and length of the peptide, the solvent and whether its termini were blocked or not[10,12,13]. NMR studies of intrinsically disordered proteins (IDPs) indicate that the content of *cis*-proline bonds in these proteins is slightly higher than in globular ones[10,14]. Recent residue-specific studies on intrinsically disordered tau[15], α-synuclein[10] and osteopontin[16] demonstrated that the *cis* population is sequence dependent, varying from 3 - 20%.

Present knowledge of isomerization processes in solution has been derived from isolated proline residues in peptides and proteins, exploiting the sensitivity of neighboring residues to the isomeric state, which is captured through multidimensional NMR experiments. However, these methods can not tackle the position-specific investigation of *cis*/*trans* isomerization in poly-proline (poly-P) repeats, which remains out of the reach of high-resolution structural biology. Indeed, poly-P is one of the most abundant homo-repeats (HRs) in eukaryotes, and often participates in protein-protein interactions[1,17–20]. For instance, the recognition of proline-rich sequences by Src homology 3 (SH3) and WW domains is fundamental for many signaling events. The only known poly-P binding protein requiring at least 6 - 8 consecutive prolines for high affinity binding is profilin. However, the functional role of its interaction is not fully understood[21–23]. In addition, proline-rich regions (PRRs) often delimitate functional/structural regions and protect from aggregation, most probably due to their inherent rigidity

and/or the conformational influence exerted to neighboring residues[14,24,25]. Indeed, poly-P tracts are found adjacent to the aggregation-prone poly-glutamine (poly-Q) motifs in proteins linked to human diseases, such as Huntington's disease and spinocerebellar ataxias SCA2 and SCA7[26].

In solution, poly-P repeats form structures known as poly-proline helices of type-I (all-*cis* right handed, PP-I) or type-II (all-*trans* left handed, PP-II), depending on the polarity of the solvent[27,28]. In aqueous solution, poly-P peptides adopt a PP-II helix with average backbone dihedral angles of (Φ, Ψ) = (-70°, +145°), resulting in an extended rod-like helical structure in which prolines in positions *i* and *i+3* are stacked on top of each other[29]. The energetic stabilization of the PP-II conformation arises from the $n \rightarrow \pi^*$ interaction established between adjacent carbonyl groups ($C_{i-1}=O_{i-1}\ldots C_i=O_i$)[30,31]. These structural features have recently been confirmed in a hexaproline peptide using high-resolution crystallography[32]. The inherent stiffness of poly-P observed by crystallographic studies prompted its use as molecular rulers for Förster resonance energy transfer (FRET) experiments[28,33]. However, this hypothesis was challenged for long poly-P peptides, which displayed larger FRET efficiencies than expected, suggesting a certain degree of flexibility[34]. In combination with NMR experiments and molecular dynamics simulations, it was demonstrated that the *cis*/*trans* isomerization was the main cause of bending of poly-P peptides[27]. Based on ^{1}H-NMR, the authors estimated ≈ 2% and ≈ 10% of *cis* conformations present in internal and C-terminal prolines of the peptides, respectively. Unfortunately, due to the severe signal overlap of the spectra, these NMR studies were not able to pinpoint the homogeneous distribution of the internal *cis*-proline population[35].

Structural studies of poly-P sequences in the protein context are even more challenging than those in peptides. The development of the ^{13}C-detected 2D CON experiment, which correlates ^{13}C' of residue *i-1* with $^{15}N^{H}$ of residue *i*, enables the direct structural interrogation of prolines, which appear in an isolated part of the spectrum[36,37]. Recently, tailored CON pulse sequences have been developed, resolving resonances of protein stretches with up to 4 consecutive prolines[38]. Although position-specific *cis* conformations of isolated prolines can be studied using such pulse sequences[16], it is not clear whether these experiments can also resolve longer poly-Ps and detect their *cis* conformations.

In this work, we have overcome the limitations arising from signal overlap and the use of peptides, and probed the residue-specific *cis*/*trans* equilibrium of prolines within and outside of poly-P tracts in the native protein context of huntingtin exon1 (httex1). By complementing traditional NMR experiments with cell-free (CF) protein synthesis[39] and site-specific isotopic labeling (SSIL)[40–42], we quantified the *cis*/*trans* populations of prolines present in the PRR of a non-pathological httex1 variant containing 16 glutamines (H16). Httex1 mutants containing an abnormally long poly-Q tract form fibrillar deposits in the striatum, causing Huntington's disease, a currently incurable neurodegenerative disorder[43]. Based on its primary sequence, httex1 can be divided into an N-terminal 17 residue long domain (N17), the poly-Q tract of variable length and a PRR, which contains three poly-P tracts of 11 (poly-P_{11}), 3 (poly-P_3) and 10 (poly-P_{10}) proline residues, respectively (Figure 2a).

Experimental evidence shows that N17 and the poly-P tracts exert opposing effects on the poly-Q tract, promoting and preventing aggregation, respectively[24,44–46].

We show that the SSIL approach is robust, yielding similar values for isolated proline residues as homogeneous isotopic labeling and multidimensional NMR. In addition, when applied to the poly-P tracts, our approach shows that in a protein context the *cis* population is dramatically reduced for inner prolines. Furthermore, the length-dependent isomerization behavior of terminal prolines substantiates a cooperative effect of the $n \rightarrow \pi^*$ stabilization forces within poly-P tracts that enhances their stiffness by reducing the *cis* population.

Results

Traditional labeling approaches do not provide position-specific information of H16 poly-P tracts

Httex1 contains a PRR with two long poly-P tracts in addition to a short proline triplet and several isolated prolines, resulting in a total number of 31 prolines (Figure 2a). In order to investigate their conformational properties, we produced a sample exclusively labeled with [^{15}N, ^{13}C]-proline (H16-Pro*) by exploiting total control over the labeling scheme offered by CF protein synthesis[39].

The ^{13}C-HSQC of this sample displayed a reduced number of peaks (Figure 2b and S1a). Two families of prolines were identified in the spectrum, those preceding a proline (Pro-*Pro*), and those preceding one of the other amino acids (Pro-*Xaa,* with *Xaa* being any non-proline residue), encompassing 21 and 10 prolines, respectively (Figure 2a,b). Indeed, each family was characterized by a distinct Cα and two Cβ peaks (Figure 2b and S1a). Conversely, Cγ and Cδ frequencies were identical for all prolines in the sample. Inspection of the upfield part of the spectrum showed the appearance of low-intensity peaks arising from the *cis* conformations of prolines in H16 (Figure 2b)[8]. Concretely, two peaks corresponding to Cγ in the *cis* conformation can be observed upfield (24.6 ppm) of the main (*trans*) Cγ peak (27.3 ppm). Moreover, four Cβ peaks corresponding to the *cis* conformations of the two proline families were observed downfield of the main (*trans*) Cβ peaks (Figure 2b).

The analysis of the intensities of the Cβ and Cγ peaks provided a direct quantification of the relative populations of the *cis* and *trans* conformations in H16 for both proline families. While Cγ intensities provided an overall estimation of 4.0% of *cis* conformations in H16, Cβ intensities gave 7.4 and 3.0% of *cis* conformations for Pro-*Xaa* and Pro-*Pro* families, respectively (Figure 2b). The difference observed between the two families is significant and must be related to the distinct nature of the neighboring residues, which can not be evaluated with present experiments[8].

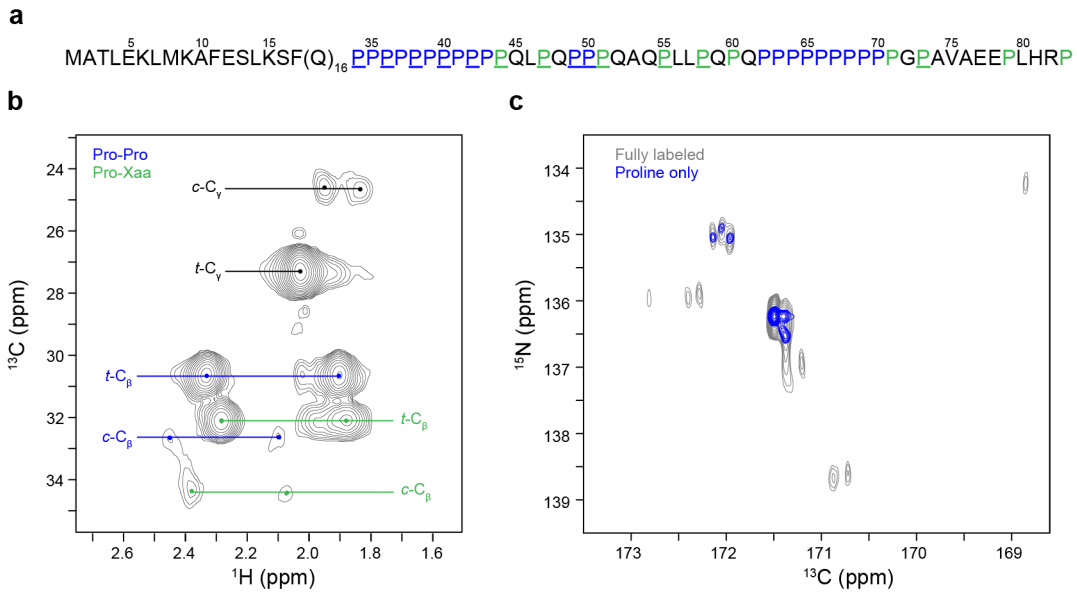


Figure 2. Proline conformations in H16. (**a**) Protein sequence of huntingtin exon1 with 16 consecutive glutamines (H16). Prolines followed by prolines (Pro-*Pro*) are highlighted in blue, while prolines followed by other residues (Pro-*Xaa*) are shown in green. Individual prolines studied by SSIL are underlined. (**b**) Zoom of the ^{13}C-HSQC spectrum for H16-Pro* (H16 sample exclusively labeled with [^{15}N, ^{13}C]-proline). The correlations for the *cis* and *trans* conformations of Cγ (black) and of the two proline families observed for Cβ, Pro-*Pro* (blue) and Pro-*Xaa* (green) are highlighted. (**c**) Zoom on the proline region of the recorded CON spectra, showing an overlay of the fully labeled spectrum (grey) with the spectrum of selectively [^{15}N, ^{13}C]-labeled prolines (blue). Even though H16 contains 21 Pro-*Pro* bonds, only six correlations were observed.

2D CON experiments have proven to be simple and effective to directly detect proline signals[38,47]. To explore if 2D CON experiments are able to resolve longer poly-P clusters, a 2D CON spectrum of a fully [^{15}N, ^{13}C]-labeled sample of H16 was recorded. The resulting spectrum was well dispersed, clearly showing the proline region (^{15}N: 134 - 137 ppm, Figure S1b). Subsequently, to exclusively examine prolines within the poly-P tract of H16, an additional 2D CON experiment of H16-Pro* was recorded. Interestingly, despite the fact that H16 contains more than 20 Pro-Pro bonds, only six groups of signals were detected (Figure 2c). This suggests that many of the Pro-Pro correlations have similar chemical shifts, resulting in a strong signal overlap. Moreover, no low-intensity signals corresponding to *cis* conformations were visible in the spectrum. Based on these observations traditional labeling schemes are not suitable to study the *cis*/*trans* conformation of prolines within poly-P tracts.

Table 1. Experimentally determined proline *cis*/*trans* conformations in H16.

	Proline context	*Data obtained from Cγ (SSIL samples)*		*Data obtained from Cβ (SSIL samples)*		*Data obtained from 3D experiments*
Proline	X_{aa}-Pro-Y_{aa}	Experimental *%cis*	Maximum estimated *%cis*[e]	Experimental *%cis*	Maximum estimated *%cis*[e]	Experimental *%cis*
34[a]	*Q*-P-*P*	16.6 ± 0.2	-	16.2 ± 0.2	-	12.3 ± 2.2
36[a]	*P*-P-*P*	*N.O.*[d]	1.9	*N.O.*[d]	1.9	-
38[a]	*P*-P-*P*	*N.O.*[d]	2.5	*N.O.*[d]	3.0	-
40[a]	*P*-P-*P*	*N.O.*[d]	1.7	*N.O.*[d]	1.8	-
42[a]	*P*-P-*P*	*N.O.*[d]	1.8	*N.O.*[d]	1.8	-
44[a]	*P*-P-*Q*	*N.O.*[d]	1.6	*N.O.*[d]	1.6	-
47	*L*-P-*Q*	6.4 ± 1.3	-	6.0 ± 1.3[f]	-	11.0 ± 1.4
49[b]	*Q*-P-*P*	9.8 ± 3.0	-	12.3 ± 1.7	-	8.4 ± 3.1
50[b]	*P*-P-*P*	*N.O.*[d]	3.4	*N.O.*[d]	2.7	-
51[b]	*P*-P-*Q*	3.3 ± 0.7	-	*N.O.*[d]	-	-
55	*Q*-P-*L*	8.9 ± 1.0	-	7.9 ± 0.9[f]	-	8.1 ± 1.3
58	*L*-P-*Q*	10.2 ± 2.1	-	10.8 ± 1.7[f]	-	9.6 ± 1.1
60	*Q*-P-*Q*	-	-	-	-	8.3 ± 0.8
62[c]	*Q*-P-*P*	-	-	-	-	11.6 ± 1.5
73	*G*-P-*A*	14.6 ± 0.9		12.9 ± 0.9[f]		12.3 ± 1.4
79	*E*-P-*L*	-	-	-		10.7 ± 1.5

a Prolines belonging to poly-P_{11}
b Prolines belonging to poly-P_3
c Prolines belonging to poly-P_{10}
d Non-observed peaks in SSIL samples
e Estimated maximum amount of *cis* population based on the experimental signal/noise level.
f Only the upfield Cβ *cis* peak was used for the calculation of the *%cis*, since the downfield peak could be contaminated by the glutamine Cγ peak appearing due to the natural abundance.

Site-specific isotopic labeling of prolines to study site-specific *cis*/*trans* equilibriums in poly-P tracts

To overcome the limitations encountered to probe the residue-specific *cis*/*trans* equilibrium within poly-P tracts in a native protein context, we applied the SSIL strategy[40,41]. This technique combines CF protein expression and nonsense suppression to site-specifically incorporate a single [^{15}N, ^{13}C]-labeled amino acid into a protein, greatly simplifying NMR spectra. Here, we labeled proline residues within and outside of the poly-P tracts of H16. Briefly, orthogonal nonsense suppressor tRNAs ($tRNA_{CUA}$) were charged with [^{15}N, ^{13}C]-labeled proline *in vitro*, using either a $tRNA_{CUA}$/prolyl-tRNA synthetase (ProRS) pair derived from *Pyrococcus horikoshii*[48] (Figure 3a) or dinitro-flexizyme (dFx)[49] (Figure 3b). The loaded $tRNA_{CUA}$ was then added to a CF reaction containing a plasmid coding for H16 fused

to sfGFP, bearing an amber stop codon (TAG) in the position coding for the targeted proline residue (Figure 3c). Six prolines, from one of the long poly-Ps (poly-P_{11}: P34, P36, P38, P40, P42, P44), the three prolines from the short tract (poly-P_3: P49, P50, P51), and four isolated prolines (P47, P55, P58 and P73) were studied using this strategy. After verifying the suppression efficiency and robustness at small scale (see Figure S2a-c and associated text), the CF reaction volume was increased to 10 mL to produce NMR samples. Despite the fact that dFx and ProRS loaded the $tRNA_{CUA}$ to a different extent (≥ 80% and ≤ 40%, respectively), similar H16 yields were obtained and both strategies were used for large scale production (see Figure S2d). Concretely, P34, P38, P47, P55, P58 were produced with dFx loaded $tRNA_{CUA,}$ and P36, P40, P42, P44, P49, P50, P51, P73 were produced with ProRS loaded $tRNA_{CUA}$ (Table S1).

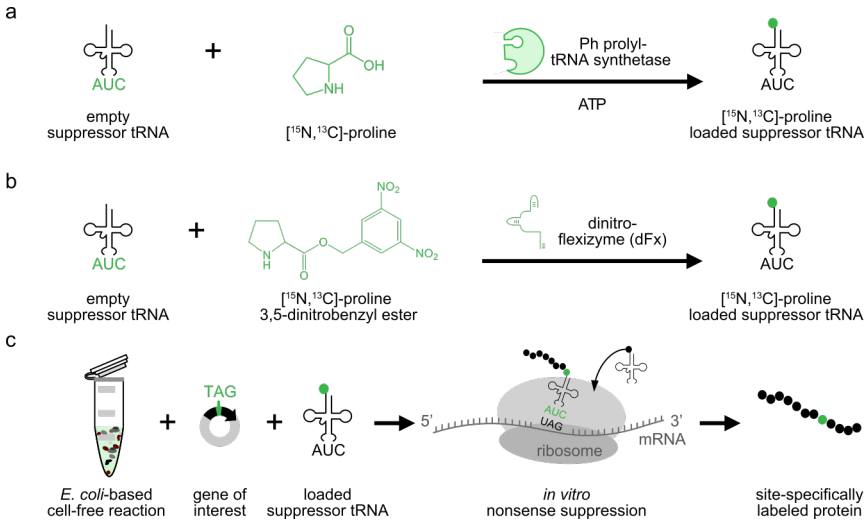


Figure 3. SSIL strategy. (**a**) Enzymatic loading of *in vitro*-transcribed suppressor $tRNA_{CUA}$ with [^{15}N, ^{13}C]-proline. (**b**) Flexizyme-mediated loading of *in vitro*-transcribed suppressor $tRNA_{CUA}$. (**c**) Cell-free suppression strategy employed for the site-specific labeling of single prolines in H16.

Site-specific estimation of *cis* population

^{13}C-HSQC spectra were recorded for the 13 site-specifically ^{13}C-labeled H16 samples (Figures 4a and S3). For each sample only a single Cγ peak and a Cβ pair of peaks, which corresponds to one of the two prolines families (Pro-*Pro* and Pro-*Xaa*, Figure 2b), were present. In order to suppress the contamination arising from natural isotopic abundance peaks, which would impact the observed intensities and the associated *cis* populations, deuterated glutamine and proline were used in the CF reaction. While the addition of deuterated proline eliminated proline-related contaminations, a residual peak of glutamine Cγ was observed in the spectra (34 ppm ^{13}C and 2.4 ppm ^{1}H) (Figure 4a), most probably due to the action of transaminases that transform protonated glutamate into glutamine[40,50,51]. Inspection of the spectra revealed the presence of Cγ and Cβ *cis* peaks only for a few of the prolines (Figures 4a and S3). With the exception of P51, the last residue of poly-P_3, *cis* peaks were only

observed for prolines preceded by a non-proline amino acid. For these samples, the relative *cis*/*trans* population was independently calculated using the intensities of the Cγ and Cβ peaks (Figure 4b and Table 1). Intensities from both carbons provided similar percentages of *cis* conformation and associated errors. For simplicity, only Cγ-derived values will be discussed. The *cis* populations derived from the experiments ranged from 3.3 ± 0.7 (P51) to 16.6 ± 0.2% (P34). With the exception of P51, the population of the *cis* conformer is above the average value obtained for all prolines in H16 (4.0%). The observed variability suggests a notable influence of the neighboring residues on the stability of the *cis* conformation, although no further conclusions could be extracted given the reduced number of examples.

We validated our results using 3D-NMR experiments, benefitting from the fact that the effects of the *cis*/*trans* equilibrium can be propagated to up to two residues from the proline[10]. We assigned the *cis* peaks of the neighboring residues in the ^{15}N-HSQC using several 3D experiments and quantified their relative intensities with respect to the main *trans* peaks. When possible, multiple neighboring residues were used for each proline (Figure S4). In total, we could unambiguously estimate the *cis* percentage for 9 prolines in H16, corresponding to the Pro-*Xaa* (six) and Pro-*Pro* (three) families (Table 1, Figure 4b). These values were in agreement with those obtained from the ^{13}C-HSQC spectra of SSIL samples, although some discrepancy was observed for P47, probably due to the lower quality of its spectrum. This overall agreement demonstrates that the site-specific labeling approach is robust and yields similar *cis*/*trans* populations as classic strategies.

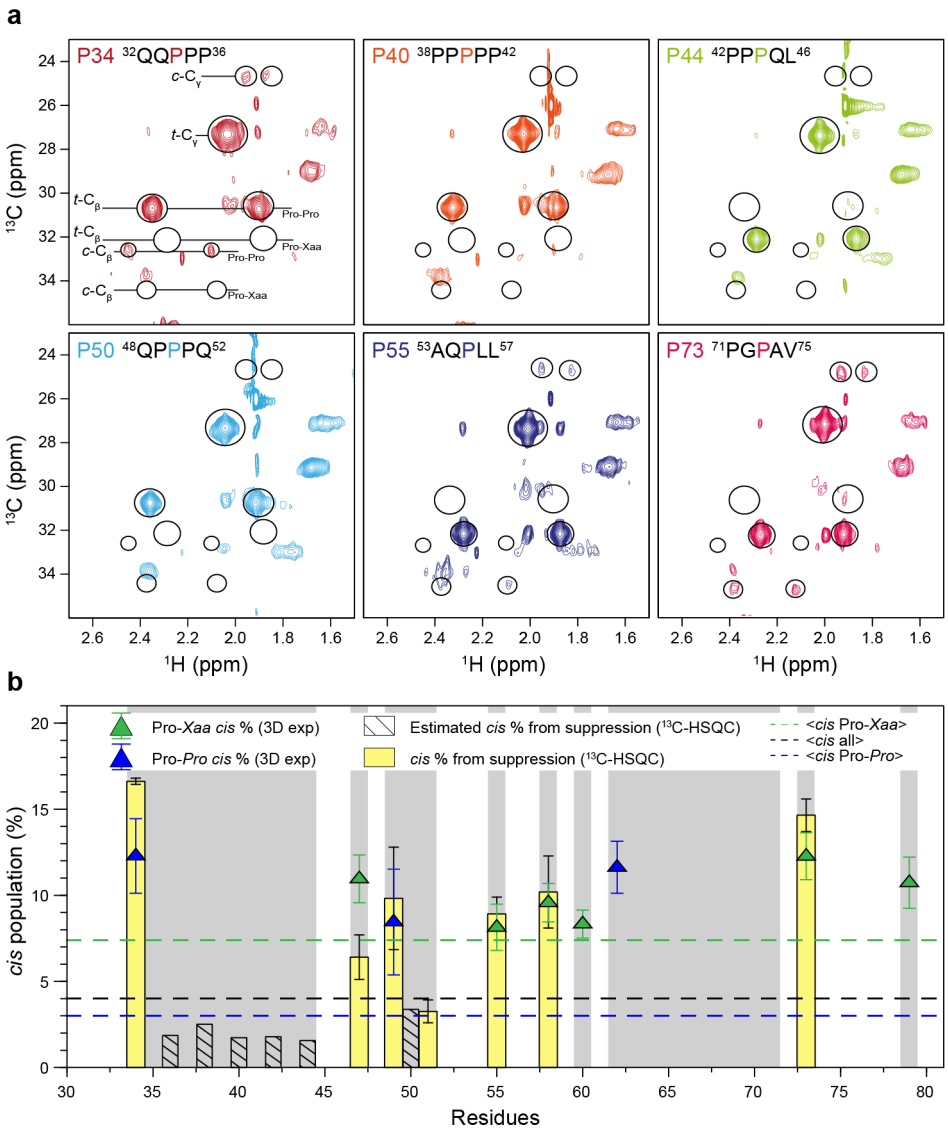


Figure 4. Investigation of proline *cis* populations using site-specifically labeled samples. (**a**) Selected ^{13}C-HSQC spectra of site-specifically [^{15}N, ^{13}C]-labeled samples. Note that all the spectra are shown in figure S3. (**b**) Population of the *cis* conformation for individual prolines in H16 extracted from SSIL ^{13}C-HSQCs (experimentally determined: yellow bars; estimated from the spectral signal/noise: dashed bars) in comparison to those obtained from 3D experiments (Pro-*Pro*: blue triangles, Pro-*Xaa*: green triangles). The averaged *cis* populations for all prolines, Pro-*Pro* and Pro-*Xaa*, obtained from the intensity ratios extracted from the ^{13}C-HSQC spectrum of Figure 2b, are displayed as horizontal dashed lines. The grey background indicates the position of all prolines present in H16, some of which were not probed in this study.

The most interesting observation of the ^{13}C-HSQC experiments is the absence of *cis* peaks for several prolines (Figure 4a and S3). In all cases, these prolines belong to poly-P tracts and are preceded by other prolines, regardless of their family (Pro-*Xaa* or Pro-*Pro*). Based on the signal to noise ratio of each individual spectrum, the maximum percentage of the *cis* conformation that could be hidden below the spectral noise was estimated (see methods section for details). The resulting values are reported in Table 1. Note that the estimations depend on the protein concentration and the spectral parameters used for each individual sample (Table S1). The resulting maximum *cis* percentage

potentially concealed by the noise ranged from 3.4 (P50) to 1.6% (P44), indicating an extremely low *cis* population within poly-P tracts. Interestingly, no differences were observed between the inner and terminal prolines in poly-P_{11}, whereas only the central proline (P50) did not show a *cis* population in poly-P_3. Taken together, these observations suggest a position and length dependency of the *cis* population in the poly-P context.

Statistical analysis of *cis* population in high-resolution crystallographic structures

Our experimental results demonstrate that proline isomerization in H16 tracts is highly context-dependent. In order to evaluate these effects in more detail, we analyzed the *cis*/*trans* populations in a 754,308 proline-centered tripeptide dataset built from coil regions of high-resolution (≤ 2.0 Å) crystallographic structures (see methods section for details). In our dataset, examples for all the 400 *Xaa*-Pro-*Yaa* tripeptides (*Xaa* and *Yaa* referring to any amino acid) were found (Table S2). On average, the *cis* conformation was found in 6.48% of the tripeptides, a value that is slightly larger than those found previously in other studies: 5.70%[52], 4.63%[53] and 4.50%[8]. This difference probably arises from the specific filtering procedures applied to build our database, excluding fragments originating from α-helices and β-strands, which are less prone to have prolines in *cis* conformation[54]. The propensity to adopt a *cis* conformation is highly heterogeneous among the 400 tripeptides, suggesting a complex relationship between residues in positions *i-1* and *i+1* in defining the conformational state of prolines (Table S3).

In the context of the study of proline homo-repeats, it is important to analyze the effects of neighboring prolines on the isomerization. The *cis* populations found in our database for the *Pro*-Pro-*Xaa* and *Xaa*-Pro-*Pro* tripeptides (where *Xaa* refers to any non-proline amino acid) were 8.16% and 9.20%, respectively. Note that these percentages are higher than those found in the whole database (6.48%), suggesting that a proline in positions *i-1* or *i+1* slightly increases the stability of the *cis* isomer. Interestingly, these influences are not additive since the percentage of the *cis* isomer dramatically decreases to 0.81% when both flanking positions are occupied by prolines (*Pro*-Pro-*Pro*). When analyzing longer poly-P fragments (*Xaa*-Pro-Pro-Pro-*Yaa* and *Xaa*-Pro-Pro-Pro-Pro-*Yaa*) extracted from coil regions of crystallographic structures, we identified that the central prolines were equally depleted in *cis* conformations. Concretely, while the population of the *cis* isomer for the N- and C-terminal prolines of *Xaa*-Pro-Pro-Pro-*Yaa* were 10.3 and 11.9%, respectively, the central one was one order of magnitude less populated (1.0%). In line with this observation, none of the central prolines of the 379 *Xaa*-Pro-Pro-Pro-Pro-*Yaa* fragments extracted from crystallographic structures adopted a *cis* conformation. Taken together, these results suggest a strong bias towards the *trans* isomer for the inner prolines of poly-P tracts.

Discussion

Despite being a common homo-repeat found in proteins[17,19,20], current structural knowledge of poly-P tracts in solution has been mainly derived from oligo-L-proline peptides using low-resolution techniques[34,35]. The use of NMR has been severely limited by the low frequency dispersion of the resulting spectra, which has hampered the position-specific investigation of poly-Ps in the protein context. In this study, we have overcome these limitations by applying SSIL to prolines in order to probe the *cis*/*trans* isomerization within two poly-P tracts of huntingtin in a position-specific manner. We have used two different procedures for $tRNA_{CUA}$ loading, an evolved ProRS from *P. horikoshii* and an engineered ribozyme, both of which resulted in very similar final protein yields, close to those reported in previous studies[40,55,56]. This observation seems in contradiction to the different level of $tRNA_{CUA}$ acylation achieved by the ProRS (≈ 80%) and the flexizyme (≈ 40%) strategies (Figure S2d). Note that the spontaneous deacylation reduces the amount of loaded tRNAs inside the CF reaction in a time-dependent manner, making the percentage of loaded $tRNA_{CUA}$ a non-critical parameter (Figure S2e). Our observations demonstrate that flexizyme[49], a less efficient but more general tRNA loading method, can become a valuable tool for the specific incorporation of isotopically labeled or non-canonical amino acids into proteins for specific structural and functional studies requiring mg-scale amounts.

Cis populations measured for H16 prolines using either SSIL samples or 3D-NMR experiments spanned from 3.3 ± 0.7 to 16.6 ± 0.2%, which are in the same range as previous observations on disordered proteins[10,15,16]. The agreement between populations derived from SSIL samples and traditional NMR approaches demonstrates our capacity to identify and quantify small amounts of the *cis* isomer. Notably, the *cis* conformers that we identified by using SSIL were present at concentrations < 0.5 μM in many of the samples, concentrations difficult to detect by traditional 3D experiments.

The variability of the *cis* populations suggests a non-negligible effect of the neighboring residues on the isomerization process. However, no clear correlation can be established between the *cis* populations measured in H16 and those calculated from crystallographic structures (Table S2). With the exception of P51 (^{50}P-P-Q^{52}), the populations of the *cis* conformer measured in solution are similar to or larger than those present in our database. This tendency can be explained by the conformational restriction found in the context of crystals used to extract the tripeptides. Despite the difficulties to quantitatively compare experimental and dataset-derived *cis* propensities, a qualitative evaluation of the effects of the flanking residues on the proline isomerization can be obtained. Our analysis suggests that the role of the flanking residues in defining the *cis* enrichment is not additive, but implies cooperative effects. As suggested previously, the effects of the flanking residues can extend beyond the immediate neighbors of prolines[53,57,58]. While these longer range effects are measured experimentally, they are not captured by our tripeptide dataset. One prominent example of the

interplay of flanking residues is the *Pro*-Pro-*Pro* tripeptide. Although the presence of a proline in position *i-1* or *i+1* (*Pro*-Pro-*Xaa* or *Xaa*-Pro-*Pro*) does not reduce the population of the *cis* conformation of the central Pro, this population is severely depleted when both flanking residues are prolines.

When investigating the *cis*/*trans* isomerization in poly-Ps, we identified three different proline types: N-terminal, inner and C-terminal prolines. The three N-terminal prolines present in H16 have similar *cis* populations (P34: 16.6 ± 0.2, P49: 9.8 ± 3.0, and P62: 11.6 ± 1.5) as isolated prolines. Therefore, they are not conformationally affected by the downstream presence of a poly-P tract. Conversely, we could not detect *cis* conformations in any of the six inner prolines studied. Our estimations of the maximum *cis* population potentially hidden under the spectral noise are low (< 2.0%), although variations among the estimated populations were observed due to sample-specific experimental conditions (Table S1). These estimated populations are in coherence with a previous study on a 40-residue long oligo-L-proline that detected an average *cis* population of 2% for inner prolines, although the position-specific percentages could not be assigned. Interestingly, the two poly-P C-terminal prolines studied provided different results. While no *cis* conformation could be detected for P44 (poly-P_{11}), a small but quantifiable population of the *cis* isomer (3.3 ± 0.7) was observed for P51 (poly-P_3). The strong depletion of the *cis* conformer of C-terminal prolines differs from the tripeptide analysis, which predicts no impact of a proline in position *i-1*. This strongly suggests that the concatenation of multiple prolines destabilizes the *cis* conformation with respect to the *trans* one. Moreover, this effect is length-dependent as only the C-terminal proline of the short tract presents a detectable amount of the *cis* conformation. The observed cumulative effect can be explained by the cooperative $n \rightarrow \pi^*$ interaction established between adjacent carbonyl groups within the homo-repeat. The asymmetric conformational behavior observed for H16 poly-P tracts suggests a cooperative stabilization of the *trans* conformation that grows along the tract and ends with the terminal proline, which will present a weaker $n \rightarrow \pi^*$ interaction with the following residue and a less stable *trans* conformation. Our observations also suggest that longer poly-P tracts lead to an enhanced stabilization of the *trans* conformation and an increase in stiffness. The enrichment of proline residues placed immediately after poly-Q regions found in huntingtin and other glutamine-rich proteins suggests that the stiffness is an important trait that has been preserved along evolution[59,60]. This is most probably due to the structure breaking properties of prolines that are enhanced when they are concatenated in homorepeats and this, for the case of huntingtin, has been proven to diminish aggregation[24,25,61].

The capacity of placing NMR-sensitive isotopes in specific positions in proteins paves the way for high-resolution studies of poly-Ps, allowing the investigation of the link between specific conformational features and biological functions. Moreover, the growing interest of proline as a building block for scaffolds and nanoscale rulers will benefit from the site-specific measurements of the structural properties[28]. Finally, in combination with the extremely versatile flexizyme strategy, the high-resolution evaluation of the impact exerted by a wide range of non-canonical amino acids is now

feasible, making it possible to decipher the structural bases of new chemically modified proteins for novel biotechnological applications.

Materials and Methods

Plasmids

All synthetic genes were ordered from Integrated DNA Technologies (IDT). For cell-free expression, constructs of wild type huntingtin exon1 with 16 consecutive glutamines (H16) or H16 carrying the amber codon (TAG) instead of a proline codon, e.g. P34 (H16P34), were cloned into pIVEX 2.3d by an In-Fusion® (Clontech) reaction. 13 amber mutants were ordered: P34, P36, P38, P40, P42, P44, P47, P49, P50, P51, P55, P58 and P73. The gene coding for *Pyrococcus horikoshii* prolyl-tRNA synthetase variant h1[48] (PhProRS) was cloned into pDB, giving rise to His_6-MBP-PhProRS. The sequence of all plasmids was confirmed by sequencing (GENEWIZ®).

In vitro transcription of suppressor $tRNA_{CUA}$

The suppressor $tRNA_{CUA}$ of a proline $tRNA_{CUA}$/tRNA synthetase pair developed by Chatterjee *et al*.[48] was produced by *in vitro* run-off transcription directly from the PCR product using the HiScribe™ T7 High Yield RNA Synthesis Kit (New England Biolabs) following the manufacturer's instructions. The resulting transcript was purified by phenol-chloroform extraction using a mixture of phenol-chloroform-isoamylalcohol (25:24:1) (Applichem), precipitated in 2 M NH_4-acetate and 2.5 volumes of 96% EtOH at -80°C and stored as dry pellets until use.

Overexpression and purification of PhProRS

Escherichia coli BL21 (DE3) cells were transformed with pDB His_6-MBP-PhProRS and grown overnight at 25°C in ZYM 5052 auto-inducing medium supplemented with 50 μg/mL kanamycin[62]. Cells were harvested by centrifugation (6,000 xg, 20 min, 4°C), the pellet was resuspended in 25 mM Tris pH 7.5, 300 mM NaCl, 5 mM $MgCl_2$, 0.5 mM DTT (PhProRS buffer A) and stored at -20°C. Cells were supplemented with a cOmplete™ EDTA free protease inhibitor tablet (Roche) and lysed by sonication. Cell debris and insoluble proteins were sedimented by centrifugation (45,000 xg, 45 min, 4°C). The supernatant was loaded onto a Ni affinity column (HisTrap Excel 5 mL, GE Life Sciences) and the column was washed with at least 10 CV PhProRS buffer A. Protein was eluted with PhProRS buffer B (PhProRS buffer A + 250 mM imidazole). Fractions were analyzed by SDS-PAGE and fractions containing MBP-PhProRS were pooled, mixed with 3C protease (1/100, w/w) and dialyzed against PhProRS buffer A overnight at 4°C. After removal of 3C protease, the cut protein mix was tested for aminoacylation activity and stored at -20°C.

Synthesis of [^{15}N, ^{13}C]-ProDBE: L-[^{15}N,$^{13}C_5$]-proline-3,5-dinitrobenzyl ester hydrochloride (3)

L-[^{15}N, ^{13}C]-proline-3,5-dinitrobenzyl ester hydrochloride (**3**) was synthesized (Figure S5) starting from commercially available L-[^{15}N, ^{13}C]-proline (CortecNet). Transformation of the carboxylic acid

into a 3,5-dinitrobenzyl ester (DBE) was carried out following a procedure similar to that described by Murakami *et al*.[49] under conditions that were previously optimized working with natural L-proline.

***N*-(*tert*-Butoxycarbonyl)-L-[^{15}N, $^{13}C_5$]-proline (1).** Di-*tert*-butyl dicarbonate (1.37 g, 6.26 mmol) was added to a solution of L-[^{15}N, $^{13}C_5$]-proline (505 mg, 4.17 mmol) and sodium bicarbonate (350 mg, 4.17 mmol) in 20 mL dioxane/water (1:1) cooled to 0°C, and the reaction mixture was stirred overnight at room temperature. After evaporation of the solvent, water (50 mL) was added and the resulting solution was washed with *n*-hexane (2x 30 mL) and then acidified with 5% aqueous potassium bisulfate to pH 2-3. The aqueous solution was extracted with methylene chloride (3x 30 mL). The combined organic layers were dried over anhydrous magnesium sulfate, filtered and evaporated to dryness to yield compound **1** as a white solid (896 mg, 4.04 mmol, 97% yield). HRMS (ESI) $^{13}C_5C_5H_{16}{}^{15}NO_4$ $[M-H]^-$: calcd. 220.1217, found 220.1211.

***N*-(*tert*-Butoxycarbonyl)-L-[^{15}N, $^{13}C_5$]-proline-3,5-dinitrobenzyl ester (2).** A solution of **1** (876 mg, 3.95 mmol) in anhydrous *N*,*N*-dimethylformamide (4 mL) was treated with triethylamine (0.82 mL, 5.93 mmol) and 3,5-dinitrobenzyl chloride (942 g, 4.35 mmol) under an inert atmosphere and was stirred at room temperature for 12 h. The resulting mixture was partitioned between diethyl ether (30 mL) and a 5% aqueous solution of potassium bisulfate (40 mL). The aqueous phase was further extracted with diethyl ether (2x 30 mL) and the combined organic extracts were washed with brine (3x 30 mL). The organic solution was dried over anhydrous magnesium sulfate, filtered and evaporated to dryness. The residue obtained was purified by column chromatography on silica gel eluting with *n*-hexane/ethyl acetate 7:3 to provide **2** as a white solid (1.51 g, 3.76 mmol, 95% yield). HRMS (ESI) $^{13}C_5C_{12}H_{21}{}^{15}NN_2NaO_8$ $[M+Na]^+$: calcd. 424.1364, found 424.1377.

L-[^{15}N, $^{13}C_5$]-proline-3,5-dinitrobenzyl ester hydrochloride (3). A 3 N solution of hydrogen chloride in ethyl acetate (10 mL) was added to compound **2** (1.50 g, 3.74 mmol) and the reaction mixture was stirred at room temperature for 1 h. After completion, the solvent was evaporated to dryness and the resulting residue was taken up in water and lyophilized to provide **3** as a white solid (1.18 g, 3.49 mmol, 94% yield). $[\alpha]_D^{28}$ –30.7 (c = 0.15, water). ^{1}H NMR (DMSO-d_6, 400 MHz) δ: 1.59-2.42 (m, 4H), 2.99 and 3.47 (d, 2H, $J_{(C,H)}$ = 145 Hz), 4.25 and 4.75 (d, 1H, $J_{(C,H)}$ = 150 Hz), 5.42-5.57 (m, 2H), 8.71-8.89 (m, 3H), 9.79 (bs, 1H). HRMS (ESI) $^{13}C_5C_7H_{14}{}^{15}NN_2O_6$ $[M+Na]^+$: calcd. 302.1020, found 302.1013.

High-resolution mass spectra were obtained on a Bruker Microtof-Q spectrometer. ^{1}H NMR spectra were recorded on a Bruker AV-400 instrument at room temperature, using the residual solvent signal as the internal standard; chemical shifts (δ) are expressed in ppm and coupling constants (J) in Hertz. Optical rotation was measured in a JASCO P-1020 polarimeter. Column chromatography was performed using 60 Å (0.04 - 0.063 mm) silica gel from Macherey-Nagel.

Aminoacylation of suppressor $tRNA_{CUA}$ using flexizyme dFx

Flexizyme dFx (5' GGAUC GAAAG AUUUC CGCAU CCCCG AAAGG GUACA UGGCG UUAGG U 3') and aminoacylated $tRNA_{CUA}$ were prepared following the protocol by Goto *et al.*[63] with minor changes. A typical flexizyme reaction contained 50 mM HEPES-KOH pH 7.5, 15.6 μM $tRNA_{CUA}$, 25 μM dFx, 600 mM $MgCl_2$ and 5 - 40 mM [^{15}N, ^{13}C]-ProDBE (L-[^{15}N, $^{13}C_5$]-proline-3,5-dinitrobenzyl ester hydrochloride (**3**)) and was incubated on ice for 4 hours. Loaded suppressor $tRNA_{CUA}$ was precipitated with 300 mM Na-acetate pH 5.2 and 2.5 volumes of 96% EtOH at -80°C. Pellets were washed twice with 70% EtOH and stored at -20°C.

Enzymatic aminoacylation of suppressor $tRNA_{CUA}$

The suppressor $tRNA_{CUA}$ was refolded in 100 mM HEPES-KOH pH 7.5, 10 mM KCl at 70°C for 5 min and a final concentration of 5 mM $MgCl_2$ was added just before the reaction was placed on ice. The refolded $tRNA_{CUA}$ was then aminoacylated with L-[^{15}N, ^{13}C]-proline (CortecNet): 20 μM $tRNA_{CUA}$, 2 μM of our PhProRS preparation, 0.1 mM [^{15}N, ^{13}C]-Pro in 100 mM Na-acetate pH 5.0, 10 mM KCl, 20 mM $MgCl_2$, 1 mM DTT and 10 mM ATP. After incubation at 37°C for 1 hour loaded suppressor $tRNA_{CUA}$ was precipitated with 300 mM Na-acetate pH 5.2 and 2.5 volumes of 96% EtOH at -80°C. Pellets were washed twice with 70% EtOH and stored at -20°C. Successful loading was confirmed by urea-PAGE (6.5% acrylamide 19:1, 8 M urea, 100 mM Na-acetate pH 5.2)[64].

Standard cell-free expression conditions

Lysate was prepared as previously described[40] and based on the *E. coli* strain BL21 Star (DE3)::RF1-CBD_3, a gift from Gottfried Otting (Australian National University, Canberra, Australia)[65]. Cell-free protein expression was performed in batch mode as described by Apponyi *et al.*[66]. The standard reaction mixture consisted of the following components: 55 mM HEPES-KOH (pH 7.5), 1.2 mM ATP, 0.8 mM each of CTP, GTP and UTP, 1.7 mM DTT, 0.175 mg/mL *E. coli* total tRNA mixture (from strain MRE600), 0.64 mM cAMP, 27.5 mM ammonium acetate, 68 μM 1-5-formyl-5,6,7,8-tetrahydrofolic acid (folinic acid), 1 mM of each of the 20 amino acids, 80 mM creatine phosphate (CP), 250 μg/mL creatine kinase (CK), plasmid (16 μg/mL) and 22.5% (v/v) S30 extract. The concentrations of magnesium acetate (5 - 20 mM) and potassium glutamate (60 - 200 mM) were adjusted for each new batch of S30 extract. A titration of both compounds was performed to obtain the maximum yield.

Optimization of cell-free suppression conditions (tRNA concentration and position screen)

To optimize the cell-free reaction for nonsense suppression, different concentrations of loaded $tRNA_{CUA}$ (0 - 40 μM final concentration of total tRNA) were added to the reaction mix and protein expression was followed by sfGFP fluorescence using a plate reader/incubator (Gen5, BioTek Instruments, 485 nm (excitation), 528 nm (emission)). Assays were carried out as triplicates in a

reaction volume of 50 μL dispensed in 96-well plates. The reactions were incubated at 23°C for 5 hours. The same setup was used to probe for possible position specific effects of the amber codon placement on the suppression efficiency. To this end, plasmids of all 13 amber mutants of wild-type H16 were tested at a constant final concentration of 10 μM $tRNA_{CUA}$.

Preparation of NMR samples

Samples for NMR studies were produced at 5 - 10 mL scale and incubated at 23°C and 550 rpm in a thermomixer for 5 hours. Uniformly labeled NMR samples were obtained by substituting the standard amino acid mix with 3 mg/mL [^{15}N, ^{13}C]-ISOGRO®[39] (an algal extract lacking four amino acids: Asn, Cys, Gln and Trp) and additionally supplying Asn, Cys, Gln and Trp (1 mM each).

The final concentration of the fully labeled sample used for 3D experiments was ~80 μM, and the one used for CON experiments was 20 μM.

A poly-Pro sample where only prolines were labeled was prepared by removing Pro from the standard amino acid mix and replacing it with 1 mM [^{15}N, ^{13}C]-Pro. The final concentration of the all-Pro HSQC sample was 9.4 μM, and the CON sample was 47.6 μM.

To produce site-specifically labeled samples 10 μM of [^{15}N, ^{13}C]-Pro suppressor $tRNA_{CUA}$ were added to the reaction. To improve the quality of the spectra by removing the natural abundance peaks of Gln and Pro, all Gln and all Pro (excluding the suppressed residue) were deuterated. Samples prepared with flexizyme-loaded $tRNA_{CUA}$: P34, P38, P47, P55, P58. Samples prepared with enzymatically-loaded $tRNA_{CUA}$: P36, P40, P42, P44, P49, P50, P51, P73.

Purification of H16

The cell-free reaction was thawed on ice and diluted 2-3 fold with buffer A (50 mM Tris-HCl pH 7.5, 500 mM NaCl, 5 mM imidazole) before loading onto a Ni gravity-flow column of 1 mL bed volume (cOmplete™ His-Tag Purification Resin, Sigma Aldrich). The column was washed with buffer B (50 mM Tris-HCl pH 7.5, 1000 mM NaCl, 5 mM imidazole) and the target protein was eluted with buffer C (50 mM Tris-HCl pH 7.5, 150 mM NaCl, 250 mM imidazole). Elution fractions were checked under UV light and fluorescent fractions were pooled and dialyzed against NMR buffer (20 mM BisTris-HCl pH 6.5, 150 mM NaCl) at 4°C using SpectraPor 1 MWCO 6-8 kDa dialysis tubing (Spectrum Labs). Dialyzed protein was then concentrated with 10 kDa MWCO Vivaspin centrifugal concentrators (3500 x g, 4°C) (Sartorius). Protein concentrations were determined by means of fluorescence using an sfGFP calibration curve. Protein integrity was analyzed by SDS-PAGE.

NMR experiments and data analysis

All NMR samples contained final concentrations of 10% D_2O and 0.5 mM 3-trimethylsilylpropane-1-sulfonic acid (DSS) and chemical shifts were referenced with respect to the H_2O signal relative to DSS using the ^{1}H/X frequency ratio of the zero point according to Markley *et al.*[67]. Experiments were

performed at 293 K on a Bruker Avance III spectrometer equipped with a cryogenic triple resonance probe and Z gradient coil. Most SSIL samples were recorded on a spectrometer operating at a ^{1}H frequency of 800 MHz, the only exception being the sample of P38, which was measured at 900 MHz. 3D-NMR experiments were recorded on a 700 MHz spectrometer. ^{15}N-HSQC and ^{13}C-HSQC were acquired in order to determine amide ($^{1}H_{N}$ and ^{15}N) and aliphatic ($^{1}H_{aliphatic}$ and $^{13}C_{aliphatic}$) chemical shifts, respectively. $^{13}C'$ and ^{15}N chemical shifts of prolines were determined by recording $^{13}C'$, ^{15}N-HCaCON spectra. To assign the *cis* peaks of residues influenced by a neighboring proline, 3D HNCaCb, HNCaCO, CbCaCONH and HNCO experiments were acquired.

Spectra acquisition parameters were set up depending on the sample concentration and the magnet strength. All spectra were processed with TopSpin v3.5 (Bruker Biospin) and analyzed using the CCPN-Analysis software[68]. The noise present in the ^{13}C-HSQC spectra of SSIL samples and the error bars associated to the *cis* populations were calculated by using the estimated noise tool of NMRFAM-SPARKY[69]. Three times the noise intensity was considered as the minimum intensity threshold above which peaks could be unambiguously identified. Therefore, when no signal could be detected, intensities were considered to have a value below this threshold. *Cis* populations obtained from 3D experiments were quantified by using *cis* and *trans* peak intensities assigned in different spectra; when possible, multiple neighboring residue peaks were averaged for each proline. From different 3D spectra and multiple neighboring residues, the standard deviations of *cis* percentages for each proline were calculated.

Tripeptide database construction

The tripeptide database was built using the entire Protein Data Bank (PDB). A total of 89,352,861 three-residue fragments (tripeptides) were extracted from the deposited protein structures and classified on the basis of their sequence (8,000 tripeptide classes). The conformations adopted by the residues were assigned using the program DSSP[70], allowing us to filter out fragments corresponding to α-helices and β-strands. After filtering out the secondary structure elements, 32,386,222 tripeptides remained, which corresponds to 36.25% of the total number of tripeptides being included in the coil database (i.e. DSSP types G, B, T, S and blank/C). The database was further filtered by only selecting proline-centred tripeptides derived from high-resolution (< 2 Å) crystallographic structures. A total of 754,308 proline-centred tripeptides was used to derive the *cis*/*trans* populations displayed in Tables S2 and S3. Note that previous statistical analyses of deposited high-resolution structures have established the influence of the sequence context on the *cis*/*trans* proline isomerization, although the effect of the N- and C-flanking residues was analyzed separately[8,52–54]. The position-specific *cis*/*trans* populations for longer fragments (*Xaa*-Pro-Pro-Pro-*Yaa* and *Xaa*-Pro-Pro-Pro-Pro-*Yaa*) were also derived from the entire PDB applying the same structural filters used for the tripeptide database.

Acknowledgements

This work was supported by the European Research Council under the European Union's H2020 Framework Programme (2014-2020) / ERC Grant agreement n° [648030], Labex EpiGenMed, an « Investissements d'avenir » program (ANR-10-LABX-12-01) awarded to PB, and the GPCteR (ANR-17-CE11-0022-01) to NS. The CBS is a member of France-BioImaging (FBI) and the French Infrastructure for Integrated Structural Biology (FRISBI), 2 national infrastructures supported by the French National Research Agency (ANR-10-INBS-04-01 and ANR-10-INBS-05, respectively). Financial support from the TGIR-RMN-THC Fr3050 CNRS, the Spanish MINECO (CTQ2013-40855-R), and Gobierno de Aragón (research group Aminoácidos y Péptidos E19_20R) for conducting the research is gratefully acknowledged.

Supporting Information. Additional NMR spectra (full ^{13}C-HSQC and CON spectra, ^{15}N-HSQC showing *cis* and *trans* peaks of residues neighboring prolines, ^{13}C-HSQCs of all studied site-specifically labeled prolines), details on the optimization of the SSIL approach, synthesis scheme of [^{15}N, ^{13}C]-Pro-DBE, a table summarizing sample concentrations and NMR parameters, a list of all *Xaa*-Pro-*Yaa* tripeptides extracted from the coil database, and data reflecting the influence of the amino acids in position *i-1* and *i+1* on the percentage of *cis* conformations.

References

(1) Morgan, A. A.; Rubenstein, E. Proline: The Distribution, Frequency, Positioning, and Common Functional Roles of Proline and Polyproline Sequences in the Human Proteome. *PLoS One* **2013**, *8* (1), e53785.

(2) Lu, K. P.; Finn, G.; Lee, T. H.; Nicholson, L. K. Prolyl Cis-Trans Isomerization as a Molecular Timer. *Nat. Chem. Biol.* **2007**, *3* (10), 619–629.

(3) Schmidpeter, P. A. M.; Schmid, F. X. Prolyl Isomerization and Its Catalysis in Protein Folding and Protein Function. *J. Mol. Biol.* **2015**, *427* (7), 1609–1631.

(4) Grathwohl, C.; Wüthrich, K. NMR Studies of the Rates of Proline Cis-Trans Isomerization in Oligopeptides. *Biopolymers* **1981**, *20* (12), 2623–2633.

(5) Wedemeyer, W. J.; Welker, E.; Scheraga, H. A. Proline Cis-Trans Isomerization and Protein Folding. *Biochemistry* **2002**, *41* (50), 14637–14644.

(6) Chen, D.; Drombosky, K. W.; Hou, Z.; Sari, L.; Kashmer, O. M.; Ryder, B. D.; Perez, V. A.; Woodard, D. R.; Lin, M. M.; Diamond, M. I.; Joachimiak, L. A. Tau Local Structure Shields an Amyloid-Forming Motif and Controls Aggregation Propensity. *Nat. Commun.* **2019**, *10* (1), 2493.

(7) Mallis, R. J.; Brazin, K. N.; Fulton, D. B.; Andreotti, A. H. Structural Characterization of a Proline-Driven Conformational Switch within the Itk SH2 Domain. *Nat. Struct. Biol.* **2002**, *9* (12), 900–905.

(8) Shen, Y.; Bax, A. Prediction of Xaa-Pro Peptide Bond Conformation from Sequence and Chemical Shifts. *J. Biomol. NMR* **2010**, *46* (3), 199–204.

(9) Alderson, T. R.; Benesch, J. L. P.; Baldwin, A. J. Proline Isomerization in the C-Terminal Region of HSP27. *Cell Stress Chaperones* **2017**, *22* (4), 639–651.

(10) Alderson, T. R.; Lee, J. H.; Charlier, C.; Ying, J.; Bax, A. Propensity for Cis-Proline Formation in Unfolded Proteins. *Chembiochem* **2018**, *19* (1), 37–42.

(11) Weiss, M. S.; Jabs, A.; Hilgenfeld, R. Peptide Bonds Revisited. *Nat. Struct. Biol.* **1998**, *5* (8), 676.

(12) Grathwohl, C.; Wüthrich, K. NMR Studies of the Molecular Conformations in the Linear Oligopeptides H-(L-Ala)n-L-Pro-OH. *Biopolymers* **1976**, *15* (10), 2043–2057.

(13) Yao, J.; Feher, V. A.; Espejo, B. F.; Reymond, M. T.; Wright, P. E.; Dyson, H. J. Stabilization of a Type VI Turn in a Family of Linear Peptides in Water Solution. *J. Mol. Biol.* **1994**, *243* (4), 736–753.

(14) Theillet, F.-X.; Kalmar, L.; Tompa, P.; Han, K.-H.; Selenko, P.; Dunker, A. K.; Daughdrill, G. W.; Uversky, V. N. The Alphabet of Intrinsic Disorder: I. Act like a Pro: On the Abundance and Roles of Proline Residues in Intrinsically Disordered Proteins. *Intrinsically Disord. Proteins* **2013**, *1* (1), e24360.

(15) Ahuja, P.; Cantrelle, F.-X.; Huvent, I.; Hanoulle, X.; Lopez, J.; Smet, C.; Wieruszeski, J.-M.; Landrieu, I.; Lippens, G. Proline Conformation in a Functional Tau Fragment. *J. Mol. Biol.* **2016**, *428* (1), 79–91.

(16) Mateos, B.; Conrad-Billroth, C.; Schiavina, M.; Beier, A.; Kontaxis, G.; Konrat, R.; Felli, I. C.; Pierattelli, R. The Ambivalent Role of Proline Residues in an Intrinsically Disordered Protein: From Disorder Promoters to Compaction Facilitators. *J. Mol. Biol.* **2019**, DOI: 10.1016/j.jmb.2019.11.015.

(17) Jorda, J.; Kajava, A. V. Protein Homorepeats. *Adv. Protein Chem. Struct. Biol.* **2010**, *79*, 59–88.

(18) Chavali, S.; Chavali, P. L.; Chalancon, G.; de Groot, N. S.; Gemayel, R.; Latysheva, N. S.; Ing-Simmons, E.; Verstrepen, K. J.; Balaji, S.; Babu, M. M. Constraints and Consequences of the Emergence of Amino Acid Repeats in Eukaryotic Proteins. *Nat. Struct. Mol. Biol.* **2017**, *24* (9), 765–777.

(19) Lobanov, M. Y.; Galzitskaya, O. V. Occurrence of Disordered Patterns and Homorepeats in Eukaryotic and Bacterial Proteomes. *Mol. Biosyst.* **2012**, *8* (1), 327–337.

(20) Mier, P.; Alanis-Lobato, G.; Andrade-Navarro, M. A. Context Characterization of Amino Acid Homorepeats Using Evolution, Position, and Order. *Proteins* **2017**, *85* (4), 709–719.

(21) Krishnan, K.; Moens, P. D. J. Structure and Functions of Profilins. *Biophys. Rev.* **2009**, *1* (2), 71–81.

(22) Metzler, W. J.; Bell, A. J.; Ernst, E.; Lavoie, T. B.; Mueller, L. Identification of the Poly-L-Proline-Binding Site on Human Profilin. *J. Biol. Chem.* **1994**, *269* (6), 4620–4625.

(23) Ostrander, D. B.; Ernst, E. G.; Lavoie, T. B.; Gorman, J. A. Polyproline Binding Is an Essential Function of Human Profilin in Yeast. *Eur. J. Biochem.* **1999**, *262* (1), 26–35.

(24) Bhattacharyya, A.; Thakur, A. K.; Chellgren, V. M.; Thiagarajan, G.; Williams, A. D.; Chellgren, B. W.; Creamer, T. P.; Wetzel, R. Oligoproline Effects on Polyglutamine Conformation and Aggregation. *J. Mol. Biol.* **2006**, *355* (3), 524–535.

(25) Darnell, G.; Orgel, J. P. R. O.; Pahl, R.; Meredith, S. C. Flanking Polyproline Sequences Inhibit Beta-Sheet Structure in Polyglutamine Segments by Inducing PPII-like Helix Structure. *J. Mol. Biol.* **2007**, *374* (3), 688–704.

(26) Darling, A. L.; Uversky, V. N. Intrinsic Disorder in Proteins with Pathogenic Repeat Expansions. *Molecules* **2017**, *22* (12), 2027.

(27) Rabanal, F.; Ludevid, M. D.; Pons, M.; Giralt, E. CD of Proline-Rich Polypeptides: Application to the Study of the Repetitive Domain of Maize Glutelin-2. *Biopolymers* **1993**, *33* (7), 1019–1028.

(28) Dobitz, S.; Aronoff, M. R.; Wennemers, H. Oligoprolines as Molecular Entities for Controlling Distance in Biological and Material Sciences. *Acc. Chem. Res.* **2017**, *50* (10), 2420–2428.

(29) Cowan, P. M.; McGavin, S. Structure of Poly-L-Proline. *Nature* **1955**, *176* (4480), 501–503.

(30) Newberry, R. W.; Raines, R. T. The N→π* Interaction. *Acc. Chem. Res.* **2017**, *50* (8), 1838–1846.

(31) Choudhary, A.; Gandla, D.; Krow, G. R.; Raines, R. T. Nature of Amide Carbonyl-Carbonyl Interactions in Proteins. *J. Am. Chem. Soc.* **2009**, *131* (21), 7244–7246.

(32) Wilhelm, P.; Lewandowski, B.; Trapp, N.; Wennemers, H. A Crystal Structure of an Oligoproline PPII-Helix, at Last. *J. Am. Chem. Soc.* **2014**, *136* (45), 15829–15832.

(33) Stryer, L.; Haugland, R. P. Energy Transfer: A Spectroscopic Ruler. *Proc. Natl. Acad. Sci. U. S. A.* **1967**, *58* (2), 719–726.

(34) Schuler, B.; Lipman, E. A.; Steinbach, P. J.; Kumke, M.; Eaton, W. A. Polyproline and the "Spectroscopic Ruler" Revisited with Single-Molecule Fluorescence. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102* (8), 2754–2759.

(35) Best, R. B.; Merchant, K. A.; Gopich, I. V; Schuler, B.; Bax, A.; Eaton, W. A. Effect of Flexibility and Cis Residues in Single-Molecule FRET Studies of Polyproline. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104* (48), 18964–18969.

(36) Bermel, W.; Bertini, I.; Felli, I. C.; Lee, Y.-M.; Luchinat, C.; Pierattelli, R. Protonless NMR Experiments for Sequence-Specific Assignment of Backbone Nuclei in Unfolded Proteins. *J. Am. Chem. Soc.* **2006**, *128* (12), 3918–3919.

(37) Bermel, W.; Bertini, I.; Felli, I. C.; Kümmerle, R.; Pierattelli, R. Novel 13C Direct Detection Experiments, Including Extension to the Third Dimension, to Perform the Complete Assignment of Proteins. *J. Magn. Reson.* **2006**, *178* (1), 56–64.

(38) Murrali, M. G.; Piai, A.; Bermel, W.; Felli, I. C.; Pierattelli, R. Proline Fingerprint in Intrinsically Disordered Proteins. *Chembiochem* **2018**, *19* (15), 1625–1629.

(39) Kigawa, T.; Yabuki, T.; Yoshida, Y.; Tsutsui, M.; Ito, Y.; Shibata, T.; Yokoyama, S. Cell-Free Production and Stable-Isotope Labeling of Milligram Quantities of Proteins. *FEBS Lett.* **1999**, *442* (1), 15–19.

(40) Urbanek, A.; Morató, A.; Allemand, F.; Delaforge, E.; Fournet, A.; Popovic, M.; Delbecq, S.; Sibille, N.; Bernadó, P. A General Strategy to Access Structural Information at Atomic Resolution in Polyglutamine Homorepeats. *Angew. Chem. Int. Ed. Engl.* **2018**, *57* (14), 3598–3601.

(41) Urbanek, A.; Elena-Real, C. A.; Popovic, M.; Morató, A.; Fournet, A.; Allemand, F.; Delbecq, S.; Sibille, N.; Bernadó, P. Site-Specific Isotopic Labeling (SSIL): Access to High-Resolution Atructural and Dynamic Information in Low-Complexity Proteins. *Chembiochem* **2019**, cbic.201900583.

(42) Yabuki, T.; Kigawa, T.; Dohmae, N.; Takio, K.; Terada, T.; Ito, Y.; Laue, E. D.; Cooper, J. A.; Kainosho, M.; Yokoyama, S. Dual Amino Acid-Selective and Site-Directed Stable-Isotope Labeling of the Human c-Ha-Ras Protein by Cell-Free Synthesis. *J. Biomol. NMR* **1998**, *11* (3), 295–306.

(43) McColgan, P.; Tabrizi, S. J. Huntington's Disease: A Clinical Review. *Eur. J. Neurol.* **2018**, *25* (1), 24–34.

(44) Pandey, N. K.; Isas, J. M.; Rawat, A.; Lee, R. V; Langen, J.; Pandey, P.; Langen, R. The 17-Residue-Long N Terminus in Huntingtin Controls Stepwise Aggregation in Solution and on Membranes via Different Mechanisms. *J. Biol. Chem.* **2018**, *293* (7), 2597–2605.

(45) Thakur, A. K.; Jayaraman, M.; Mishra, R.; Thakur, M.; Chellgren, V. M.; Byeon, I.-J. L.; Anjum, D. H.; Kodali, R.; Creamer, T. P.; Conway, J. F.; Gronenborn, A. M.; Wetzel, R. Polyglutamine Disruption of the Huntingtin Exon 1 N Terminus Triggers a Complex Aggregation Mechanism. *Nat. Struct. Mol. Biol.* **2009**, *16* (4), 380–389.

(46) Shen, K.; Calamini, B.; Fauerbach, J. A.; Ma, B.; Shahmoradian, S. H.; Serrano Lachapel, I. L.; Chiu, W.; Lo, D. C.; Frydman, J. Control of the Structural Landscape and Neuronal Proteotoxicity of Mutant Huntingtin by Domains Flanking the PolyQ Tract. *Elife* **2016**, *5*, 1–29.

(47) Csizmok, V.; Felli, I. C.; Tompa, P.; Banci, L.; Bertini, I. Structural and Dynamic Characterization of Intrinsically Disordered Human Securin by NMR Spectroscopy. *J. Am. Chem. Soc.* **2008**, *130* (50), 16873–16879.

(48) Chatterjee, A.; Xiao, H.; Schultz, P. G. Evolution of Multiple, Mutually Orthogonal Prolyl-TRNA Synthetase/TRNA Pairs for Unnatural Amino Acid Mutagenesis in Escherichia Coli. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109* (37), 14841–14846.

(49) Murakami, H.; Ohta, A.; Ashigai, H.; Suga, H. A Highly Flexible TRNA Acylation Method for Non-Natural Polypeptide Synthesis. *Nat. Methods* **2006**, *3* (5), 357–359.

(50) Yokoyama, J.; Matsuda, T.; Koshiba, S.; Tochio, N.; Kigawa, T. A Practical Method for Cell-Free Protein Synthesis to Avoid Stable Isotope Scrambling and Dilution. *Anal. Biochem.* **2011**, *411* (2), 223–229.

(51) Tonelli, M.; Singarapu, K. K.; Makino, S.; Sahu, S. C.; Matsubara, Y.; Endo, Y.; Kainosho, M.; Markley, J. L. Hydrogen Exchange during Cell-Free Incorporation of Deuterated Amino Acids and an Approach to Its Inhibition. *J. Biomol. NMR* **2011**, *51* (4), 467–476.

(52) Pal, D.; Chakrabarti, P. Cis Peptide Bonds in Proteins: Residues Involved, Their Conformations, Interactions and Locations. *J. Mol. Biol.* **1999**, *294* (1), 271–288.

(53) Singh, J.; Hanson, J.; Heffernan, R.; Paliwal, K.; Yang, Y.; Zhou, Y. Detecting Proline and Non-Proline Cis Isomers in Protein Structures from Sequences Using Deep Residual Ensemble Learning. *J. Chem. Inf. Model.* **2018**, *58* (9), 2033–2042.

(54) Pahlke, D.; Freund, C.; Leitner, D.; Labudde, D. Statistically Significant Dependence of the Xaa-Pro Peptide Bond Conformation on Secondary Structure and Amino Acid Sequence. *BMC Struct. Biol.* **2005**, *5*, 8.

(55) Ellman, J. A.; Volkman, B. F.; Mendel, D.; Schultz, P. G.; Wemmer, D. E. Site-Specific Isotopic Labeling of Proteins for NMR Studies. *J. Am. Chem. Soc.* **1992**, *114* (20), 7959–7961.

(56) Peuker, S.; Andersson, H.; Gustavsson, E.; Maiti, K. S.; Kania, R.; Karim, A.; Niebling, S.; Pedersen, A.; Erdelyi, M.; Westenhoff, S. Efficient Isotope Editing of Proteins for Site-Directed Vibrational Spectroscopy. *J. Am. Chem. Soc.* **2016**, *138* (7), 2312–2318.

(57) Frömmel, C.; Preissner, R. Prediction of Prolyl Residues in Cis-Conformation in Protein Structures on the Basis of the Amino Acid Sequence. *FEBS Lett.* **1990**, *277* (1–2), 159–163.

(58) Song, J.; Burrage, K.; Yuan, Z.; Huber, T. Prediction of Cis/Trans Isomerization in Proteins Using PSI-BLAST Profiles and Secondary Structure Information. *BMC Bioinformatics* **2006**, *7*, 124.

(59) Ramazzotti, M.; Monsellier, E.; Kamoun, C.; Degl'Innocenti, D.; Melki, R. Polyglutamine Repeats Are Associated to Specific Sequence Biases That Are Conserved among Eukaryotes. *PLoS One* **2012**, *7* (2), e30824.

(60) Mier, P.; Elena-Real, C.; Urbanek, A.; Bernadó, P.; Andrade-Navarro, M. A. The Importance of Definitions in the Study of PolyQ Regions: A Tale of Thresholds, Impurities and Sequence Context. *Comput. Struct. Biotechnol. J.* **2020**, *18*, 306–313.

(61) Darnell, G. D.; Derryberry, J.; Kurutz, J. W.; Meredith, S. C. Mechanism of Cis-Inhibition of PolyQ Fibrillation by PolyP: PPII Oligomers and the Hydrophobic Effect. *Biophys. J.* **2009**, *97* (8), 2295–2305.

(62) Studier, F. W. Protein Production by Auto-Induction in High Density Shaking Cultures. *Protein Expr. Purif.* **2005**, *41* (1), 207–234.

(63) Goto, Y.; Katoh, T.; Suga, H. Preparation of Materials for Flexizyme Reactions and Genetic Code Reprogramming. Protocol (Version 1). *Protoc. Exch.* **2011**, DOI: 10.1038/protex.2011.209.

(64) Walker, S. E.; Fredrick, K. Preparation and Evaluation of Acylated TRNAs. *Methods* **2008**, *44* (2), 81–86.

(65) Loscha, K. V; Herlt, A. J.; Qi, R.; Huber, T.; Ozawa, K.; Otting, G. Multiple-Site Labeling of Proteins with Unnatural Amino Acids. *Angew. Chem. Int. Ed. Engl.* **2012**, *51* (9), 2243–2246.

(66) Apponyi, M. A.; Ozawa, K.; Dixon, N. E.; Otting, G. Cell-Free Protein Synthesis for Analysis by NMR Spectroscopy. In *Structural Proteomics. Methods in Molecular Biology™*; Kobe, B., Guss, M., Huber, T., Eds.; Humana Press, 2008; Vol. 426, pp 257–268.

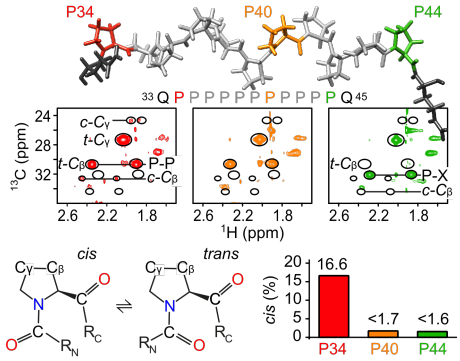
(67) Markley, J. L.; Bax, A.; Arata, Y.; Hilbers, C. W.; Kaptein, R.; Sykes, B. D.; Wright, P. E.; Wüthrich, K. Recommendations for the Presentation of NMR Structures of Proteins and Nucleic Acids. *J. Mol. Biol.* **1998**, *280* (5), 933–952.

(68) Vranken, W. F.; Boucher, W.; Stevens, T. J.; Fogh, R. H.; Pajon, A.; Llinas, M.; Ulrich, E. L.; Markley, J. L.; Ionides, J.; Laue, E. D. The CCPN Data Model for NMR Spectroscopy: Development of a Software Pipeline. *Proteins* **2005**, *59* (4), 687–696.

(69) Lee, W.; Tonelli, M.; Markley, J. L. NMRFAM-SPARKY: Enhanced Software for Biomolecular NMR Spectroscopy. *Bioinformatics* **2015**, *31* (8), 1325–1327.

(70) Kabsch, W.; Sander, C. Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers* **1983**, *22* (12), 2577–2637.

TOC graphic


P34
P40
P44
33 Q P P P P P P P P P P Q 45
24
28
32
^{13}C (ppm)
c-Cγ
t-Cγ
t-Cβ
P-P
c-Cβ
t-Cβ
P-X
c-Cβ
2.6
2.2
1.8
^{1}H (ppm)
cis
trans
cis (%)
20
15
10
5
0
16.6
<1.7
<1.6
P34
P40
P44