SUPPLEMENTARY INFORMATION

Multi-Site Specific Isotopic Labeling Accelerates High-Resolution Structural Investigations of Pathogenic Huntingtin Exon-1

Carlos A. Elena-Real,[a] Annika Urbanek,[a] Xamuel L. Lund,[ab] Anna Morató,[a] Amin Sagar,[a] Aurélie Fournet,[a] Alejandro Estaña,[ac] Tracy Bellande,[d] Frédéric Allemand,[a] Juan Cortés,[c] Nathalie Sibille,[a] Ronald Melki[d] and Pau Bernadó*[a]

[a]- Centre de Biologie Structurale (CBS), Université de Montpellier, INSERM, CNRS, France.

[b]- Institut Laue Langevin, 38000-Grenoble, France.

[c]- LAAS-CNRS, Université de Toulouse, CNRS, 31400 Toulouse, France.

[d]- Institut François Jacob, Molecular Imaging Center (MIRCen), Commissariat à l'Energie Atomique et aux Energies Alternatives (CEA) and Laboratory of Neurodegenerative Diseases, Centre National de la Recherche Scientifique (CNRS), Université Paris-Saclay, Fontenay-aux-Roses, France.

Experimental Methods and Supplemental Figures

Unless specified otherwise, all chemicals were obtained from Sigma-Aldrich (St. Quentin Fallavier, France).

Huntingtin exon-1 constructs

All plasmids were prepared as previously described[1]. Synthetic genes of wild-type huntingtin exon1 (HTTExon1) with 16, 36 and 46 consecutive glutamines (H16, H36 and H46, respectively) or H16 and H36 carrying amber codons (TAG) instead of the glutamine codons e.g. Q20 (H16 Q20), were ordered from Integrated DNA Technologies (IDT, Leuven, Belgium) or GeneArt® (ThermoFisher Scientific, Illkirch, France). H16 double-suppression (H16 Q20Q32) and triple-suppression (H16 Q20Q28Q32) mutants and 9 H36 triple-suppression mutants were ordered: H36 Q18Q36Q50, H36 Q50Q24Q53, H36 Q53Q39Q20, H36 Q20Q47Q52, H36 Q52Q33Q19, H36 Q19Q42Q65, H36 Q65Q45Q21, H36 Q21Q30Q51 and H36 Q51Q27Q18. All genes were cloned into pIVEX 2.3d 3C-sfGFP-His_6, giving rise to pIVEX-H36-3C-sfGFP-His_6 and mutants. The sequence of all plasmids was confirmed by sequencing by GENEWIZ® (Leipzig, Germany).

Standard cell-free expression conditions

Lysate was prepared as previously described[1] and based on the *Escherichia coli* strain BL21 Star (DE3)::RF1-CBD_3, a gift from Gottfried Otting (Australian National University, Canberra, Australia)[2]. Cell-free protein expression was performed in batch mode as described by Apponyi *et al.*[3]. 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, 250 µg/mL creatine kinase, 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.

Preparation and aminoacylation of suppressor $tRNA_{CUA}$

A $tRNA_{CUA}$/tRNA synthetase pair based on the Gln2 tRNA[4] and glutamine ligase GLN4 from *Saccharomyces cerevisiae* was prepared in house as previously described[1]. Briefly, the artificial suppressor $tRNA_{CUA}$ was transcribed *in vitro* and purified by phenol-chloroform extraction. Prior to

use, 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 [^{15}N, ^{13}C]-glutamine (CortecNet, Les Ulis, France) in a standard aminoacylation reaction: 20 µM $tRNA_{CUA}$, 0.5 µM GLN4, 0.1 mM [^{15}N, ^{13}C]-Gln in 100 mM HEPES-KOH pH 7.5, 10 mM KCl, 20 mM $MgCl_2$, 1 mM DTT and 10 mM ATP[1]. After incubation at 37°C for 1 hour GLN4 was removed by addition of glutathione beads and loaded suppressor $tRNA_{CUA}$ was precipitated with 300 mM sodium acetate pH 5.2 and 2.5 volumes of 96% EtOH at -80°C and stored as dry pellets at -20°C. Successful loading was confirmed by urea-PAGE (6.5% acrylamide 19:1, 8 M urea, 100 mM sodium acetate pH 5.2).

Optimization of cell-free suppression conditions

To optimize the CF reaction for multi-site nonsense suppression, different concentrations of loaded $tRNA_{CUA}$ (0–30 µM final concentration of total $tRNA_{CUA}$) were added to the reaction mix. Protein expression was followed by sfGFP fluorescence using a plate reader/incubator (Gen5 v3.03.14, BioTek Instruments, Colmar, France) at 485 nm (excitation) and 528 nm (emission). Assays were carried out in triplicates in a reaction volume of 50 µL dispensed in 96-well plates and incubated at 23°C for 5 h.

Preparation of NMR samples

Samples for NMR studies were produced at 5 mL scale and incubated at 23°C and 450 rpm in a thermomixer for 4 h. Uniformly labeled NMR samples were obtained by substituting the standard amino acid mix with 3 mg/mL [^{15}N, ^{13}C]-labeled ISOGRO®[5] (an algal extract lacking four amino acids: Asn, Cys, Gln and Trp) and additionally supplying [^{15}N, ^{13}C]-labeled Asn, Cys and Trp (1 mM each) and 4 mM Gln (CortecNet, Les Ulis, France). The use of potassium glutamate buffer in CF reaction impairs the labeling of Glu residues, that do not appear in the NMR spectra. To produce site-specifically labeled samples, the standard reaction mixture was slightly modified. Instead of adding 1 mM of each amino acid, proline and glutamine were substituted by deuterated versions (Eurisotop, Saint-Aubin, France) and used at 2 or 4 mM, respectively. 10 µM of [^{15}N, ^{13}C]-Gln suppressor $tRNA_{CUA}$ were added to facilitate single-site suppression, while 20 µM of [^{15}N, ^{13}C]-Gln suppressor $tRNA_{CUA}$ were added to double and triple-site suppression reactions.

HTTExon1 purification

Purification of HTTExon1 with different poly-Q stretch lengths and suppression mutants was performed at 4°C. The cell-free reaction was thawed on ice and diluted 10 fold with buffer A (50 mM

Tris-HCl pH 7.5, 1000 mM NaCl) before incubating it 1 h with 1.5 mL of Ni-resin (cOmplete™ His-Tag Purification Resin). The matrix was packed by gravity-flow and 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, protease inhibitors were added (cOmplete EDTA-free protease inhibitor cocktail) and the sample was dialyzed against NMR buffer (20 mM BisTris-HCl pH 6.5, 150 mM NaCl) at 4°C using SpectraPor 4 MWCO 12-14 kDa dialysis tubing (Fisher Scientific, Illkirch, France). Dialyzed protein was then concentrated with 10 kDa MWCO Vivaspin centrifugal concentrators (3500 xg, 4°C) (Sartorius, Göttingen, Germany). Protein concentrations were determined by means of fluorescence using a sfGFP calibration curve. Final NMR sample concentrations ranged from 10 to 40 µM for WT H16, H36 and H46; and from 4 to 15 µM for SSIL samples. Protein integrity was analyzed by SDS-PAGE.

Production and purification of Hsc70

Recombinant N-terminally-tagged hexahistidine Hsc70 was expressed in BL21(DE3) *E. coli* and purified as described previously[6]. Hsc70, in 50 mM Tris-HCl, pH 7.5, 150 mM KCl was stored at 80°C. The activity of the purified Hsc70 was assessed using a luciferase refolding assay. Briefly, firefly luciferase (Sigma) at 1 mg/mL was denatured in 7 M guanidine hydrochloride for 2 h at room temperature. Denatured luciferase was diluted in refolding buffer (25 mM HEPES, pH 7.5, 50 mM KCl, 5 mM MgCl2, 2 mM DTT) containing or not Hsc70 (2 µM). The resulting mixtures were incubated at 30°C and the refolding activity assessed quantitatively with time in the absence or presence of 2 mM ATP or ADP, by withdrawing 5 µL aliquots and mixing with 95 µL of luciferase assay reagent (Promega) at different time intervals. Luminescence was quantified using a Cary Eclipse fluorescence spectrophotometer (Varian Inc., Palo Alto, CA) in bioluminescence mode at 550 nm. Native luciferase activity was taken as 100%. The ATPase activity of Hsc70 alone or in the presence of unfolded luciferase was also monitored as previously described[7].

NMR experiments and data analysis

All NMR samples contained final concentrations of 10% D_2O and 0.5 mM 4,4-dimethyl-4-silapentane-1-sulfonic acid (DSS). Experiments were performed at 293 K on a Bruker Avance III spectrometer (Bruker Biospin, Wissembourg, France) equipped with a cryogenic triple resonance probe and Z gradient coil, operating at a ^{1}H frequency of 800 MHz. ^{15}N-HSQC and ^{13}C-HSQC were acquired for each sample in order to determine amide ($^{1}H_N$ and ^{15}N) and aliphatic ($^{1}H_{aliphatic}$ and $^{13}C_{aliphatic}$) chemical

shifts, respectively. 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, Wissembourg, France) and analyzed using CCPN-Analysis software v2.4[8]. Chemical shifts were referenced with respect to the H_2O signal relative to DSS using the 1H/X frequency ratio of the zero point according to Markley et *al.*[9]

Random coil chemical shifts were predicted using POTENCI, a pH, temperature and neighbor corrected IDP library (https://st-protein02.chem.au.dk/potenci/)[10]. Secondary chemical shifts (SCS) were obtained by subtracting the predicted value from the experimental one ($SCS=\delta_{exp}-\delta_{pred}$).

Hsc70 titration experiments

H36 Q18-Q36-Q50 and Q21-Q30-Q51 samples were prepared using ISOGRO to ensure the labeling of N17 and PRR. 10 mL CF were prepared and protein purified as explained above. ^{15}N-HSQC spectra of free samples were measured at 8-10 μM. Then, increasing concentrations of Hsc70 were added in order to reach H36:Hsc70 ratios of 1:0.5, 1:1, 1:1.5 and 1:2. Independent ^{15}N-HSQC spectra were acquired for each point of the titration. Signal intensities were measured and corrected according the sample dilution factor.

Model building and Cα chemical shift ensemble optimization

Ensemble models for the two families capturing the conformational influences of the flanking regions, N→C and N←C, were constructed with the algorithm previously described[11], which uses a curated database of three-residue (tripeptide) fragments extracted from high-resolution protein structures. The model building strategy consecutively appends a single residue that can be considered to be either fully disordered or partially structured. For fully disordered residues, amino acid specific ϕ/ψ angles are randomly selected from the database without considering the sequence context. For partially structured residues, the nature and the conformation of the flanking residues on both sides are taken into account when selecting the conformation of the incorporated residue.

Two families of ensembles were built. For the first one (N→C ensembles), starting with the ^{10}AFESLKS[16] region of N17 as partially structured, multiple ensembles of 5,000 conformations were built by successively including an increasing number of glutamines in the poly-Q tract (from F17 to Q53) as partially structured, while the rest of the chain was considered to be fully disordered. An equivalent strategy was followed for the second family of ensembles (N←C ensembles) for which glutamines were considered successively as partially structured from the poly-P tract (from Q53 to Q18). Note that, following this strategy, secondary structural elements present in the flanking regions

are propagated towards the poly-Q tract. Two tripeptide databases were used to generate the conformational ensemble models. Both databases were constructed from the protein domains in the SCOP (Structural Classification of Proteins) repository[12,13] filtered to 95% of sequence identity: (1) an "unfiltered" database containing all the tripeptides extracted from all protein domains, and (2) a "coil" database that only includes tripeptides not participating in α-helices or β-strands. For the N→C ensembles, the best results were obtained when using the "unfiltered" and "coil" databases to sample the partially structured and the fully disordered sections, respectively. For the N←C ensembles, the "coil" database yielded the best results. For the resulting 37 ensembles of each family, and after placing the side chains with the program SCWRL4[14], averaged Cα CSs were computed with SPARTA+[15], and used for the optimization. The optimized ensemble model of H36 was built by reweighting the populations of the pre-computed ensembles, minimizing the error with respect to the experimental Cα CSs by reweighting the populations of the pre-computed ensembles, minimizing the error with respect to the experimental Cα CSs by performing a χ^2 test. For this, we implemented a simple stochastic optimization algorithm inspired by the Monte Carlo Simulated Annealing method. Starting from random values for the populations of the ensembles, at each iteration one of the populations was randomly selected and perturbed within a range of 10% from its current value. This new population was accepted or rejected based on a Metropolis test for the χ^2 value. In our implementation, the temperature of the Metropolis test was a self-adaptive parameter aimed at balancing exploration and exploitation. The algorithm was iterated until convergence, which was estimated based on the evolution of the χ^2 value. The algorithm was run several times from different starting sub-ensemble populations and it converged to approximately the same values for the optimized populations, with deviations of less than 1%

In order to capture the influence of the flanking regions, glutamines within the tract were divided in two groups: those influenced by N17 and those influenced by the poly-P tract, whose chemical shifts were fitted with the N→C and N←C ensembles, respectively. The limit between both families was systematically explored by computing the agreement between the experimental and optimized CSs through a χ^2 value. An optimal description of the complete CS profile was obtained when Q45 was chosen as the last residue structurally connected with N17. Finally, an ensemble of 11,000 conformations was built using the optimized weights and it was used to derive secondary structure population with SS-map[16].

References

1. Urbanek, A., Morató, A., Allemand, F., Delaforge, E., Fournet, A., Popovic, M., Delbecq, S., Sibille, N. & Bernadó, P. (2018). A General Strategy to Access Structural Information at Atomic Resolution in Polyglutamine Homorepeats. *Angew. Chem. Int. Ed. Engl.* **57**, 3598–3601.
2. Loscha, K. V, Herlt, A. J., Qi, R., Huber, T., Ozawa, K. & Otting, G. (2012). Multiple-Site Labeling of Proteins with Unnatural Amino Acids. *Angew. Chem. Int. Ed. Engl.* **51**, 2243–2246.
3. Apponyi, M. A., Ozawa, K., Dixon, N. E. & Otting, G. (2008). Cell-Free Protein Synthesis for Analysis by NMR Spectroscopy. In *Structural Proteomics: High-Throughput Methods*; Kobe, B., Guss, M., Huber, T., Eds.; Humana Press: Totowa, NJ; pp 257–268.
4. Whelihan, E. F. & Schimmel, P. (1997). Rescuing an Essential Enzyme-RNA Complex with a Non-Essential Appended Domain. *EMBO J.* **16**, 2968–2974.
5. Kigawa, T., Yabuki, T., Yoshida, Y., Tsutsui, M., Ito, Y., Shibata, T. & Yokoyama, S. (1999). Cell-Free Production and Stable-Isotope Labeling of Milligram Quantities of Proteins. *FEBS Lett.* **442**, 15–19.
6. Pemberton, S., Madiona, K., Pieri, L., Kabani, M., Bousset, L. & Melki, R. (2011). Hsc70 Protein Interaction with Soluble and Fibrillar Alpha-Synuclein. *J. Biol. Chem.* **286**, 34690–34699.
7. Melki, R., Carlier, M. F. & Pantaloni, D. (1990). Direct Evidence for GTP and GDP-Inorganic Phosphate Intermediates in Microtubule Assembly. *Biochemistry*. **29**, 8921–8932.
8. 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. (2005). The CCPN Data Model for NMR Spectroscopy: Development of a Software Pipeline. *Proteins*. **59**, 687–696.
9. Markley, J. L., Bax, A., Arata, Y., Hilbers, C. W., Kaptein, R., Sykes, B. D., Wright, P. E. & Wüthrich, K. (1998). Recommendations for the Presentation of NMR Structures of Proteins and Nucleic Acids. *J. Mol. Biol.* **280**, 933–952.
10. Nielsen, J. T. & Mulder, F. A. A. (2018). POTENCI: Prediction of Temperature, Neighbor and PH-Corrected Chemical Shifts for Intrinsically Disordered Proteins. *J. Biomol. NMR*. **70**, 141–165.
11. Estaña, A., Sibille, N., Delaforge, E., Vaisset, M., Cortés, J. & Bernadó, P. (2019). Realistic Ensemble Models of Intrinsically Disordered Proteins Using a Structure-Encoding Coil Database. *Structure*. **27**, 381–391.
12. Andreeva, A., Howorth, D., Chothia, C., Kulesha, E. & Murzin, A. G. (2014). SCOP2 Prototype:

A New Approach to Protein Structure Mining. *Nucleic Acids Res.* **42**, D310–D314.

13. Andreeva, A., Kulesha, E., Gough, J. & Murzin, A. G. (2020). The SCOP Database in 2020: Expanded Classification of Representative Family and Superfamily Domains of Known Protein Structures. *Nucleic Acids Res.* **48**, D376–D382.
14. Krivov, G. G., Shapovalov, M. V & Dunbrack Jr., R. L. (2009). Improved Prediction of Protein Side-Chain Conformations with SCWRL4. *Proteins Struct. Funct. Bioinforma.* **77**, 778–795.
15. Shen, Y. & Bax, A. (2010). SPARTA+: A Modest Improvement in Empirical NMR Chemical Shift Prediction by Means of an Artificial Neural Network. *J. Biomol. NMR*. **48**, 13–22.
16. Iglesias, J., Sanchez-Martínez, M. & Crehuet, R. (2013). SS-Map: Visualizing Cooperative Secondary Structure Elements in Protein Ensembles. *Intrinsically Disord. Proteins*. **1**, e25323.

Table S1. Chemical shifts for the glutamines measured in this study and concentrations of the m-SSIL samples used.

	Chemical Shifts (ppm)								Sample information	
Residue	HN	N	Hα	Cα	Hε1	Hε2	Nε		**H36 Triplet**	Concentration (μM)
Q18	8.345	120.376	4.107	57.463	7.539	6.896	112.01		**Q18 Q36 Q50**	4.7
Q19	8.24	120.426	4.162	57.465	7.546	6.887	112.323		**Q24 Q50 Q53**	4.4
Q20	8.279	120.734	4.18	57.336	7.513	-	112.152		**Q20 Q39 Q53**	2.0
Q21	8.281	120.288	4.129	57.256	7.395	6.861	112.7925		**Q20 Q47 Q52**	4.7
Q24	8.312	120.553	4.215	57.188	7.526	6.866	112.13		**Q19 Q33 Q52**	3.2
Q27	8.349	120.698	4.224	57.157	7.538	6.877	112.215		**Q19 Q42 Q65**	6.0
Q30	8.363	120.756	4.233	57.117	7.542	6.88	12.265		**Q21 Q45 Q65**	4.4
Q33	8.372	120.823	4.239	56.964	7.541	6.88	112.22		**Q21 Q30 Q51**	5.5
Q36	8.382	120.872	4.238	56.968	7.543	6.884	112.262		**Q18 Q27 Q51**	5.7
Q39	8.385	120.923	4.253	56.833	7.543	-	112.288			
Q42	8.397	121.01	4.256	56.734	7.59	6.916	112.55			
Q45	8.408	121.113	4.269	56.543	7.547	6.885	112.39			
Q47	8.414	121.171	4.279	56.361	7.548	-	112.412			
Q50	8.439	121.431	4.301	56.095	7.558	6.892	112.508			
Q51	8.45	121.572	4.313	55.92	7.561	6.89	112.53			
Q52	8.47	121.931	4.316	55.78	7.568	6.887	112.6905			
Q53	8.503	123.216	4.605	53.628	7.573	6.906	112.8125			
Q65	8.436	120.584	4.318	55.434	7.613	6.911	113.04			

Figure S1

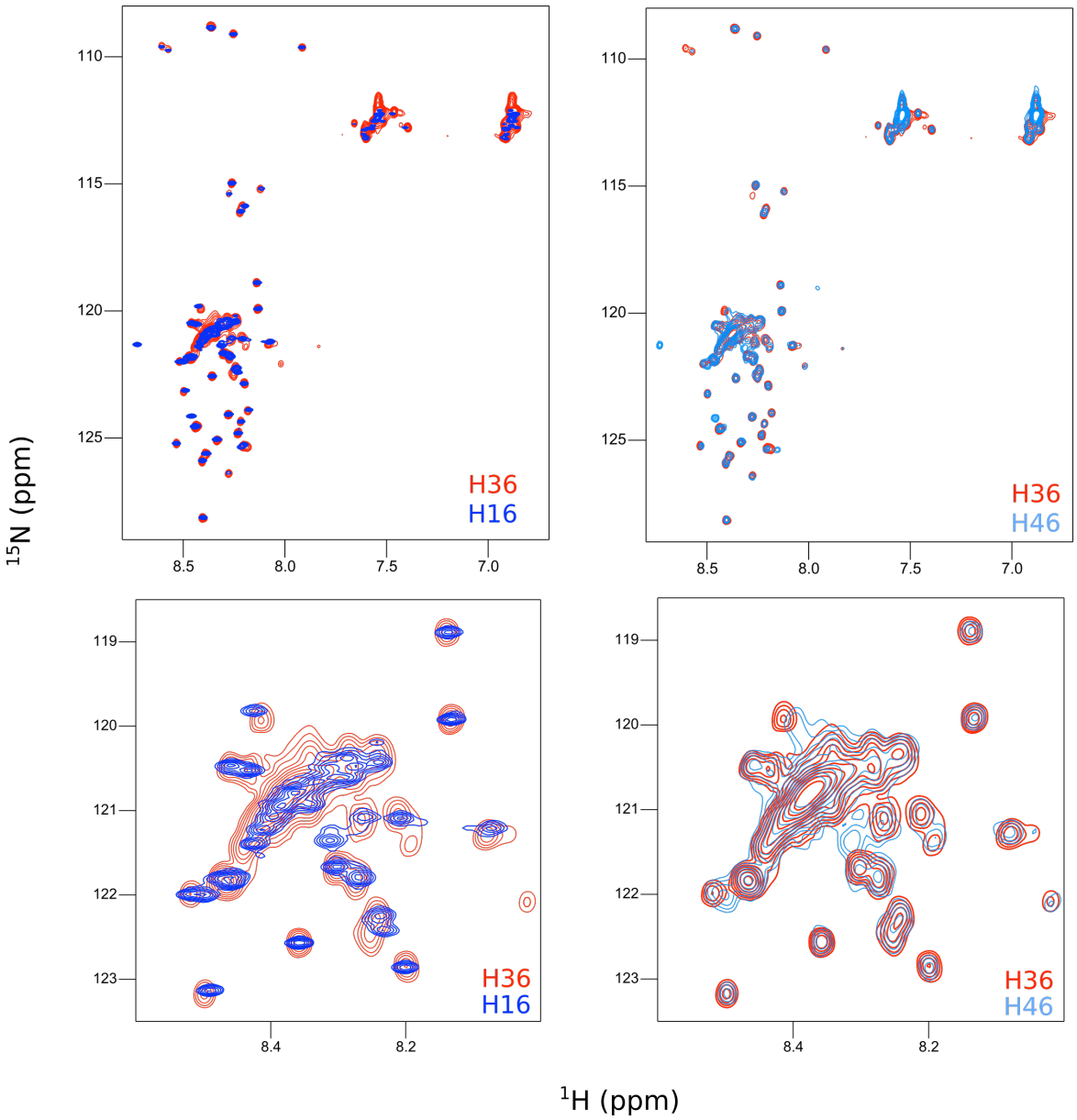

Figure S1. NMR spectrum of fully labeled H36 and comparison with H16 and H46. *Upper panels:* overlays of the ^{15}N-HSQC spectra of fully labeled H36 (red) with H16 (blue, left panel) or with H46 (light blue, right panel). *Lower panels*: zoom of the poly-Q region of the ^{15}N-HSQC spectra shown in upper panels.

Figure S2

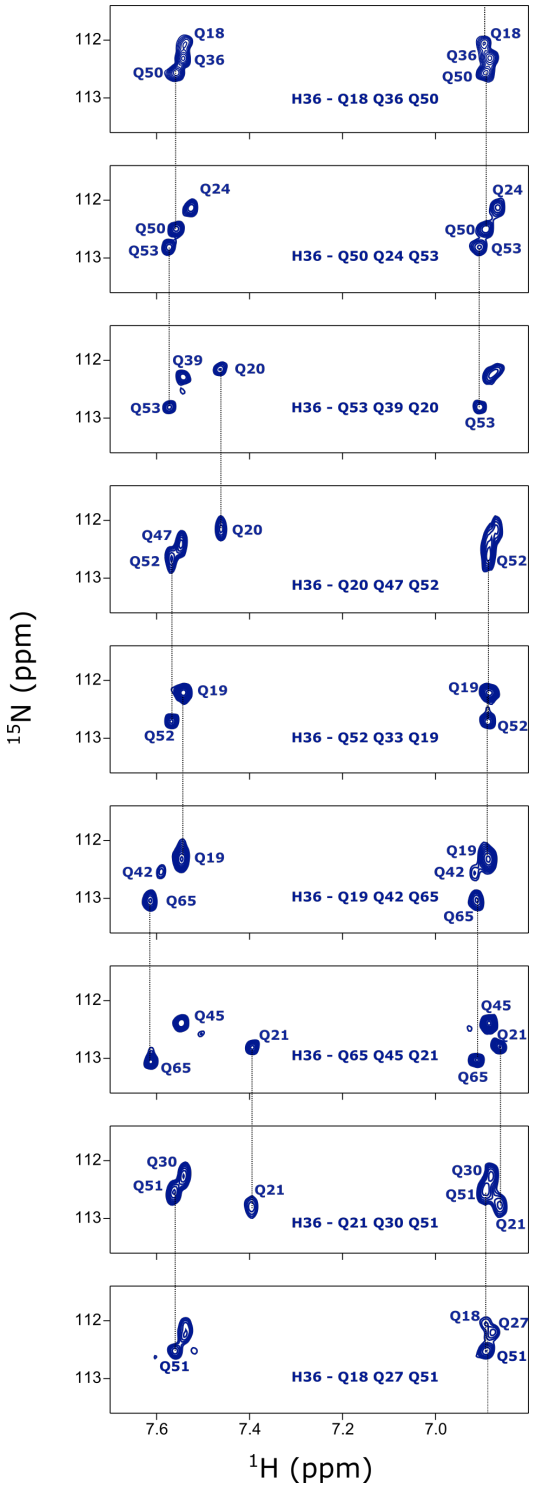

Figure S2. Concatenated analysis of H36 NMR frequencies for poly-Q side chains. Zoom of the ^{15}N-HSQC spectra showing the NHε correlations displayed for the nine H36 triple m-SSIL samples. Repeated residues in different mutants are connected by dotted lines to visualize the signal overlap that allows the sequential assignment.

Figure S3

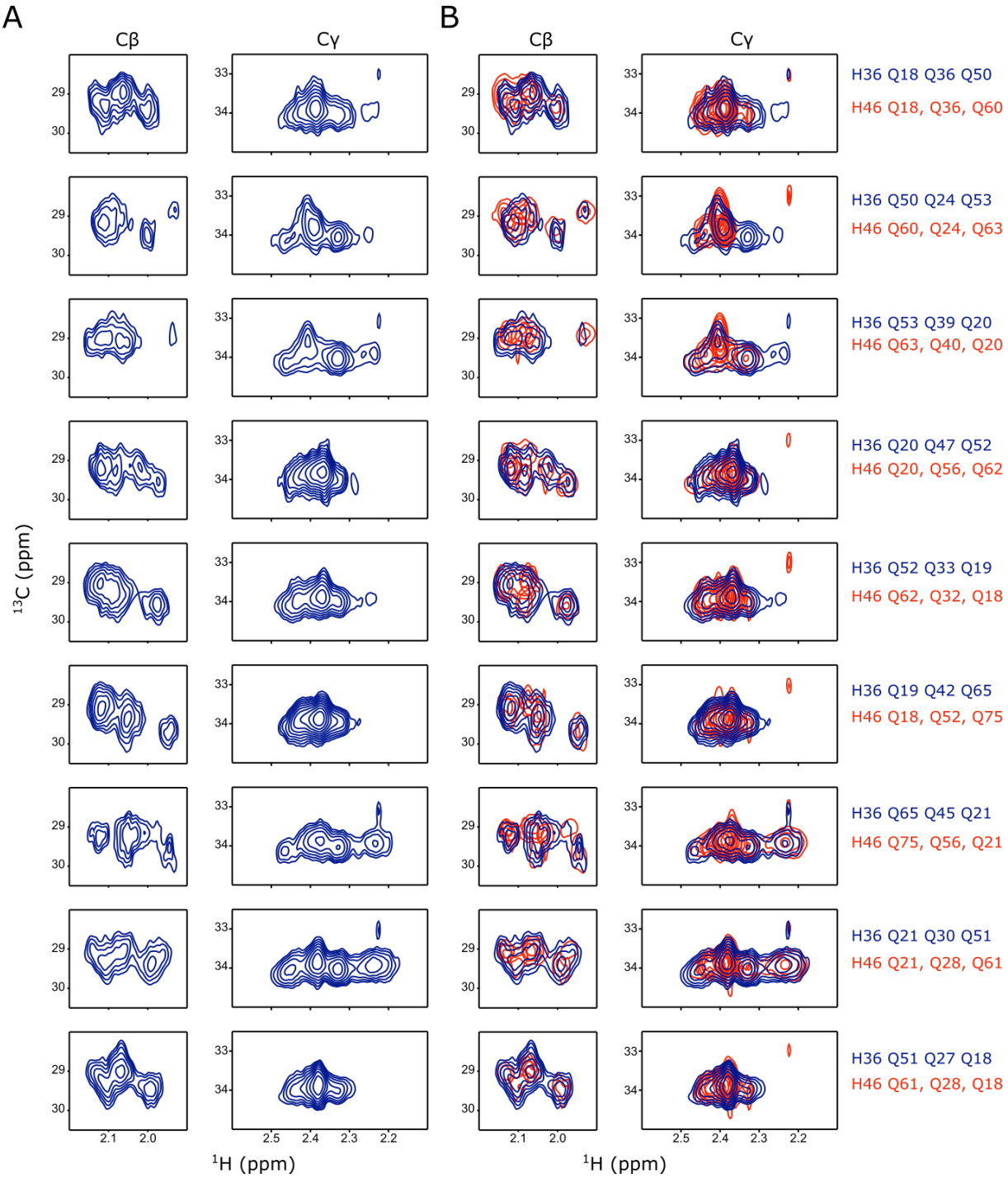

Figure S3. Analysis of H36 NMR frequencies for poly-Q side chains. **(A)** Zoom of the ^{13}C-HSQC spectra showing the Cβ-Hβ (left) and Cγ-Hγ (right) correlations for the nine H36 m-SSIL samples. **(B)** The same spectra than in panel **A** (blue) overlaid to the most similar residues measured for H46 (red) (ref. 20 in the main text). The excellent overlap indicates that the structural properties of the glutamine side chains are equivalent in both HTTExon1 variants.

Figure S4

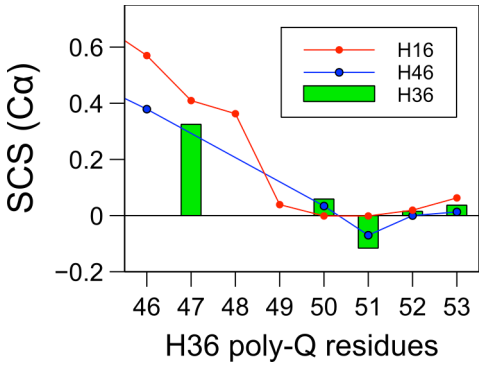

Figure S4. Secondary structure in the poly-Q tract of H36. Cα Secondary Chemical Shift (SCS) profile of H36 (green bars) for the last glutamines of the tract in comparison with SCS data of H16 (reference 19 in Main Text) (red line with dots) and H46 (blue line with dots) (ref. 20 in the main text) aligned from their C-termini. Due to the severe overlap of the Cβ signals, only Cα data was used in the SCS calculation.

Figure S5

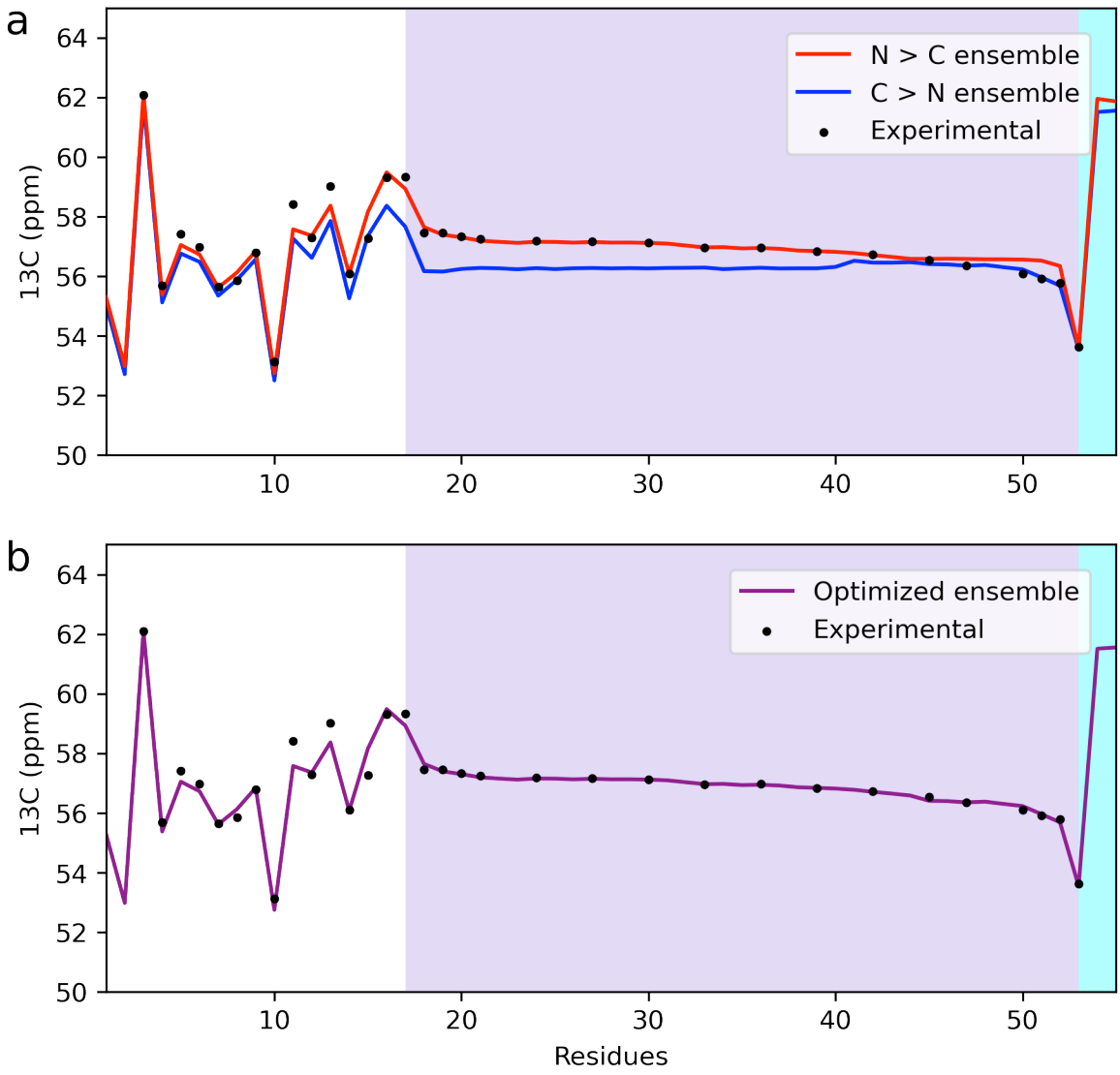

Figure S5. Chemical shift based ensemble refinement of H36. (a) Experimental (black) *vs.* ensemble-optimized (red for N→C and blue for N←C ensembles) chemical shifts for H36. **(b)** The final optimized ensemble was built using the M1-Q45 and the Q45-P103 fragments of the N→C and N←C ensembles, respectively. The poly-Q tract is shaded in purple. Notice that we have measured the chemical shifts for 17 glutamines of the 36-glutamine long tract. No experimental data for prolines are available. The PRR is shaded in blue.

Figure S6

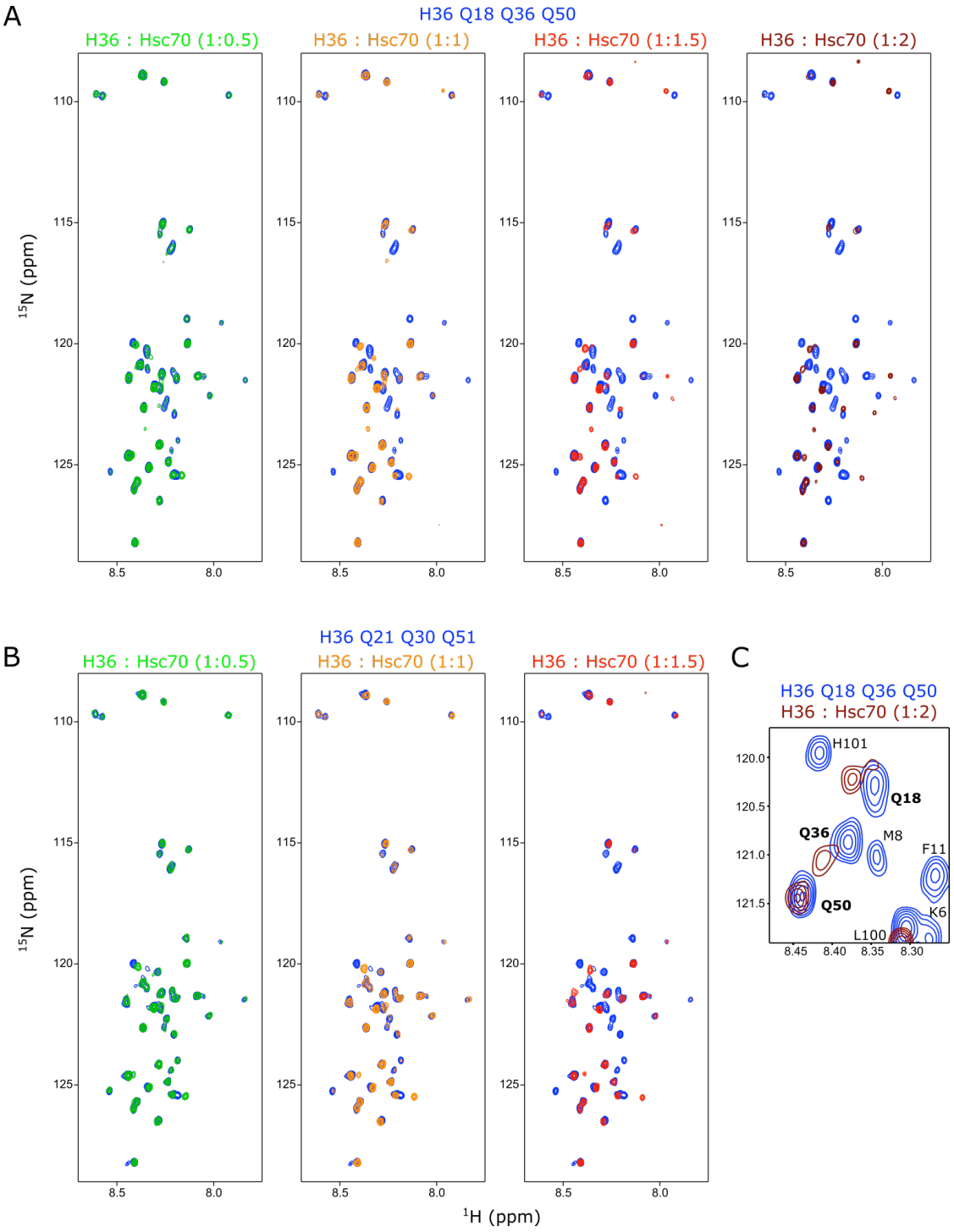

Figure S6. NMR assessment of the interaction of H36 with Hsc70. ^{15}N-HSQC spectra of H36 Q18-Q36-Q50 (**A**) and Q21-Q30-Q51 (**B**) upon addition of increasing concentrations of Hsc70. **C)** Zoom of ^{15}N-HSQC of H36 Q18-Q36-Q50 either free (blue) or upon addition of Hsc70 at ratio 1:2 (dark red).

Figure S7

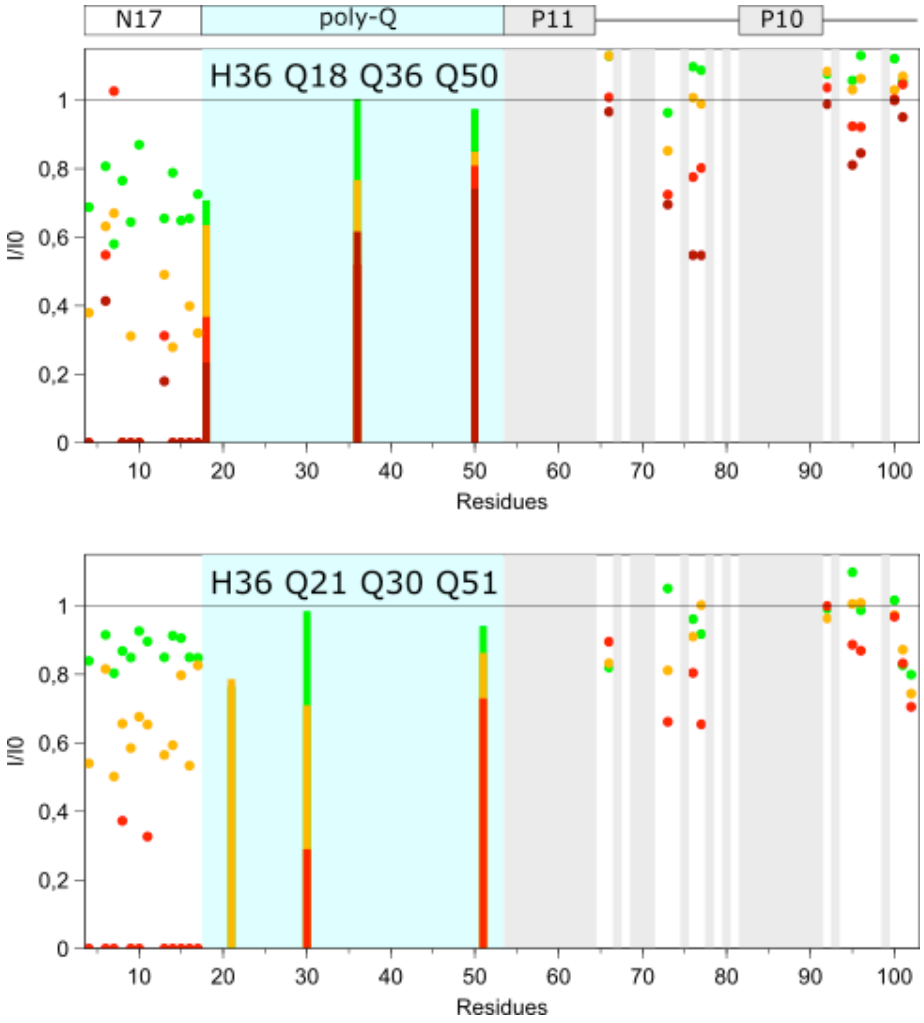

Figure S7. Interaction between H36 and Hsc70. Intensity ratios of H36 Q18-Q36-Q50 (upper panel) and Q21-Q30-Q51 (lower panel) sample peaks with increasing amounts of Hsc70. Green, orange, red and dark red correspond to H36:Hsc70 of 1:0.5, 1:1, 1:1.5 and 1:2 ratios, respectively. Light blue and grey indicate the poly-Q region and prolines, respectively.