Are Long-Range Structural Correlations Behind the Aggregration Phenomena of Polyglutamine Diseases?

Mahmoud Moradi, Volodymyr Babin, Christopher Roland, Celeste Sagui*

Center for High Performance Simulations (CHiPS) and Department of Physics, North Carolina State University, Raleigh, North Carolina, United States of America

Abstract

We have characterized the conformational ensembles of polyglutamine Q_n peptides of various lengths n (ranging from 6 to 40), both with and without the presence of a C-terminal polyproline hexapeptide. For this, we used state-of-the-art molecular dynamics simulations combined with a novel statistical analysis to characterize the various properties of the backbone dihedral angles and secondary structural motifs of the glutamine residues. For Q_{40} (*i.e.*, just above the pathological length $\simeq 36$ for Huntington's disease), the equilibrium conformations of the monomer consist primarily of disordered, compact structures with non-negligible α-helical and turn content. We also observed a relatively small population of extended structures suitable for forming aggregates including β- and α-strands, and β- and α-hairpins. Most importantly, for Q_{40} we find that there exists a *long-range* correlation (ranging for at least 20 residues) among the backbone dihedral angles of the Q residues. For polyglutamine peptides below the pathological length, the population of the extended strands and hairpins is considerably smaller, and the correlations are short-range (at most 5 residues apart). Adding a C-terminal hexaproline to Q_{40} suppresses both the population of these rare motifs and the *long-range* correlation of the dihedral angles. We argue that the long-range correlation of the polyglutamine homopeptide, along with the presence of these rare motifs, could be responsible for its aggregation phenomena.

Citation: Moradi M, Babin V, Roland C, Sagui C (2012) Are Long-Range Structural Correlations Behind the Aggregration Phenomena of Polyglutamine Diseases? PLoS Comput Biol 8(4): e1002501. doi:10.1371/journal.pcbi.1002501

Editor: James M. Briggs, University of Houston, United States of America

Received January 12, 2012; **Accepted** March 18, 2012; **Published** April 26, 2012

Funding: This work was supported by the NSF grants FRG-0804549 and 1021883. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: sagui@ncsu.edu

Introduction

Polyglutamine (polyQ) diseases involve a set of nine late-onset progressive neurodegenerative diseases caused by the expansion of CAG triplet sequence repeats [1]. These repeats result in the transcription of proteins with abnormally long polyQ inserts. When these inserts expand beyond a normal repeat length, the affected proteins form toxic aggregates [2] leading to neuronal death. PolyQ aggregation takes place through a complex multistage process involving transient and metastable structures that occur before, or simultaneously, with fibril formation [3–9]. Experimental findings suggest that the therapeutic target for polyQ diseases should be the soluble oligomeric intermediates, or the conformational transitions that lead to them [9,10], and not the insoluble ordered fibrils. These findings, common to all amyloid diseases [11], have spurred efforts to understand the structural attributes of soluble oligomers and amyloidogenic precursors.

The free energy landscapes of polyQ aggregates display countless minima of similar depth that correspond to a great variety of metastable and/or glassy states. The aggregation kinetics of pure polyQ have been described as a nucleation-growth polymerization process [4–6,12], where soluble expanded glutamine requires a considerable time lag for the creation of a critical nucleus, which then readily converts into a sheet in the presence of a template [13]. However, the "time lag" seems to properly be associated with the formation of the fully aggregated precipitates, since soluble aggregates – sometimes called "protofibrils" – that form during the putative lag phase have been reported [14,15]. The variety of polyQ soluble and insoluble aggregates might correlate with the conformational flexibility of monomeric (non-aggregate single-chain) polyQ regions, which are influenced by the conformations of neighboring protein regions [4,16–18]. One striking example of this conformational wealth – and still a source of controversy– is given by the polyQ expansion in the N-terminal of the huntingtin protein that is encoded in the exon 1 (EX1) of the gene. The N-terminal amino acid sequence consists of a seventeen, mixed residue sequence, the polyQ region of variable length, two polyproline regions of 11 and 10 residues separated by a region of mixed residues, and a C-terminal sequence. Toxicity develops after the polyQ expansion exceeds a threshold of approximately 36 repeats, leading to Huntington's disease. The flanking sequences have been shown to play a structural role in polyQ sequences, both in synthetic and natural peptides, and both in monomeric or aggregate form [4,16,17,19]. In particular, a polyproline (polyP) region immediately adjacent to the C-terminal of a polyQ region has been shown to affect the conformation of the polyQ region; the resulting conformations depend on the lengths of both the polyQ and polyP sequences [16,17,20,21].

In this work, we set out to obtain a conceptual and quantitative understanding of the role played by a polyP sequence that is placed at the C-terminal of a polyQ peptide, which is relevant for the understanding of the behavior of the EX1 segment in the huntingtin protein. Sedimentation aggregation kinetics experiments [17] show

Author Summary

Nine neurodegenerative diseases are caused by polyglutamine (polyQ) expansions greater than a given threshold in proteins with little or no homology except for the polyQ regions. The diseases all share a common feature: the formation of polyQ aggregates and eventual neuronal death. Using molecular dynamics simulations, we have explored the conformations of polyQ peptides. Results indicate that for Q_{40} peptides (*i.e.*, just above the pathological length for Hungtington's disease), the equilibrium conformations were found to consist primarily of disordered, compact structures with a non-negligible α-helical and turn content. We also observed a small population of extended structures suitable for forming aggregates. For peptides below the pathological length, the population of these structures was found to be considerably lower. For longer Q_{40} peptides, we found evidence for long-range correlations among the dihedral angles. This correlation turns out to be short-range for the smaller polyQ peptides, and is suppressed (along with the extended structural motifs) when a C-terminal polyproline tail is added to the peptides. We believe that the existence of these long-range correlations in above-threshold polyQ peptides, along with the presence of rare motifs, could be responsible for the experimentally observed aggregation phenomena associated with polyQ diseases.

that the introduction of a $\mathbf{P}_{10}$ sequence C-terminal to polyQ in synthetic peptides decreases both the rate of formation and the apparent stability of the associated aggregates. The polyP sequence can be trimmed to $\mathbf{P}_6$ without altering the suppression effect, but a $\mathbf{P}_3$ sequence is ineffective. There are no effects when the polyP sequences are attached to the N-terminal or via a side-chain tether [17]. These experiments were complemented with CD spectra for monomeric peptides, where the presence of polyP at the C-terminal of Q_{40} showed remarkable changes in the spectra. Analysis of their data led the authors to propose that addition of the C-terminal $\mathbf{P}_{10}$ sequence does not alter the aggregation mechanism, which is nucleated growth by monomer addition with a critical nucleus of 1 monomer (for Q_{40}), but destabilizes both the α-helical and the (still unknown) aggregation-competent conformations of the monomer. These experimental results were unexpected: although a single proline residue interrupting an amyloidogenic sequence can decrease the propensity of that sequence to aggregate [22,23], Pro replacements in amyloidogenic sequences placed in turns or disordered regions do not alter the aggregate core [23].

Here, we consider monomeric polyQ and polyQ-polyP chains, and quantify changes brought about in the conformations of the polyQ sequences by the addition of the polyP sequences at their C-terminal. In order to assess these changes, one must first characterize the conformation of pure monomeric polyQ in water. Wildly diverse conformations have been postulated experimentally for monomeric polyQ, including a totally random coil, β-sheet, α-helix, and PPII structures. At present there is growing experimental evidence that single polyQ chains are mainly disordered [6,13–15]. The solvated polyQ disorder, however, is different from a total random coil or a protein denatured state. In particular, atomic X-ray experiments [18] show that single chains of polyQ (in the presence of flanking sequences) present isolated elements of α-helix, random coil and extended loop. Single-molecule force-clamp techniques were used to probe the mechanical behavior of polyQ chains of varying lengths spanning normal and diseased polyQ expansions [24]. Under the application of force, no extension was observed for any of the polyQ constructs. Further analysis led the authors to propose that polyQ chains collapse to form a heterogeneous ensemble of globular conformations that are mechanically stable.

Simulations results for the monomer conformation have also been contradictory [25–31]. It is interesting that in the search for soluble prefibrillar intermediates, an α-sheet was proposed to play a role in polyQ toxicity [32,33]. In these molecular dynamics simulations, polyQ monomers of various lengths were found to display transient α-strands of four residues or less. The authors proposed that fibril formation in polyQ may proceed through α strands intermediates [33]. More recently, a molecular dynamics study of hexamers of $Ace-\mathbf{GQ}_8\mathbf{G}-Nme$ in explicit water showed that α-sheet aggregates are very stable (more stable than β-sheets) [34]. These results strongly support the idea that α-sheet may either be a stable, a metastable, or at least a long-lived transient, secondary structure of polyQ aggregates. Coming back to the monomeric polyQ conformation, further simulation evidence [35–38] supports the experimental findings that monomeric polyglutamine of various lengths is a disordered statistical coil in solution. The disorder is inherently different from that of denatured proteins and the average compactness and magnitude of conformational fluctuations increase with chain length [35]. In addition, the coils may present considerable α-helical content [38], but there are acute entropic bottlenecks for the formation of β-sheets.

The molecular dynamics results presented here for single polyQ and polyQ-PolyP chains consisting of 6, 9, 12, 18, 30, and 40 glutamine residues are in qualitative agreement with the experimental and simulation results mentioned above: polyQ is primarily disordered, with non-negligible α-helical content and a small population of other secondary structures including both β and α strands. The addition of polyP reduces the population of the α_R region of Ramachandran plot [39], and increases the population of β and PPII Ramachandran regions for all PolyQ lengths. If one considers secondary structure motifs (i.e., hydrogen-bonds patterns in addition to dihedral angles), the addition of the polyP segment increases the populations of the PPII helices and turns, and decreases the α-helical content of all peptides but Q_{40} (which may have a protective effect against aggregation, as discussed later). The addition of polyP does not change the average radius of gyration of polyQ, but changes the radius of gyration distribution function for Q_{40}, that becomes dependent on the prolyl bond isomerization state. Most importantly, the addition of polyP decreases the population of small β and α strands, and β and α hairpins.

Since the extended strands and hairpins in both β and α forms are found only in a small fraction of the structures, we used a novel statistical measure based on the odds ratio construction [40] to quantify to study the secondary structural propensities [41,42], thereby learning about the possibility of the growth of such secondary structures under nucleation conditions. This study, also supported by more conventional linear correlation analysis, provides evidence that among all the peptides studied here, only Q_{40} exhibits a *long-range* correlation between all glutamine residue pairs that favors formation of both α and β-strands. This correlation is suppressed by the addition of only six proline residues to the C-terminal of the peptide, which suggests a mechanism in which nucleation starts at these scarcely populated secondary structures (mainly β_3, β_4, α_3 and α_4 strands, as well as β-hairpins and α-hairpins) and can only spread through positive correlations in polyQ peptides of approximately 40 residues or longer.

This paper is organized as follows. The *Methods* section details our simulation methodology and analysis. Specifically, we discuss the generalized Replica Exchange scheme used here for enhanced sampling, the simulation details, our clustering techniques to identify the Ramachandran regions and the secondary structural motifs, and the odds ratio construction, used here to study the correlations between residues. In the *Results* section, we present our results with a focus on a statistical analysis of the equilibrium conformations based on (i) Ramachandran regions (ii) secondary structure (iii) correlation analysis and (iv) radius of gyration. A discussion of our results and a short summary of this work is given in the last section.

Methods

In this section, we briefly describe the generalized replica exchange molecular dynamics [41–44] approach used to generate the equilibrium conformations. In addition, we describe our quantification of the secondary structural content, and review the odds ratio [40] construction for correlations between residues. For a more detailed description of our simulation methods and the clustering approach used to classify the secondary structure motifs of the peptides, please see the Supporting Information section.

Sampling Protocol

Room temperature, regular molecular dynamics (MD) simulations are often too computationally limited to carry out a full sampling of the conformational space of a biomolecular system and generate a reliable statistical ensemble. Thus, in order to deal with the sampling issue, we make use of a replica exchange scheme [43,45]. In the replica exhange molecular dynamics (REMD) [43,46] method, one considers several replicas of a system subject to some sort of ergodic dynamics based on different Hamiltonians, and attempts to exchange the trajectories of these replicas at a predetermined rate to increase the barrier crossing rates (*i.e.*, decrease the ergodic time scale). One possibility is to successively increase the temperatures of the replicas [46]. This method, known as parallel tempering, is here referred to as Temperature REMD (T-REMD). Another possibility [43] is to construct the replicas by adding a biasing potential to the original Hamiltonian that acts on some collective variable that describes the slow modes of the system that need "acceleration". This method can be referred to as Hamiltonian REMD (H-REMD). In practice, T-REMD is used to promote the barrier crossing events in a generic way but the use of H-REMD allows one to directly focus on specific slow modes of the system, such as the cis-trans isomerization of proline amino acids which involves a barrier of 10 to 20 Kcal/mol [47]. A combination of the two methods, known as Hamiltonian-Temperature REMD (HT-REMD) [41–44] provides for a practical way to reduce the computational costs associated with REMD sampling, since it facilitates the sampling by both means.

In this work, we used the T-REMD and HT-REMD methods for polyQ and polyQ-polyP peptides, respectively. In the T-REMD method, one replica runs at room temperature and the rest of the replicas run at higher temperatures. Care must be taken with respect to the choice of the number of replicas and their temperatures. The performance of the setting can be checked by monitoring the exchange rate between the neighboring replicas (*i.e.*, with closest temperatures) as well as the ergodic time scale of the "hottest" replica. The equilibrium conformational ensemble is then generated by taking the structures at a predetermined rate from the trajectory of the replica at the lowest (room) temperature.

In the HT-REMD method, the replicas have different biasing potentials. The biasing potential is usually described in terms of a *collective variable* $\sigma = \sigma(\mathbf{r})$, defined as a smooth function of the atomic positions $\mathbf{r} = \mathbf{r}_1, \ldots, \mathbf{r}_N$. The corresponding free energy or potential of mean force (PMF) [48], $f(\xi) = -\mathrm{k}_B T \ln < \delta[\xi - \sigma(\mathbf{r})] >$ (where the angular brackets denote the equilibrium ensemble average), provides for an ideal biasing potential. Indeed, if the biasing potential is exactly $U(\mathbf{r}) = -f[\sigma(\mathbf{r})]$, then the probabilities of different values of the collective variable would all be equal, since there are no barriers present. Although the true free energy $f(\xi)$ is typically unknown in advance, a roughly approximate $f(\xi)$ is often sufficient to improve the sampling considerably in an H-REMD or HT-REMD setting. Such free energies can be computed in a variety of ways [48]. For the polyQ-polyP systems, some of the slow modes originate in the cis-trans isomerization of the prolyl bonds, that occur when polyproline is in solution. We have recently carried out extensive work on proline-rich systems [41,42,44,47,49] and can take advantage of the free energy profiles previously obtained for polyproline of various lengths [44], calculated using the Adaptively Biased Molecular Dynamics (ABMD) [50,51] method. The ABMD method is an umbrella sampling method with a time-dependent biasing potential, which can be used in conjunction with the REMD protocol, by combining different collective variables and/or temperatures on a per-replica basis [43,50]. Currently, the ABMD method has been implemented into the AMBER v.10,11 simulation package [52]. Details of the calculation of the polyproline potentials are given elsewhere [41,42,44,47].

The HT-REMD simulations proceeded in several stages. We recycled the previously computed free energies associated with a collective variable that "captures" the cis-trans transitions of the prolyl bonds of polyproline peptides of different lengths in implicit water at different temperatures.

The collective variable used for these calculations is defined based on the backbone dihedral angle ω of prolyl bonds, $\Omega = \sum_b \cos\omega_b$ (here sum runs over all the prolyl bonds b). The dihedral angle ω takes the values around 0 and 180° for cis and trans conformations, therefore Ω can "capture" different patterns of the cis/trans conformations in any proline-containg peptide. The biasing potentials, transfered from our previous calculations were then refined for the polyQ-polyP peptides using similar simulation settings. Next, several additional replicas running at the lowest temperature T_0 were introduced into the setup. One of these replicas is completely unbiased, and therefore samples the Boltzmann distribution at $T = \mathrm{T}_0$. The other replicas, also at $T = \mathrm{T}_0$, are subject to a reduced bias (*i.e.*, these biasing potentials are scaled down by a constant factor). The purpose of these "proxy" replicas is to ensure adequate exchange rates between the conformations, and thereby enhance the mixing [43]. Data was then taken from the unbiased replica at a suitable, predetermined rate.

Simulation Details

Simulations were carried out for the peptides with sequence $\mathrm{Ace}-(\mathrm{Gln})_n-\mathrm{NH}_2$ (denoted as Q_n) and $\mathrm{Ace}-(\mathrm{Gln})_n-(\mathrm{Pro})_6-\mathrm{NH}_2$ (denoted as $\mathrm{Q}_n\mathrm{P}_6$). These peptides include Q_{40}, $Q_{40}P_6$, Q_{30}, Q_{18}, $Q_{18}P_6$, Q_{12}, $Q_{12}P_6$, Q_9, Q_9P_6, Q_6, and Q_6P_6. In each case, we refer to the ith glutamine and jth proline residues as Q^i and P^j, respectively. The simulations were carried out using the AMBER [52] simulation package with the ff99SB version of the Cornell *et al* force field [53] with an implicit water model based on the Generalized Born approximation (GB) [54,55] including the surface area contributions computed using the LCPO model [56] (GB/SA). For more simulation details, our implementation of the REMD scheme and a discussion of convergence issues, please see the Supporting Information (Text S1).

Secondary Structure

We used the (ϕ, ψ) dihedral angles (see Fig. 1 for their definition) to identify different regions [57] of the Ramachandran map [39]. Table 1 provides the corresponding definition for these regions.

Although this delineates clear regions for the dihedrals of most residues, it turns out that the populations may overlap around the borders. In order to handle this situation, we used a clustering technique as explained in the Supporting Information (Text S1) to classify the conformations, rather than strictly enforcing the sharp boundaries between the defined regions.

Although the backbone dihedral angles of all the residues forming a right-handed α-helix fall into the α_R region of Ramachandran map, many of the residues in this region do not actually form α-helices. As a matter of fact, several other secondary structural motifs, such as 3_{10} and π helices as well as random coil and turn are characterized by or may involve backbone dihedral angles falling in the same region. An interesting example is provided by polyglutamine itself. It has been suggested recently [32–34] that an α-sheet, whose backbone dihedral angles alternate between the α_R and α_L helical regions, can be a stable, metastable, or at least a long-lived transient secondary structure in oligomers.

In general, for a residue to be considered to belong to a given secondary structure, it is not enough to identify the Ramachandran region of its dihedral angles. Thus, we used the secondary structure prediction program DSSP [58,59] that uses not only the backbone diheral angles, but also the inter-residual hydrogen bonding as well as the relative position of the C_α atoms to identify secondary structural motifs. For our peptides, the DSSP secondary structures with highest probabilities were: (i) helices, including α and 3_{10} types, (ii) turns, including H-bonded turns and bends, (iii) coils. There are also isolated residues involved in β bridges and extended strands, participating in the β ladders with small probabilities. Since DSSP does not specifically identify isolated α or β strands (*i.e.*, strands not H-bonded to another strand of their type) or α hairpins, we used a combination of H-bonding results from DSSP analysis and the Ramachandran regions from the clustering analysis to define β and α strands and hairpins. A β_N strand is defined here as at least N adjacent residues all falling into the β region of Ramachandran plot. A β_N strand is referred to as isolated if none of its N residues is H-bonded. A β hairpin is defined as two adjacent β_3 strands with a turn in between and at least one H-bond between the two strands. The turn between the two strands of a hairpin could be H-bonded or not and is of any length but it has to have the geometrical form of a turn, (*i.e.*, identified as bend by DSSP). Each of the two strands has at least three adjacent residues in β region to ensure the structure is relatively extended. At least one of these three β residues are H-bonded to another β residue in the other strand. We define an α_{RL} repeat as two adjacent residues, whose backbone dihedral angles alternate between α_R and α_L regardless of the order (*i.e.*, this includes both $\alpha_R\alpha_L$ and $\alpha_L\alpha_R$). An α_N strand is formed from N adjacent residues, involving N alternating α_R and α_L repeats. In this definition, an α_3 strand is either $\alpha_R\alpha_L\alpha_R$ or $\alpha_L\alpha_R\alpha_L$ and an α_4 strand is either $\alpha_R\alpha_L\alpha_R\alpha_L$ or $\alpha_L\alpha_R\alpha_L\alpha_R$ but not $\alpha_L\alpha_R\alpha_R\alpha_L$. An isolated α strand is defined as an α strand not H-bonded to another strand, and the α hairpin is defined as two adjacent α_3 strands with a turn in between and at least one H-bond between the two strands, similar to the β hairpin. Another relatively extended secondary structure is PPII that is defined here as adjacent residues whose dihedral angles fall into the PPII region of Ramachandran plot. A PPII_N structure, is defined as a structure having N adjacent PPII residues. A summary of these secondary structures is given in Table 1.

Finally, we determined the type of turn from both the DSSP analysis and our Ramachandran region clustering analysis. DSSP distinguishes between H-bonded turns and geometrical bends that do not involve any H-bonding. The DSSP analysis can be also used to identify β and γ types based on the number of residues involved, which is 4 and 3 respectively. The dihedral angles of the two middle residues of β turns (*i.e.* the second and the third residues) can be used to partition β turns into more types such as I, I′, II, II′, etc. but we will only consider type I-β that involves an $\alpha_R\alpha_R$ sequence and the "other" type β turns that involve other combinations of dihedral angles. Since the population of "other" combinations is relative small, we group these all together.

Odds Ratio

To quantify how the secondary structures of Gln residues influence each other we made use of the odds ratio (OR)

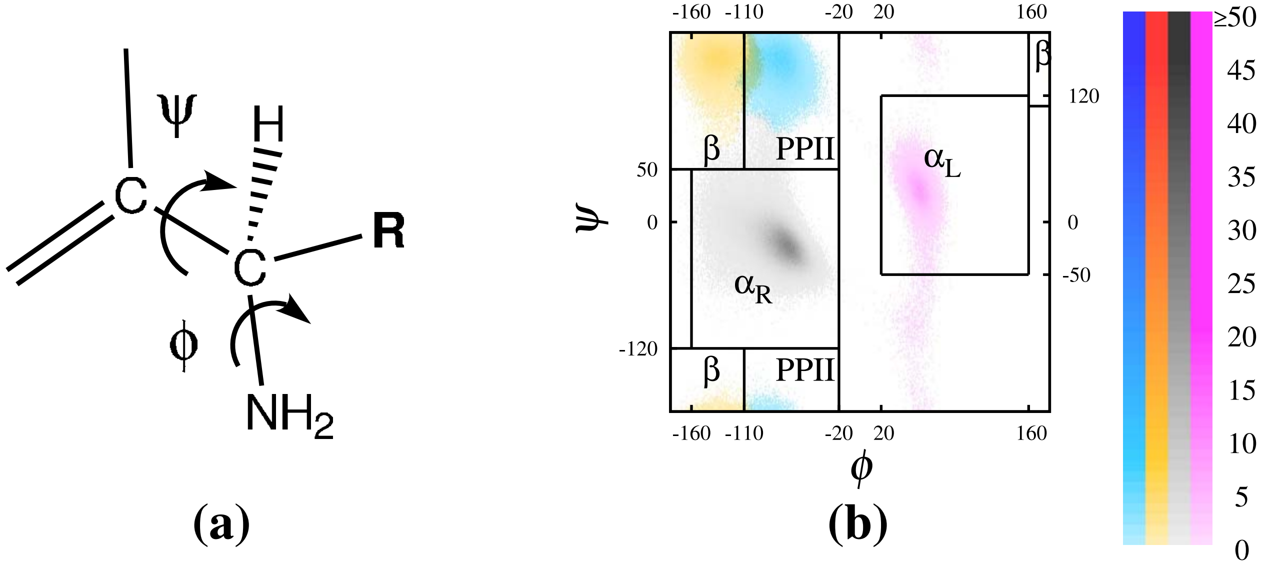

Figure 1. (a) Schematic of amino acid backbone dihedrals ϕ and ψ, and (b) a corresponding Ramachandran plot. In a typical Ramachandran plot of a glutamine residue, each pixel represents a $1^\circ \times 1^\circ$ bin, whose intensity represents its relative population, ranging from 1,2,. . .,9, and 10 or more conformations, sampled in our simulations. Blue, yellow, grey, and pink clusters identify PPII, β, α_R, and α_L regions, respectively.

doi:10.1371/journal.pcbi.1002501.g001

Table 1. Secondary structure definitions.

Ramachandran regions	
α_R	$-160^\circ < \phi < -20^\circ, -120^\circ < \psi < 90^\circ$
α_L	$20^\circ < \phi < 160^\circ, -50^\circ < \psi < 110^\circ$
PPII	$-110^\circ < \phi < -20^\circ$,($90^\circ < \psi < 180^\circ$ or $-180^\circ < \psi < -120^\circ$)
β	$-180^\circ < \phi < -110^\circ$,($90^\circ < \psi < 180^\circ$ or $-180^\circ < \psi < -120^\circ$) or $160^\circ < \phi < 180^\circ$,$120^\circ < \psi < 180^\circ$
extended secondary structures	
α_N strand	N or more adjacent residues, alternating in α_R and α_L regions
β_N strand	N or more adjacent residues, all in β region
$\alpha(\beta)$ isolated strand	$\alpha(\beta)$ strand, not H-bonded to any other $\alpha(\beta)$ strand
α hairpin	α_3- turn $-\alpha_3$ with at least one H-bond between the two strands
β hairpin	β_3- turn $-\beta_3$ with at least one H-bond between the two strands

For a detailed description see *Methods*.
doi:10.1371/journal.pcbi.1002501.t001

construction [40–42]. The OR is a descriptive statistic that measures the strength of association, or non-independence, between two binary values. The OR is defined for two binary random variables (denoted as X and Y) as:

$$\mathrm{OR} = \frac{\mathrm{p}_{11}\mathrm{p}_{00}}{\mathrm{p}_{10}\mathrm{p}_{01}}, \tag{1}$$

where $\mathrm{p}_{ab} = \mathrm{p}(X=a, Y=b)$ is the joint probability of the $(X=a, Y=b)$ event (with a and b taking on binary values of 0 and 1). For the purposes of this study, we can think of X and Y as being some characteristic properties describing the conformations of different residues. For example, the variables could be assigned values of 1 or 0 depending on whether the backbone dihedral angles of corresponding residue falls into the β region of Ramachandran plot or not. We denote this definition of OR as OR_β. Similarly one can define $\mathrm{OR}_{\alpha_{RL}}$ based on the involvement of residues in α_{RL} repeats. In this case, to define the $\mathrm{OR}_{\alpha_{RL}}$ of two given residues x and y, the probabilities $p(X,Y)$ are defined such that the variables X and Y take the values 1 or 0 depending on whether or not the corresponding residue is involved in an α_{RL} repeat as defined in the last subsection. For instance, $X=1$ if and only if residue x either is in the α_R region and is neighboring a residue in the α_L region, or it is in the α_L region and is neighboring a residue in the α_R region. Note that in general, to calculate the $\mathrm{OR}_{\alpha_{RL}}$ of two residues, dihedral angles of not only the two residues but also their neighbors are needed, *i.e.*, up to 6 residues could be involved.

The usefulness of the OR in quantifying the influence of one binary random variable upon another can be readily seen. If the two variables are statistically independent, then $\mathrm{p}_{ab} = \mathrm{p}_a\mathrm{p}_b$ so that $\mathrm{OR}=1$. In the opposite extreme case of $X=Y$ (complete dependence) both p_{10} and p_{01} are zero, and the OR is infinite. Similarly, for $X=\overline{Y}$ $\mathrm{p}_{00}=\mathrm{p}_{11}=0$ rendering $\mathrm{OR}=0$. To summarize, an OR of unity indicates that the values of X are equally likely for both values of Y (*i.e.*, $Y=1{,}0$, X and Y are therefore independent); an OR greater than unity indicates that $X=1$ is more likely when $Y=1$ (X and Y are positively correlated), while an OR less than unity indicates that $X=1$ is more likely when $Y=0$ (X and Y are negatively correlated).

It is convenient to recast the log of the OR in terms of free energy language. If one expresses the probability of the $(X=x, Y=y)$ events in terms of a free energy G_{xy}:

$$\mathrm{p}_{xy} \propto \mathrm{e}^{-\mathrm{G}_{xy}/\mathrm{k}_B T}, \tag{2}$$

then the ratio of probabilities $\mathrm{p}_{xy}/\mathrm{p}_{xz}$ translates into a free energy difference:

$$\ln\frac{\mathrm{p}_{xy}}{\mathrm{p}_{xz}} = -\left(\mathrm{G}_{xy} - \mathrm{G}_{xz}\right)/\mathrm{k}_B T. \tag{3}$$

Clearly, the logarithm of the OR then maps onto the difference of those differences, *i.e.*,

$$\Delta\Delta G = \mathrm{k}_B T \ln \mathrm{OR}. \tag{4}$$

For the case of statistically independent properties, $\Delta\Delta G = 0$; otherwise, this quantity takes on either positive or negative values, whose magnitude depends on the mutual dependence between the two variables. The standard error in its asymptotic approximation is:

$$SE(\Delta\Delta G) = \mathrm{k}_B T\sqrt{\frac{1}{\mathrm{n}_{11}} + \frac{1}{\mathrm{n}_{00}} + \frac{1}{\mathrm{n}_{10}} + \frac{1}{\mathrm{n}_{01}}}. \tag{5}$$

in which n_{xy} is the total number of independent $(X=x, Y=y)$ events sampled. While this development may be perceived as purely formal, the use of an OR analysis couched in terms of free energy language provides for a useful and intuitive measure of the inter-residual correlations, as has been illustrated before [41,42].

In this work, our OR-based correlation analysis is supported by the conventional linear correlation analysis. We have used the correlation coefficient (also know as cross-correlation or Pearson correlation) of ψ dihedral angles of glutamine residues to measure the correlation of glutamine residues in different situations. We emphasize that in the context of secondary structural propensities, the odds ratio analysis is more powerful than the correlation coefficient since it eliminates the noise associated with the dihedral angles. This noise may dominate the linear correlation results such that even substantial correlations may be completely ignored. The OR-based correlation analysis, combined with the clustering technique explained here takes into account both nonlinearity and multivariate components of amino acid correlations in a peptide chain, although in some particular cases a conventional univariate linear correlation may reveal a correlation as we will report in the results. In the context of this paper, the multivariate component is particularly evident when the correlation of α_{RL} repeats is considered, since this may involve ϕ and ψ angles of up to six residues for each single odds ratio calculation.

Results

We generated 5×10^5 equilibrium structures of the Q_{40} and $Q_{40}P_6$ peptides, 2×10^5 structures of Q_{30}, Q_{18}, and $Q_{18}P_6$, and 10^5 structures of Q_{12}, $Q_{12}P_6$, Q_9, Q_9P_6, Q_6, and Q_6P_6 peptides at 300 K to compute the probabilities of different secondary structural motifs and thereby characterize the conformational ensemble of these peptides.

Here, we present our results in terms of (i) the regions of the Ramachandran map occupied by each individual glutamine residue, (ii) the secondary structures identified based not only by the backbone dihedral angles but also by the inter-residual hydrogen bonds and positions of the C_α atoms, (iii) a correlation analysis on the dihedral angles of glutamine residues, and (iv) the ensemble distribution of the radius of gyration, describing the overall compactness of the structures. Figures 1–8 (and Figures S1, S2, S3) and Tables 2–3 (and Table S1) summarize these results.

Ramachandran Regions

Figure 1b shows the Ramachandran plot of a typical glutamine residue, for which the clusters in the different regions are computed according to the protocol described in the *Methods* section. Four clusters can be identified in these plots including PPII (blue), β (yellow), α_R (gray), and α_L (pink). Figures S2 and S3 show the Ramachandran plots of all 40 glutamine residues of both Q_{40} and $Q_{40}P_6$. Considering these, as well as similar plots for other peptides (not shown here), we observe the following trends: (i) The dominant region of most residues is the α_R cluster that is present in all residues, except for the glutamines immediately followed by a proline, for which this region is precluded; (ii) PPII and β clusters are present in almost all residues; (iii) The α_L cluster is present in more than half of the residues but its population is often very small; (iv) Compared to Q_{40}, $Q_{40}P_6$ displays regions with higher non-α_R intensities, particularly for the α_L cluster (see Q^{14}, Q^{20}, Q^{21}, and Q^{40}).

Figure 2 plots the percent population of the β, PPII, and α_{RL} regions of glutamine residues (top, middle and bottom rows, respectively) in terms of the residue number. The left column shows results for Q_{18} [red] and $Q_{18}P_6$ [blue] and the right column for Q_{40} [red] and $Q_{40}P_6$ [blue]. Table 2 presents the population of the different Ramachandran regions (averaged over all glutamine residues) and the α_{RL} repeats, the secondary structure motifs, and the "extended structures" including hairpins. The residue populations in the Ramachandran plot show that, on average, 67–87% of the residues are in the α_R region of the Ramachandran plot, 5–13% of the residues are in the PPII region and 5–17% of the residues are in the β region. The PPII and β regions are almost always equally probable, as can be seen in Figs. 2,S2,S3. The lowest population belongs to the α_L region, comprising only 3–6% although in certain residues it could be as high as 38% as, for instance, in Q^{20} in $Q_{40}P_6$ where the content of α_L correlates with the presence of turns. The addition of P_6 decreases the population of the α_R Ramachandran region and increases that of the β and PPII regions, while leaving the small population of α_L approximately invariant. In Q_nP_6 peptides, proline residues are excluded from the statistical analysis so that only Q residue propensities are compared (for instance, when we state that the average helical content of $Q_{40}P_6$ is 43%, it means that 43% of *all Q residues* are in a helix – the P residues are not counted in the statistic).

Figure 2 shows that the populations of the PPII and β regions are always higher at the two ends of the polyQ peptides, particularly at the C-terminal. When a short proline segment is added at the C-terminal of polyQ, the population of these regions in the neighboring glutamines increases even more. For Q_n peptides shorter than $n<18$ (not shown here), the population of the PPII-β region decreases in the middle of the peptide, but for Q_{18} (red line) we see a small peak in the middle of the peptide for both PPII and β regions. In Q_{40}, we have two small peaks (rather than a single peak) centered around residues 13 and 25 for both the β and PPII regions. The presence of the prolines at the C-terminal of a polyglutamine can drastically alter the population distribution. Fig. 2 shows that the few relatively wide peaks of the β-PPII regions in both Q_{18} and Q_{40} are replaced by several narrow peaks of larger heights. Regarding the residues involved in α_{RL} repeats, one can see from Fig. 2e,f that the distribution of these repeats throughout these peptides depends both on the position of glutamine residues and the presence or absence of the C-terminal prolines although, as seen in Table 2, the average α_{RL} content is similar (6–7%) in all four peptides: Q_{40}, $Q_{40}P_6$, Q_{18}, and $Q_{18}P_6$. We note that the distribution of α_{RL} content in the peptide is mostly determined by the α_L content as the α_R content is abundant in these peptides and most α_L residues are involved in an α_{RL} repeat. One can compare Fig. 2e,f with Figs. S2,S3 and observe similar behaviour, *i.e.*, the residues with high α_{RL} content (Fig. 2e,f) have more intense α_L clusters (pink clusters in Figs. S2,S3).

Secondary Structure

When one considers not only the backbone dihedral angles *i.e.*, the (ϕ,ψ) regions occupied by individual glutamine residues, but also the inter-residual hydrogen bonding and the relative positions of the C_α atoms, one can identify different secondary structures, particularly α-helical segments in many of the sampled conformations. Short 3_{10} helices are also possible but the majority of the residues are either in a turn or a coil conformations according to both DSSP [58,59] and STRIDE [60] analysis. Figure 3 plots the helical, turn, and coil content of the individual glutamine residues against their residue numbers for Q_{18}, $Q_{18}P_6$, Q_{40}, and $Q_{40}P_6$. Figure 4 shows plots of select conformations of Q_{40} and $Q_{40}P_6$ peptides, as generated by VMD [61] using STRIDE [60] for the secondary structure assignment. Table 2 lists the population of helix, turn, and "other" secondary structures as obtained from DSSP, averaged over all residues. The "other" secondary structure category includes mainly what DSSP identifies as "loop or irregular" – sometimes called "coil" in other programs – but which may also include a very small population of other secondary structures such as extended β strand and "isolated β-bridge". We use the protocols explained in *Methods* section to further identify these, as well as other extended structures (Tables 2 and 3).

When the population of residues in the α_R region is compared to the actual helical content, one realizes that the majority of the residues in the α_R region do not form α_R or any other type of helices. Many of these residues in the α_R region are followed and/or preceded by a residue in a different Ramachandran region, such as α_L, as discussed in the previous subsection, forming an α_{RL} repeat. Similarly an α_{RL} repeat does not necessarily form an α strand. Table 2 gives the population of the structures (or conformations) having at least one segment in one of the extended conformation forms, as defined in *Methods* section, including β and α strands either in the isolated form of length 3 (or length 4 in parenthesis) or in the hairpin form as well as PPII structures of length 3 (or length 4 in parenthesis). Note that unlike the other populations in part (a) and (b) in Table 2, the population of extended secondary structures in part (c) is not averaged over the residues. Instead, we counted all the conformations having *at least one* such secondary structures in the polyQ portion of the molecule and divided this number by the total number of sampled conformations. These structures are less common than helices or turns, but they are possible and form a small subpopulation of the secondary structures. Indeed, one can see that a non-negligible portion of the structures has at least one such segment. In particular, isolated α_3 strands are quite common, although they may simply be considered as part of a random coil. The isolated β_3 and PPII strands form the second most populated extended structures. Similarly, these structures may also be considered as part of a random coil. However $PPII_4$, α_4 and β_4 strands form extended structures that are unlikely to be considered random coil elements.

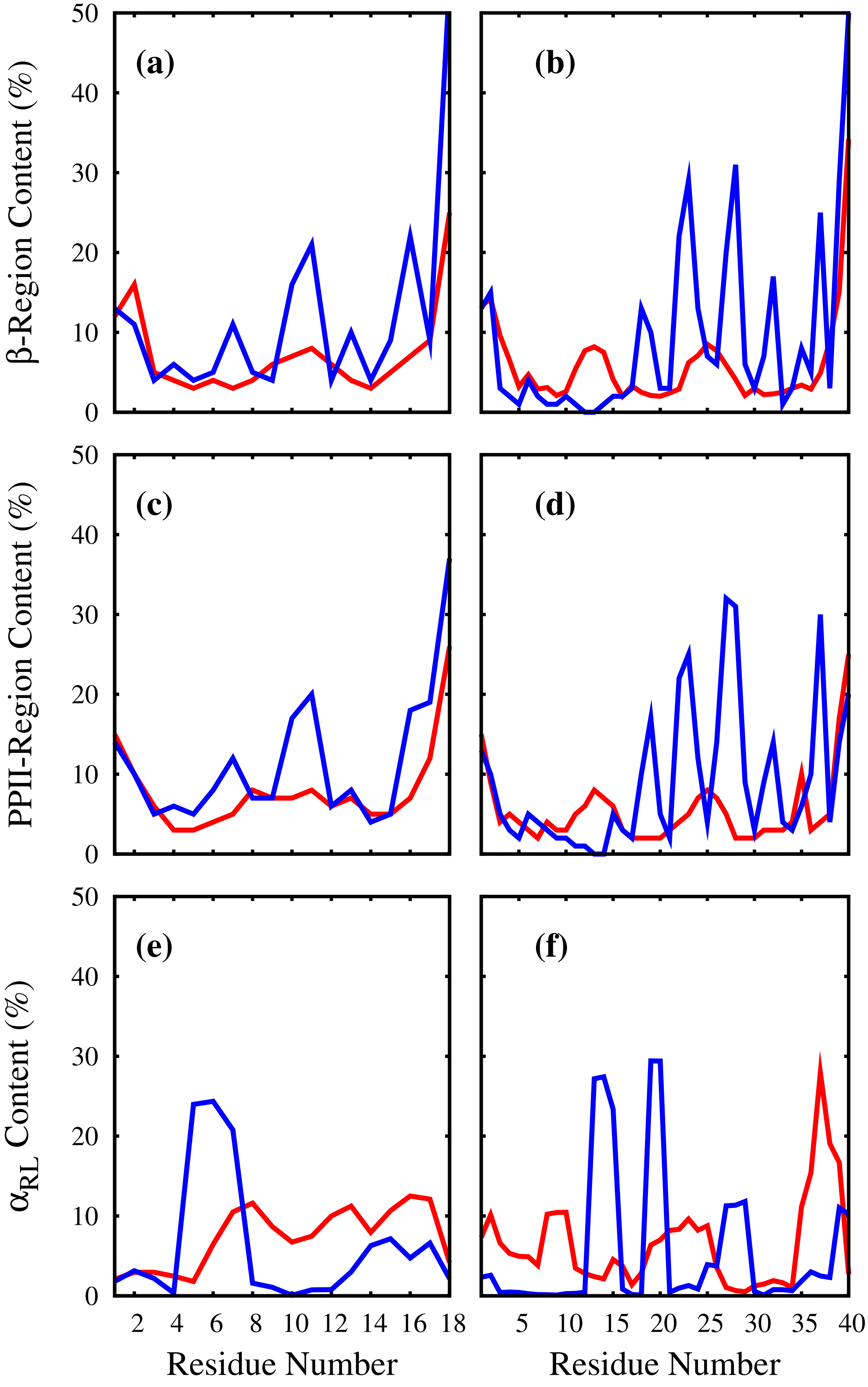

Figure 2. *β*, PPII and α_{RL} content of selected polyQ peptides. Here, given are the contents (as a percentage) of individual glutamine residues found in: (a,b) β-region (c,d) PPII-region (e,f) α_{RL}. These percentages are plotted against the Glu residue numbers for (a,c,e) Q_{18} [red], $Q_{18}P_6$ [blue] and (b,d,f) Q_{40} [red], $Q_{40}P_6$ [blue]. These percentages are obtained from clustering the conformations based on their dihedral angles in the Ramachandran plot.
doi:10.1371/journal.pcbi.1002501.g002

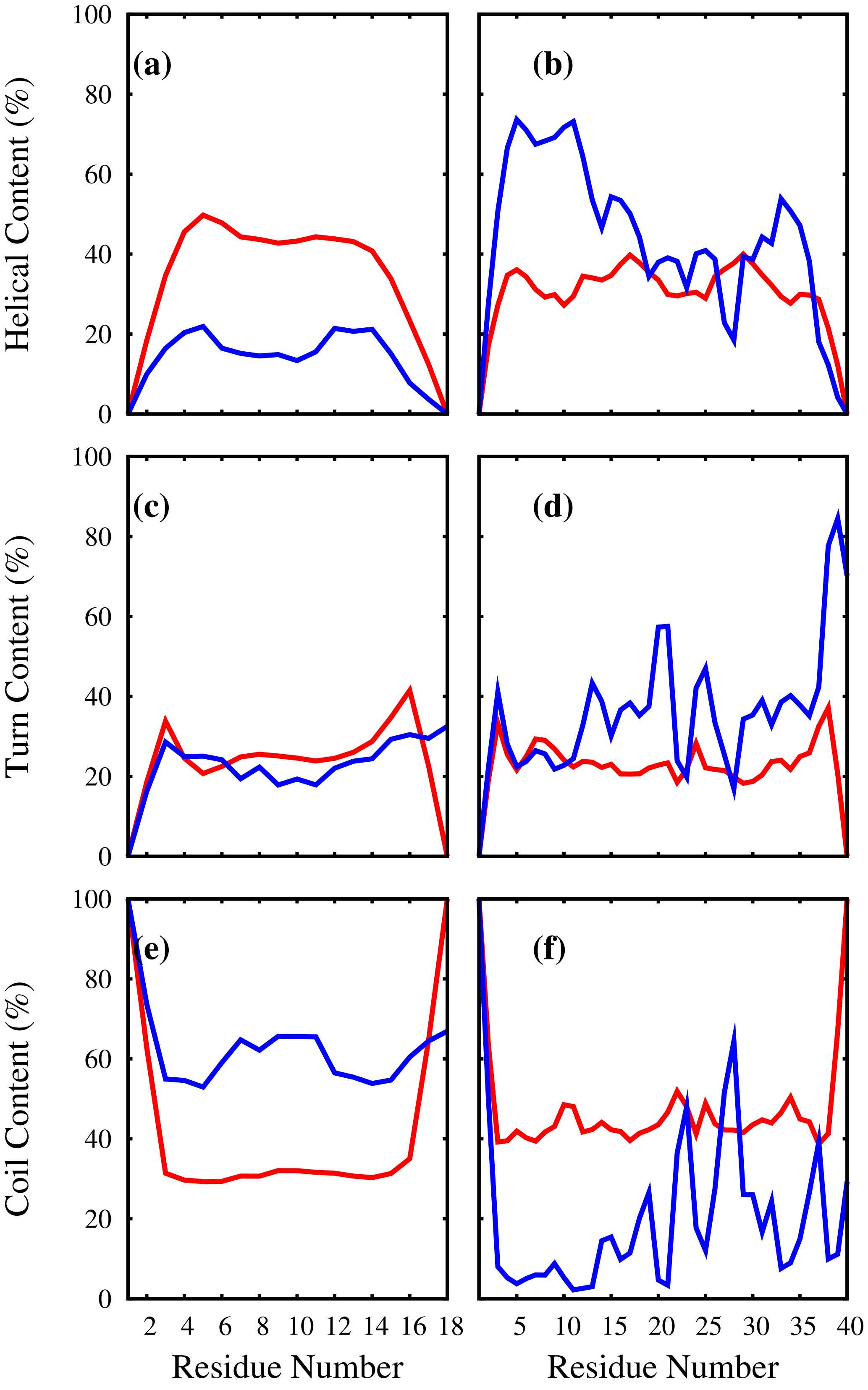

Figure 3. Helical, turn and coil content of selected polyQ peptides. Here, given are the contents (as a percentage) of individual glutamine residues found in the following conformations: (a,b) helical (α,3_{10}) (c,d) turn (H-bonded,bend) (e,f) coil. These percentages are plotted against the Glu residue numbers for (a,c,e) Q_{18} [red],$Q_{18}P_6$[blue] and (b,d,f) Q_{40} [red], $Q_{40}P_6$ [blue]. These percentages are obtained from the DSSP [58,59] analysis code.

doi:10.1371/journal.pcbi.1002501.g003

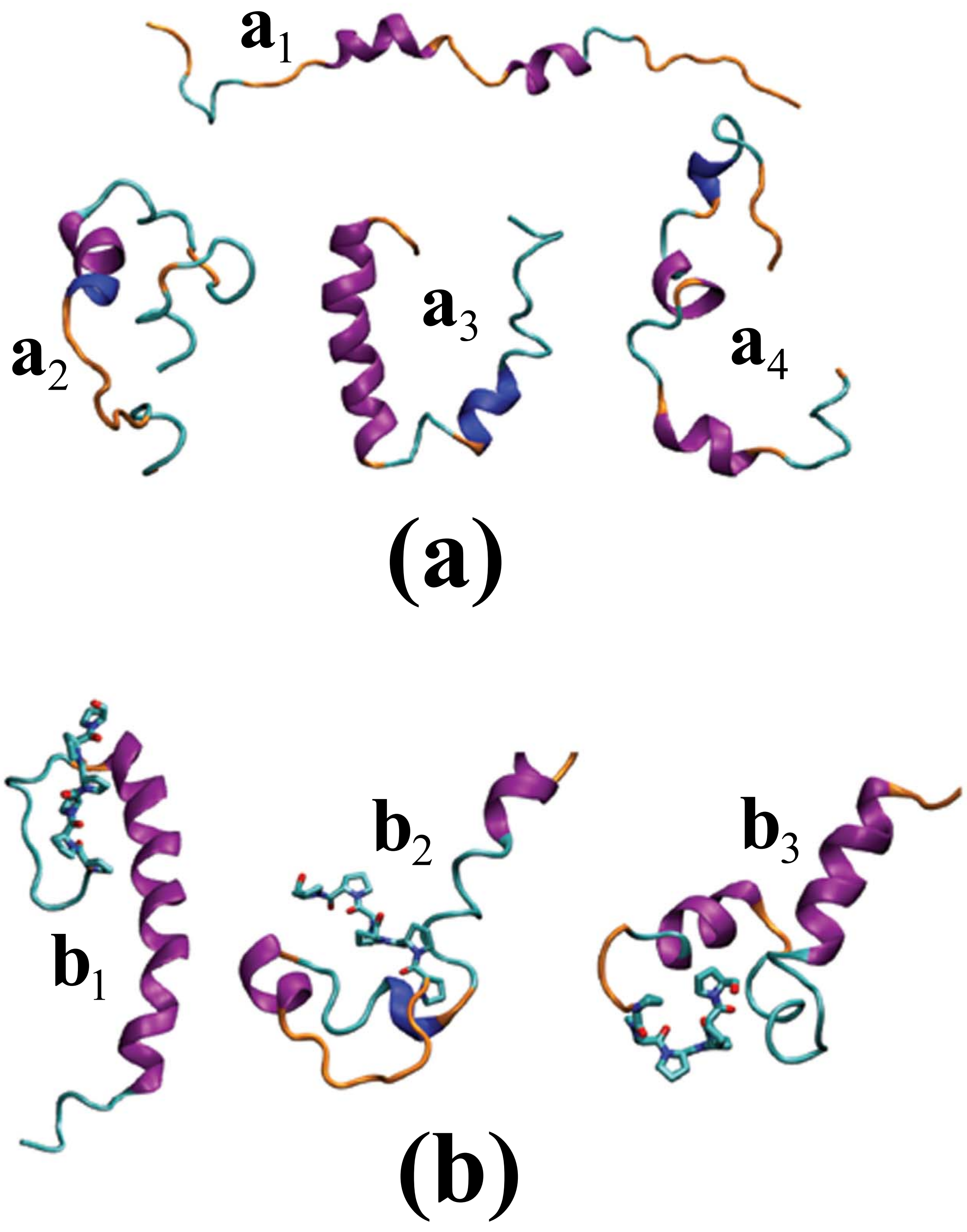

Figure 4. Sample conformations of Q_{40} and $Q_{40}P_6$. Cartoon representation of sample conformations of (a) Q_{40} and (b) $Q_{40}P_6$. Purple, blue, cyan, and orange represent α-helix, 3_{10}-helix, turn, and coil secondary structural motifs, respectively. The licorice-like representation of the proline segment of $Q_{40}P_6$ is given in (b). These structures are plotted by VMD [61] using STRIDE [60] for secondary structure prediction.
doi:10.1371/journal.pcbi.1002501.g004

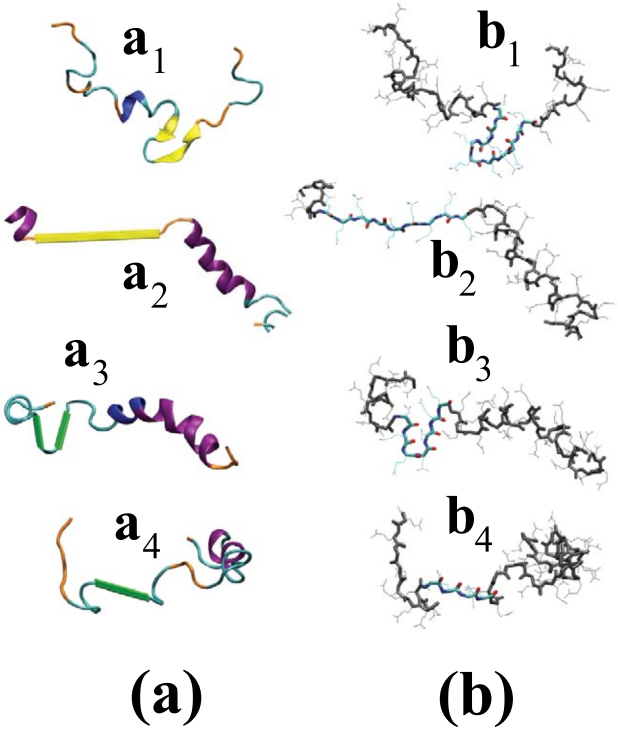

Figure 5. Selected extended conformations of Q_{40} peptides. Here, we give (a) cartoon and (b) licorice-like representation of select conformations of the Q_{40} peptide with (a_1,b_1,a_2,b_2) β and (a_3,b_3,a_4,b_4) α strands. (a) The coloring is similar to Fig. 4 with yellow and green representing β and α strands respectively. We used a dihedral angle-based algorithm to detect the α strands and for other secondary structures in these plots we used STRIDE [60] distributed with VMD [61]. (b) The residues involved in (b_1) β-hairpin, (b_2) isolated β-strand, (b_3) α-harpin, and (ib_4) isolated α-strand are highlighted. The rest of residues are grey and all the side chains are represented by thin lines.
doi:10.1371/journal.pcbi.1002501.g005

Figure 4 shows some examples of isolated and adjacent extended structures in both α and β forms.

Remarkably, among all the sequences presented here, Q_{40} has the highest percentage of extended structures. This peptide shows a significantly higher propensities for the extended structures, particularly the α strands. The population of the structures having at least one α-hairpin is almost 2%, and is higher than the number of structures having at least one β-hairpin. However, the β-hairpin rate is still the highest among all the peptides studied here. Adding the proline segment to the Q_{40} peptide reduces the chance of forming α or β extended structure dramatically, especially in the case of α-hairpins and isolated strands of length four or more. However, PPII propensity is increased in the peptides of length $n>9$ by adding the proline segment.

Table 3 gives more details on the helices and turns observed in the polyQ and polyQ-polyP structures. The helices are found mostly in the right-handed α form except for Q_6 and Q_6P_6 that favor 3_{10} helices due to their short length. This Table also shows the percentage of helical segments present in a given peptide. A helical segment is defined as a series of residues adjacent in the sequence whose secondary structure has been identified as helical by DSSP. Thus helical segments can have varying lengths, and the table lists the number of helical segments (independent of their length). Thus, among Q_{40} conformations, 31% do not have any helical segment but when the prolines are added 99% form at least one helical segment (in particular, 40% of the structures in $Q_{40}P_6$ have 3 helical segments). The addition of P_6 to Q_{40} increases the helical content from 30% in Q_{40} to 43% in $Q_{40}P_6$ (the highest helical content in all peptides), while the addition of polyP decreases the helical content in all other peptides. Comparing Q_{40} and $Q_{40}P_6$ structures, the population of the structures having more than one helix increases.

The select Q_{40} and $Q_{40}P_6$ structures given in Fig. 4a,b illustrate various conformations, for which a statistical description is given in Figs. 2,3 and the Tables 2–3. In particular, the left column of Fig. 3 indicates that adding a polyP segment to Q_{18} reduces the helical content but increases the coil content (while the turn content stays the same). Instead, adding a polyP to Q_{40} (right column of Fig. 3) results in an increase of the helical content in the N-terminal of $Q_{40}P_6$, farther away from the polyP segment. The addition of P_6 to Q_{40} increases not only the number of helical segments but also their length, particularly in the N-terminal half. The population of the structures having short helices (less than 7 residues) is very similar in Q_{40} (26%) and $Q_{40}P_6$ (27%) but 72% of $Q_{40}P_6$ conformations have longer helices (7 residues or more) as compared to only 43% in Q_{40}. Also 37% of the $Q_{40}P_6$ conformations have a helical segment longer than 9 residues while only 20% of Q_{40} conformations do.

Adding the polyP segment generally increases the turn content (both of β and γ types), except for Q_{18}, where the total population of turns stays constant. The majority of turns are of I-β type but there is a smaller population of other types of β turns as well as γ turns. The increase in the γ-turn content of polyQ-polyP peptides can explain why adding the polyP to polyQ sometimes increases

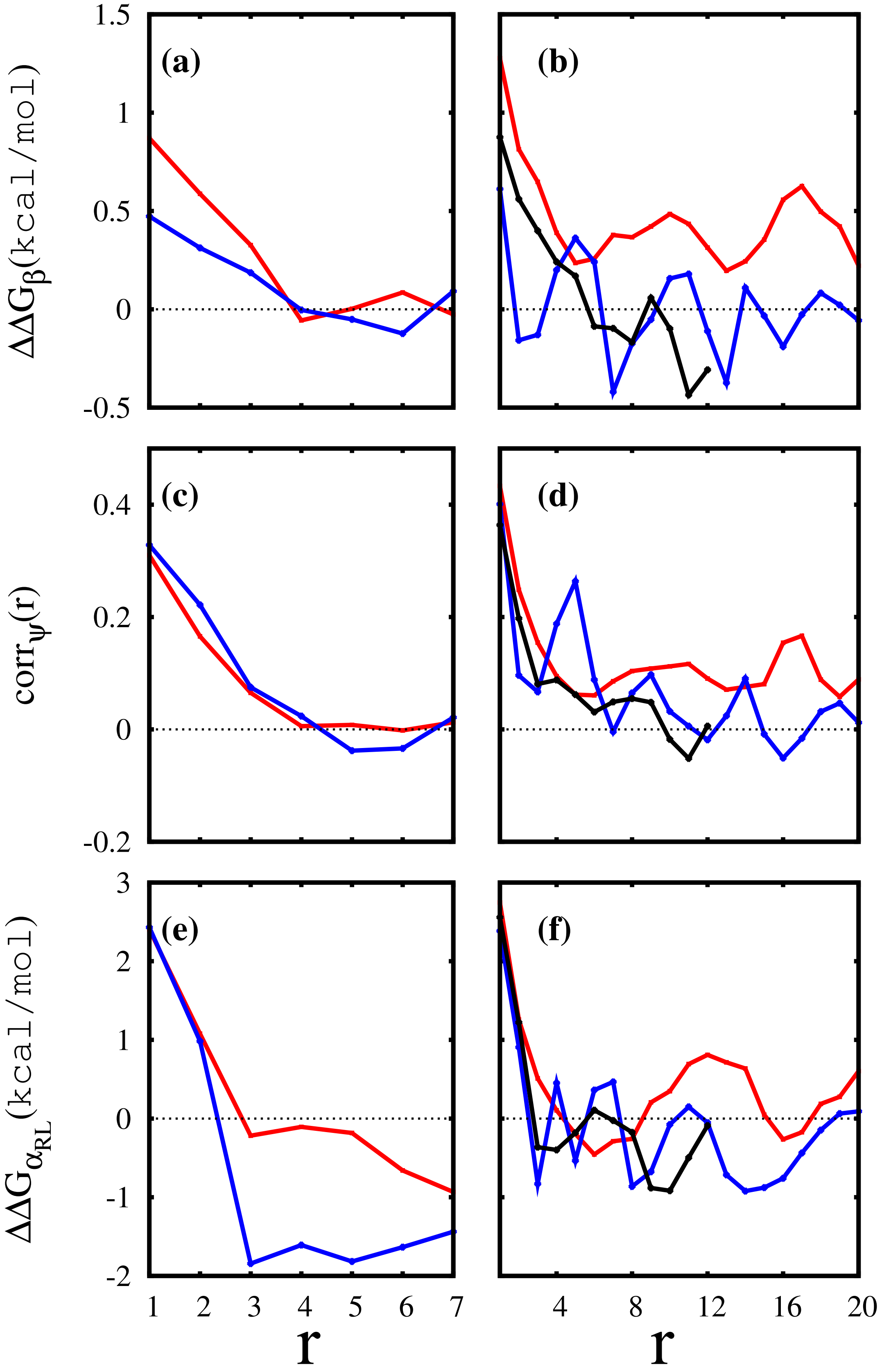

Figure 6. Correlation analysis results for selected polyQ peptides. Here is given the (a) odds ratio based $\Delta\Delta G_\beta$ between any two glutamine residues (Q_n^i and Q_n^j) of Q_{18} [red] and $Q_{18}P_6$ [blue] in terms of ($r = i - j$). From each side of the peptide 5 ending residues are omitted in the calculations to reduce the end effects. (b) Similar to (a) for Q_{40} [red], $Q_{40}P_6$ [blue], and Q_{30} [black]. Here 8 residues from each end are omitted. (c,d) Correlation coefficient between ψ dihedral angles of any two glutamine residues (Q_n^i and Q_n^j) in terms of ($r = i - j$) for (c) Q_{18} [red], $Q_{18}P_6$ [blue] and

(d) Q_{40} [red], $Q_{40}P_6$ [blue], and Q_{30} [black]. The end residues were omitted according to the same protocol used for odds ratio analysis. (e,f) Similar to (a,b) but with the odds ratio calculated using the probabilities that residues belong or not to an α_{RL} repeat region.
doi:10.1371/journal.pcbi.1002501.g006

the α_L content, as α_L residues are involved in most of γ-turns. For instance, one finds more α_L content in the residues of $Q_{40}P_6$ compared to Q_{40} but there are fewer residues in $Q_{40}P_6$ involved in α_{RL} repeats. There is no contradiction here as part of the α_L content is involved in γ turns rather than α-strands. Finally, Fig. 5 presents examples of (rare) extended conformations in the Q_{40} peptides. In particular, the figure shows β hairpins and isolated strands, and α hairpins and isolated strands.

Correlation Analysis

An odds ratio analysis based on the Ramachandran regions was conducted, and results summarized in Figures 6 and 7 for Q_{18}, $Q_{18}P_6$, Q_{30}, Q_{40}, and $Q_{40}P_6$ peptides. We defined the OR as a function of sequence distance $r=j-i$ between two glutamine residues Q^i and Q^j. OR$_X$ indicates an OR based on the X region of Ramachandran plot. These figures display $\Delta\Delta G_X = k_B T ln(\mathrm{OR}_X)$, for a better intuitive illustration. $\Delta\Delta G_X(r)$ measures how the presence or absence of Q^i in the X region can influence the presence or absence of Q^{i+r} in the X region. Here, to reduce the end effects, i only runs between m and $n-m-r$, with $m=5$ for $n=18$ and $m=8$ for $n=30{,}40$.

In Fig. 6a, Q_{18} shows higher correlation $\Delta\Delta G_\beta(r)$ for $r=1,2,3$ than $Q_{18}P_6$. In other words, Q_{18} would have a greater chance of forming β strands if the population of β residues increases.

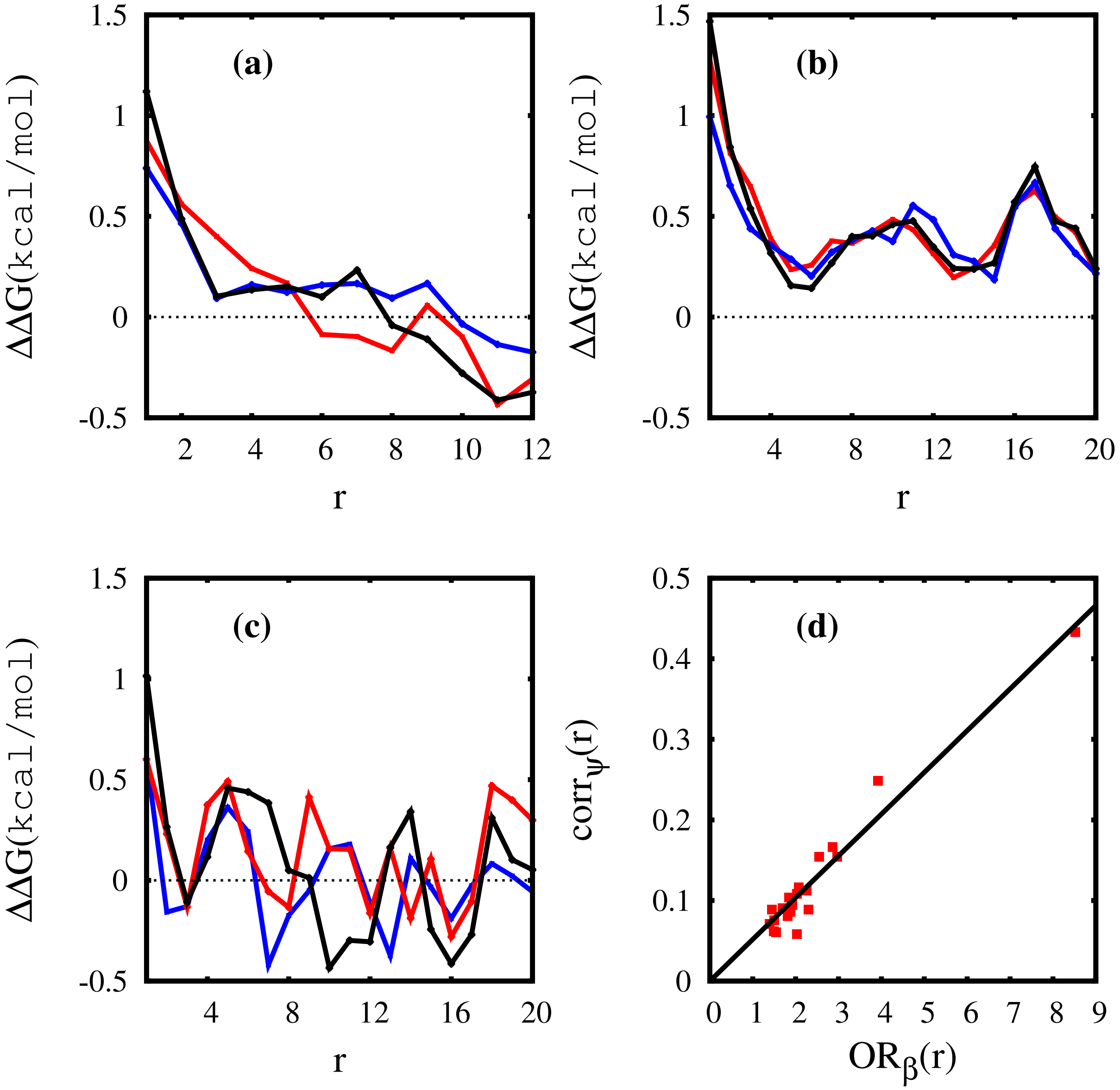

Figure 7. Correlation analysis results for selected polyQ peptides. Specifically, we give $\Delta\Delta G$ for (a) Q_{30} (b) Q_{40} and (c) $Q_{40}P_6$ based on OR(β)[red] OR(PPII)[blue] and OR(α_R)[black]. (d) To compare the linear and OR-based results we plotted OR$_\beta$(r) versus the correlation coefficient corr$_\psi$(r) for Q_{40} that suggests an almost linear behavior with a correlation coefficient of 0.97.
doi:10.1371/journal.pcbi.1002501.g007

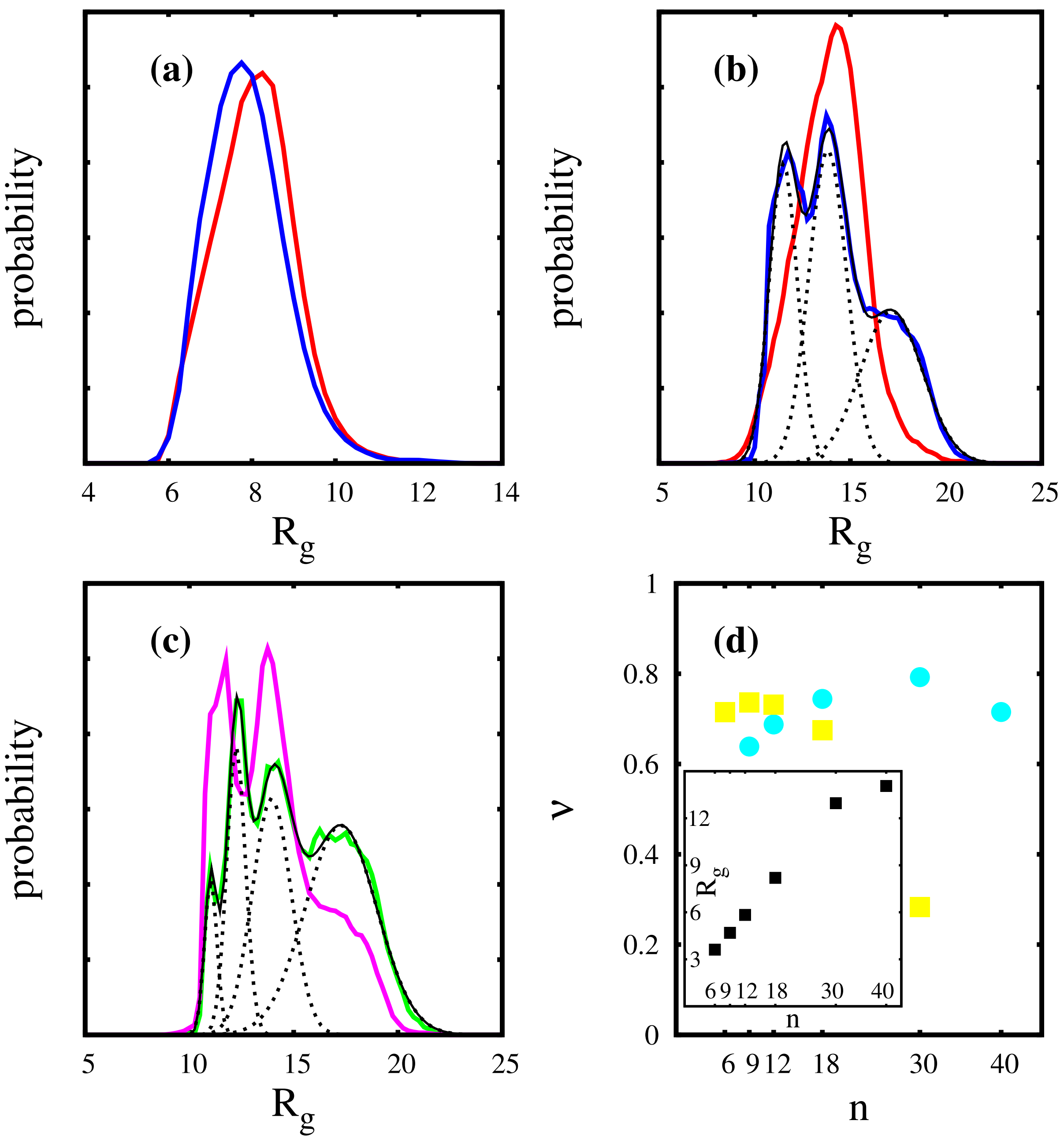

Figure 8. Distribution of radius of gyration of polyQ peptides. (a) The estimated R_g distribution for Q_{18} [red] and $Q_{18}P_6$ [blue]. (b) The estimated R_g distribution for Q_{40} [red] and $Q_{40}P_6$ [blue]. The blue curve can be estimated as the sum [black] of three Gaussian distributions [dotted]. (c) The estimated R_g distribution for $Q_{40}P_6$, considering only the structures with an all-trans proline segment [green]. Similarly the green curve can be estimated as the sum [black] of four Gaussian distributions [dotted]. Considering only the structures that at least have one cis-proline results in the magenta curve for the R_g distribution. All the histograms are obtained using a window of width $\Delta R_g = 0.25$ A. (d) The exponent v in $\frac{R_g(Q_n)}{R_g(Q_m)} = (\frac{n}{m})^v$ relation estimated from select pairs of n (x axis) and m ($m=6$ for blue circles and $m=40$ for yellow squares). Inset: The average R_g (in A) of Q_n peptides for $n=6,9,12,18,30,40$.

doi:10.1371/journal.pcbi.1002501.g008

However the correlation range between the β residues in both Q_{18} and $Q_{18}P_6$ is about $r=3$ since for $r>3$ there is no significant deviation from $\Delta\Delta G_\beta(r)=0$, the expected value for independent events. This situation changes with polymer length. Q_{30} in Fig. 6b has a correlation length of about $r=5$, after which it quickly loses correlation (it even becomes "anti-correlated"). Once again, Q_{40} exhibits unique behavior since $\Delta\Delta G(r)$ does not decay to zero but oscillates around 0.4 kcal/mol and more importantly, the oscillation does not seem to be damped by increasing r (ignoring the smaller r values). This indicates a long-range correlation between the glutamine residues of Q_{40}. (Oscillations can be seen for $Q_{40}P_6$ as well, but they are around zero).

The results of the OR analysis can be further confirmed by conducting a direct correlation analysis on the ψ angles of the glutamine residues. We used the correlation coefficient (also known as cross-correlation or Pearson correlation) as a measure of linear

Table 2. Secondary structure analysis of the polyQ peptides.

	(a) Ramachandran regions					(b) secondary structures			(c) extended structures				
peptide	α_R	β	PPII	α_L	α_{RL}	helix	turn	other	PPII	β-s	β-h	α-s	α-h
Q_{40}	87	5	5	3	7	30	23	47	6.5 (1.3)	7.1 (1.2)	1.1	42 (1.6)	1.9
$Q_{40}P_6$	78	9	9	4	6	43	36	21	8.9 (3.3)	3.9 (0.1)	0.5	25 (0.1)	0.1
Q_{30}	80	8	9	3	7	37	32	31	7.3 (1.3)	4.2 (0.5)	0.5	19 (0.1)	0.7
Q_{18}	81	7	8	4	7	34	23	43	2.4 (0.3)	1.5 (0.5)	0.5	19 (0.1)	0.2
$Q_{18}P_6$	72	13	12	3	6	14	23	63	7.3 (1.0)	0.9 (0.3)	0.1	15 (0.1)	0.1
Q_{12}	79	8	9	4	8	38	31	31	1.9 (0.2)	0.9 (0.2)	0.3	19 (0.8)	0.1
$Q_{12}P_6$	70	14	12	4	6	26	38	36	2.3 (0.3)	1.6 (0.2)	0.1	8 (0.5)	0.0
Q_9	78	9	9	4	8	31	31	38	1.5 (0.2)	0.6 (0.1)	0.4	14 (0.0)	0.0
Q_9P_6	68	15	11	6	10	23	51	26	1.1 (0.1)	1.1 (0.2)	0.6	17 (0.2)	0.0
Q_6	73	12	13	2	4	18	29	53	1.1 (0.2)	1.3 (0.2)	0.0	2 (0.0)	0.0
Q_6P_6	67	17	13	3	8	10	50	40	1.3 (0.2)	1.4 (0.2)	0.0	2 (0.1)	0.0

Here, we give the (a) population (as a percentage) of the residues in the different Ramachandran regions (α_R, β, PPII, and α_L), as well as the population of residues involved in α_{RL} repeats; (b) the population (as a percentage) of residues in different secondary structures (helix, turn, and other secondary structures); (c) the percentage of *conformations* having at least one PPII, α, or β extended secondary structures including isolated strands and hairpins. The isolated PPII_3, α_3, or β_3 (PPII_4, α_4, or β_4) strands – identified in the table as PPII-s, α-s, β-s – are defined based on at least three (four) adjacent residues with the backbone dihedral angles falling into the region associated with these structures; and not involved in any inter-residual hydrogen bonding. Similarly a hairpin – identified in the table as PPII-h, α-h, β-h – is defined based on two adjacent strands of at least three residues with one or more hydrogen bonds between the two strands and a turn in between. For more details of this analysis, that is based on both DSSP [58,59] and dihedral-based clustering, see *Methods*.

doi:10.1371/journal.pcbi.1002501.t002

correlation between the ψ angles of Gln residues of sequence distance r, using the same protocol explained above for odds ratio analysis (*i.e.*, omitting the end residues) and verified the same unique behavior of Q_{40}. First, the ψ dihedral angles were shifted $+40$ degrees (with the assumption of periodic boundary condition at ± 180), then the correlation coefficient of ψ of the residues with a sequence distance r, $\mathrm{corr}_\psi(\mathrm{r})$, was calculated. Note that this correlation measure does not involve any clustering and ignores any dependence on the ϕ dihedral angle, however, it confirms the OR predictions. Although in general both ϕ and ψ angles are needed to identify the Ramachandran region of an amino acid, the linear correlation analysis on ψ angles is still able to detect a long-range, positive correlation for Q_{40} (Figs. 6c,d).

An OR-based correlation analysis for $X = \alpha_{RL}$ is illustrated in Fig. 6e,f. Here, a residue is considered to be an α_{RL} residue if it is involved in an α_{RL} repeat. In the case of Q_{18} and $Q_{18}P_6$ there is an even shorter positive correlation range (compared to OR_β) for both peptides, with a significant negative correlation when r increases. Q_{40} shows a somewhat similar oscillatory behavior around a non-zero average, with negative troughs. Note that the Pearson correlation coefficient cannot be used here for the α_{RL} analysis (in its univariate form) due to the fact that the definition of

Table 3. Helix and turn populations of the polyQ peptides.

	helical content				turn content				
	helix type		helical segments		H-bonding		turn type		
peptide	α	3_{10}	0	1,2,3,4,5	H-bonded	bend	I-β	other β	γ
Q_{40}	23	7	31	3,16,27,18,4	15	7	18	1	4
$Q_{40}P_6$	31	12	1	3,21,40,28,6	23	13	24	3	9
Q_{30}	28	9	11	15,37,30,6	22	10	23	2	7
Q_{18}	27	7	28	39,31,2	18	6	16	2	5
$Q_{18}P_6$	10	4	61	25,13,1	13	10	12	2	9
Q_{12}	29	9	15	76,9	25	6	25	2	4
$Q_{12}P_6$	20	6	30	66,4	23	15	28	3	7
Q_9	22	9	32	67	25	6	24	1	6
Q_9P_6	15	8	48	52	31	20	37	4	10
Q_6	7	11	69	31	24	5	19	2	8
Q_6P_6	3	7	82	18	25	25	36	4	10

The helical content is partitioned into α- and 3_{10}-helix populations. The structures are also categorized based on the number of their helical segments. The population of each category (0,1,2,. . .) is given if greater than 0%. The turn content is partitioned based on both the hydrogen-bonding and turn types. For the secondary structure prediction, the DSSP analysis code [58,59] was used along with the protocols discussed in *Methods*.

doi:10.1371/journal.pcbi.1002501.t003

an α_{RL} repeat is highly dependent on the dihedral angles of both adjacent residues, involving four residues in the correlation analysis instead of two. The ϕ angles are also quite important for the α_R/α_L distinction.

Finally, Fig. 7 compares the behavior of OR-based $\Delta\Delta G_X$ in Q_{30}, Q_{40}, and $Q_{40}P_6$ peptides for $X = \beta, PPII, \alpha_R$. In Q_{30} there are differences between these different regions, but they all decay by increasing r, as expected for short correlations. However, in Q_{40} we see an almost identical behaviour for all three Ramachandran regions. This clearly indicates that the dihedral angles of most of the glutamine residues are correlated in an indirect manner, influencing each other. We compared the OR_β of glutamine residues based on their distance r and the correlation coefficients of their ψ angles for Q_{40}. Fig. 7d shows that the two vary similarly for different r and have a correlation coefficient of about 0.97, suggesting that OR and corr are linearly correlated.

In terms of the error estimate, we note that the estimated standard error for these calculations is different not only for different plots but also for different data points (varying by r) in one plot. The latter is the result of having fewer samples with larger r than shorter r but the former is due to the difference between the population of secondary structures, the number of residues in each peptide, and the number of sampled conformations for each peptide. However, the standard error remains less than 0.02 kcal/mol in most cases. In some exceptions in Fig. 6e,f the standard error could be as high as 0.1 kcal/mol.

Radius of Gyration

Here we consider the statistical ensemble results concerning the radius of gyration and its distribution. The radius of gyration R_g gives a simple and intuitive measure of the overall structure of the polyQ peptides as the collapsed (stretched) structures are associated with smaller (larger) values of R_g. Table S1 gives the R_g of the C_α atoms of the Gln residues in Q_n and Q_nP_6. The proline segments are not included in the calculation of R_g so that the polyQ sequences are compared on equal footing. The averages are accompanied by the standard deviation that somewhat estimates the width of the distribution, if it is close to a normal distribution. The averages do not show much difference between Q_n and Q_nP_6 peptides. The standard deviation is also very similar between the two in most cases except for the case $n = 40$. Fig. 8a shows the R_g distribution of Q_{18} [red] and $Q_{18}P_6$ [blue] peptides that is close to a normal distribution with a longer tail on the right as expected for a random-coil structure. $Q_{18}P_6$ is only slightly more compact. The normal distribution with a slightly longer tail as a characteristic distribution of random coil is seen for all of these peptides except for $Q_{40}P_6$. Fig. 8b shows that although Q_{40} follows the same distribution, $Q_{40}P_6$ can be estimated as the sum of three distinct Gaussian distributions.

We used the Marquardt-Levenberg [62] algorithm to estimate the probability distribution of $Q_{40}P_6$ as the sum of three Gaussian distributions (see Fig. 8b), each representing one class of structures covering 24, 44, and 32% of the samples distributed around an R_g of 11.41, 13.65, and 17.08 A, respectively. The fitting resulted in a reduced χ^2 smaller than 10^{-6}, indicating that this model explains the probability distribution of $Q_{40}P_6$ well. Examining the structures of each class shows that the P_6 segment is responsible for this clear difference between the three classes. The structures distributed around $R_g = 17.07$A, accounting for almost one third of the samples, have relatively stretched conformations (see Fig. 4b$_1$), and this correlates with the presence of all-trans prolyl bonds in P_6. In these proline isomers, P_6 forms a rigid stretched helical segment, in contrast with a proline segment including one or more cis-isomers, particularly in the middle of the segment (see Fig. 4b$_3$). Table S1 shows the trans content of each of the prolyl bonds of P_6 as well as the population of the P_6 isomers with all-trans prolyl bonds. There is a clear difference between $Q_{40}P_6$ and the rest of proline-containing peptides in terms of cis-trans isomerization. Although, 73–77% of the residues are in trans conformation in the shorter peptides, only 12–23% of the structures are all-trans. In $Q_{40}P_6$ 60% of the structures are stretched all-trans conformations. What is more interesting is that the distribution of radius of gyration is meaningfully different for the all-trans proline sub-ensemble as shown in Fig. 4c. Green curve is the R_g distribution of this sub-ensemble and magenta curve is the R_g distribution, obtained from the rest of the structures (*i.e.*, cis-containing polyP). Here we somewhat recognize four normal distributions. We use a similar method as explained above to fit these Gaussians. We find four clusters with 6, 17, 29, and 48% of the population centered around $R_g = 11.02$, 12.24, 13.94, and 17.27 respectively. The conclusion is that all-trans prolines increase the population of the stretched cluster considerably. This somewhat explains why we do not observe this partitioning of the clusters with proline segment in shorter peptides (see Fig. 8a) because in those cases the population of all-trans conformations is not large enough to affect the overall R_g distribution.

As the peptides Q_n grow with residue number n, their structure becomes more collapsed. In particular, the average radius of gyration for Q_{40} is only about 1.1 Å larger than for Q_{30}. The inset in Fig. 8d illustrates the dependence of the radius of gyration on the length of the peptide. Assuming $R_g(Q_n) \propto n^{\nu}$ one can estimate ν using any pair of peptides such as Q_n and Q_m from $(\frac{n}{m})^{\nu} = \frac{R_g(Q_n)}{R_g(Q_m)}$. Fig. 8d gives examples of the estimated ν_{nm} for different pairs of n and m: n is given by the indices in the x axis and m is $m = 6$ (cyan circles) or $m = 40$ (yellow squares). There is an abrupt collapse of the structure ($\nu = 0.28$) on going from $n = 30$ to $m = 40$.

Discussion

Our atomistic simulations show the disordered nature of monomeric polyglutamine peptides, in agreement with experimental conclusions [6,13–15] and with previous all-atom MD simulations [35–38]. Our simulations are also in agreement with recent experiments [18] in that the monomeric polyQ is different from a total random coil or a protein denatured state, with a significant presence of short α-helices. Therefore polyglutamine is a disordered peptide that is somewhat preorganized, containing short rigid segments [63,64]. Contrary to certain coarse-grained models [27–29,31], our atomistic simulations provide no evidence for a large β content in monomeric polyglutamines.

We observed that the Q_{40} peptide forms an ensemble of mostly compact structures with an average radius of gyration only about 1.1 Å larger than that of Q_{30}. This agrees with the conclusions from single-molecule force-clamp experiments [24] that polyQ chains collapse to form a heterogeneous ensemble of globular conformations that are mechanically stable. For the radius of gyration of the shorter peptides, we observed an exponent ν slightly larger than that of a random-coil in a good solvent (*i.e.* about 0.6, [65]). However, we have not been able to simulate a large enough range of peptide sizes in order to get a good estimate of ν. This may not be necessary, since the simulations suggest that the radius of gyration does not follow a power law anyway (see Fig. 8d).

The addition of a short C-terminal proline segment to the Q_{40} peptide changes the distribution of the radius of gyration from a Gaussian-like function with a longer tail for larger R_g – a characteristic of a random coil, seen also in all the other peptides studied here – to a combination of three distinct Gaussians. The way the proline segment affects the R_g distribution is closely correlated with the cis-trans pattern of its prolyl bonds. An all-trans proline segment (the most common pattern in $Q_{40}P_6$) results in the multi-modal distribution of Fig. 8. Instead, proline isomers with cis bonds are abundant in shorter peptides which results in the normal R_g distribution. We note that prolyl bond isomerization requires crossing barriers of 10–20 kcal/mol, which can only be accomplished with special enhanced-sampling techniques such as used here [44,47,49].

The addition of the polyP segment to polyQ introduces position dependent features among the Gln residues. This is readily seen in Fig. 3. The fluctuations observed cannot be explained as "noise" resulting from sampling limitations. As explained in the previous section, sampling of independent data produces the same features, which suggests a sensitive dependence on the position of the residue in the sequence. Interestingly, polyP induces helix formation in the further residues in the N-terminal of Q_{40}, while creating more turns in the nearer Gln residues. As a result of the polyP addition, the overall α-helical content of Q_{40} increases. This is in contrast with the shorter peptides in which the α-helical content drops considerably by adding the polyP segment.

Experimentally, it has been claimed that the addition of polyP to polyQ decreases the α-helical content of polyQ for all polyQ lengths [17]. A superficial comparison might indicate that this is in contradiction with our results for Q_{40}. Our results are, however, in agreement with the experimental data, which is based on the CD spectra of these peptides. These CD spectra identify the distribution of individual backbone dihedral angles rather than the actual α-helical content, a quantity not only dependent on the individual residues but also the way they are aligned. Our simulations are in total agreement with this observation as we see a decrease in the population of the α_R cluster (*i.e.*, the residues falling into the α_R region of Ramachandran plot) in all the peptides studied here, as we add a P_6 segment to the C-terminal (Table 2). As we have pointed out before [41,42], care is needed in the interpretation of the CD data. Table 2 shows that the majority of the residues in the α_R cluster are not involved in any form of helix in either polyQ or polyQ-polyP peptides, and while the helical content of all other peptides decreases, that of Q_{40} actually increases with the addition of P_6. While this effect for Q_{40} cannot be ruled out as an defficiency of the force field, it is interesting to note that this would represent quite an effective way of neutralizing Q_{40}, since the rather stable α helix will not be prone to aggregation.

In addition to α and 3_{10} helices, as well as β and γ turns, one can identify a small but non-negligible population of extended secondary structures of β and α strands, particularly in the Q_{40} peptides. PolyP increases the β-region content in the Ramachandran plot, but decreases the β-strand content (as explained before, several β residues need to be adjacent in order to form a β-strand). For Q_{40}, the addition of polyP dramatically decreases the content of β_3, β_4, α_3 and α_4 strands. On the other hand, relatively short PPII helices in polyQ form another extended secondary structure that happens to be more common in Q_nP_6 peptides than Q_n peptides for $n>9$. The PPII strands do not form inter-residual hydrogen bonds (hairpins,sheets) and would not favor aggregation.

In this work we used an odds ratio analysis to quantify the dependencies among certain properties of the molecules. Regarding the β-strand formation in Q_{40}, the graph for $\Delta\Delta\mathrm{G}_\beta$ in Fig. 6 shows a positive, long-range correlation in sequence distance. In other words, the chances of two glutamine residues falling into the β region of the Ramachandran map correlate positively with each other, even if they are distant in the sequence. This long range correlation was not seen in any other peptide but Q_{40}. Interestingly, this long-range correlation for the Q_{40} peptide is not limited to the β-region but it is also seen in other regions such as α_R and PPII. In particular, $\Delta\Delta G$ *scales* for the β, α_R and PPII regions as shown in Fig. 7. A linear correlation analysis on ψ dihedral angle verifies the very same long-range correlation between glutamine residues of Q_{40} peptide, a correlation that is absent in other peptides studied here. This surprising phenomenon could be interpreted as the possibility of the growth of any of these secondary structures in the long polyQ peptides, especially if the conformation were "seeded" with a given secondary structure. In a polymeric form of polyglutamine, the nucleation of α or β strands could result in further growth of those strands or could induce growth in adjacent strands resulting in the the growth of α or β sheets. Interestingly, the "period" for the oscillations of $\Delta\Delta G$ is approximately 7–8 residues, which is also the optimal experimental extended chain length in an aggregate [7].

The populations of α-strand, β-strand, α-hairpin, and β-hairpin (Table 2) decrease and the long-range correlations $\Delta\Delta\mathrm{G}_\beta$ and $\Delta\Delta\mathrm{G}_{\alpha_{RL}}$ are disrupted by the presence of the C-terminal proline residues in $Q_{40}P_6$. For shorter peptides, the corresponding populations are much lower, and the $\Delta\Delta G$ correlations are short-ranged. Taken together, these results indicate that for $Q_{40}P_6$ (but not for the shorter peptides) nucleation could start in one of these strands or hairpins (that can align two strands) and then grow from there, favored by the positive correlations generated by the longer peptide.

We can summarize the main findings of this work as follows:

1. *Monomeric Q_{40} peptide forms an ensemble of disordered, mostly compact structures with non-negligible α helical content and other secondary structures, and with a very slow growth of the radius of gyration with the number of peptides for longer polyQ peptides*. This is in agreement with previous experimental and simulation results [6,13–15,24,35–38]. The average radius of gyration of Q_{40} is only about 1.1 Å larger than that of Q_{30}.
2. *The average radius of gyration for polyQ does not vary with the addition of polyP, but its distribution in Q_{40} is affected by the isomerization states of the polyP segment.*
3. *For peptides of all lengths, the population of the α_R region in the Ramachandran plot decreases while the populations of the β and PPII Ramachandran regions increase with the addition of polyP.*
4. With respect to secondary structures (i.e., dihedrals angles and hydrogen bonds, *the addition of polyP increases the PPII and turn contents, and decreases the helical content in all peptides but Q_{40}*. These effects probably disfavor aggregation as PPII structures dislike backbone H-bonding, turns increase disorder, and the increase of helical content in Q_{40} may also disfavor aggregation as helices are quite stable, with all their H-bonds properly engaged.
5. *Although small, the populations of β_3, β_4, α_3 and α_4 strands, as well as β-hairpins and α-hairpins, are considerably larger for Q_{40} than for smaller peptides*. These populations decrease when polyP is added. These small secondary structures are good candidates to initiate nucleation: the strands might "attract" other strands to hydrogen bond and the hairpins help to align two strands. Their suppression by the presence of polyP would disfavor aggregation.
6. An odds-ratio based correlation function $\Delta\Delta G$ describes how the chances of two Gln residues of falling into a given region of the Ramachandran plot correlate. *Only Q_{40} shows positive,* **long-range** *correlation in sequence space for various regions of the Ramachandran plot. The addition of polyP destroys this long-range correlation for $\Delta\Delta\mathrm{G}_\beta$ and $\Delta\Delta\mathrm{G}_{\alpha_{RL}}$*. In particular, $\Delta\Delta G$ *scales* for the β, α_R and PPII regions. Together with the results described

in (6) above, this could be interpreted as the possibility of the growth of the α or β strands or hairpins already present in disordered Q_{40} (or longer polyQ peptides). Interestingly, the "period" for the oscillations of $\Delta\Delta G$ is approximately 7–8 residues, which is also the optimal experimental extended chain length in an aggregate [7]. A linear correlation analysis on ψ dihedral angles confirms this period is a "universal" feature of correlations in long polyQ peptides.

Our careful statistical analysis has revealed a wealth of very subtle effects that are far from obvious. Secondary structures such as α helices, β-sheets, α-sheets, PPII helices, and coils have all been reported in the literature. The picture that is emerging is that if one can induce the nucleation of one of these structures, or provide a template for it, a long enough polyQ polymer or an aggregate will probably continue growing in the given conformation, even if it is not the absolute thermodynamic minimum. In this sense, the wealth of conformations of polyQ is reminiscent of the different phases that appear in 'inorganic' systems with short-range attractive interactions and long-range electrostatics interactions such as Langmuir monolayers or block copolymers, where *kinetics* effects also play a fundamental role in determining the final phase of the system. PolyQ is a very special homopeptide due to its long side changes and the dipoles at the ends. The van der Waals packing of the side chains provides the source of short-range attractive interactions, while the carboxamide groups provide the long-range dipolar interactions [34]. In this sense, the only other peptide that would exhibit similar behavior is asparagine, with one methyl group less in its side chain [34]. The "collapsed" random coil would just represent the *frustration* between different phases.

Supporting Information

Figure S1 α-helical content of Q_{18} and $Q_{18}P_6$ peptides. Here, we give (a,b) the α-helical content (as a percentage) of individual glutamine residues plotted against their residue numbers for Q_{18} [red] and $Q_{18}P_6$ [blue] as obtained from the last 100 *ns* of two 200 *ns* long independent simulations; (c,d) The α-helical content (as a percentage) of individual glutamine residues plotted against their residue numbers for Q_{40} [red] and $Q_{40}P_6$ [blue] as obtained from the third (c) and the fourth (d) 250 *ns* of 1000 *ns* REMD simulations.
(EPS)

Figure S2 Ramachandran plots of Gln residues in the Q_{40} peptide. On these plots, each pixel represents a $1^\circ \times 1^\circ$ bin, whose intensity represents its relative population, ranging from 1,2,…, 49, and 50 or more samples out of 5×10^5 conformations. Color scheme is as in Fig. 1.
(EPS)

Figure S3 Ramachandran plots of Gln residues in the $Q_{40}P_6$ peptide. See Figures 1 and S2 for the details.
(EPS)

Table S1 Radius of gyration and cis-trans isomerization.
(PDF)

Text S1 This text includes a description of our simulation details, secondary structure assignments, and radius of gyration analysis.
(PDF)

Acknowledgments

We thank the NC State HPC Center for extensive computational support.

Author Contributions

Conceived and designed the experiments: MM VB CR CS. Performed the experiments: MM. Analyzed the data: MM. Contributed reagents/materials/analysis tools: MM. Wrote the paper: MM CR CS.

References

1. Zoghbi HY, Orr HT (2000) Glutamine repeats and neurodegeneration. Ann Rev Neurosci 23: 217–247.
2. Davies SW, Turmaine M, Cozens BA, DiFiglia M, Sharp AH, et al. (1997) Formation of neuronal intranuclear inclusions underlies the neurological dysfunction in mice transgenic for the hd mutation. Cell 90: 537–548.
3. Michalik A, Van Broeckhoven C (2003) Pathogenesis of polyglutamine disorders: aggregation re- visited. Hum Mol Genet 12: R173–186.
4. Scherzinger E, Lurz R, Turmaine M, Mangiarini L, Hollenbach B, et al. (1997) Huntingtin-encoded polyglutamine expansions form amyloid-like protein aggregates in vitro and in vivo. Cell 90: 549–558.
5. Scherzinger E, Sittler A, Schweiger K, Heiser V, Lurz R, et al. (1999) Self-assembly of polyglutamine-containing huntingtin fragments into amyloid-like fibrils: Implications for huntingtons disease pathology. Proc Natl Acad Sci U S A 96: 4604–4609.
6. Chen S, Berthelier V, Yang W, Wetzel R (2001) Polyglutamine aggregation behavior in vitro supports a recruitment mechanism of cytotoxicity. J Mol Biol 311: 173–182.
7. Thakur AK, Wetzel R (2002) Mutational analysis of the structural organization of polyglutamine aggregates. Proc Natl Acad Sci U S A 99: 17014–17019.
8. Wacker JL, Zareie MH, Fong H, Sarikaya M, Muchowski PJ (2004) Hsp70 and Hsp40 attenuate formation of spherical and annular polyglutamine oligomers by partitioning monomer. Nat Struct Mol Biol 11: 1215–1222.
9. Nagai Y, Inui T, Popiel HA, Fujikake N, Hasegawa K, et al. (2007) A toxic monomeric conformer of the polyglutamine protein. Nat Struct Mol Biol 14: 332–340.
10. Bodner RA, Outeiro TF, Altmann S, Maxwell MM, Cho SH, et al. (2006) Pharmacological promotion of inclusion formation: A therapeutic approach for Huntington's and Parkinson's diseases. Proc Natl Acad Sci U S A 103: 4246–4251.
11. Glabe CG, Kayed R (2006) Common structure and toxic function of amyloid oligomers implies a common mechanism of pathogenesis. Neurology 66: S74–S78.
12. Kar K, Jayaraman M, Sahoo B, Kodali R, Wetzel R (2011) Critical nucleus size for disease-related polyglutamine aggregation is repeat-length dependent. Nat Struct Mol Biol 18: 328–36.
13. Chen S, Ferrone FA, Wetzel R (2002) Huntington's disease age-of-onset linked to polyglutamine aggregation nucleation. Proc Natl Acad Sci U S A 99: 11884–11889.
14. Lee CC, Walters RH, Murphy RM (2007) Reconsidering the mechanism of polyglutamine peptide aggregation. Biochemistry 46: 12810–12820.
15. Walters RH, Murphy RM (2009) Examining polyglutamine peptide length: A connection between collapsed conformations and increased aggregation. J Mol Biol 393: 978–992.
16. Nozaki K, Onodera O, Takano H, Tsuji S (2001) Amino acid sequences flanking polyglutamine stretches influence their potential for aggregate formation. Neuroreport 12: 3357–3364.
17. Bhattacharyya A, Thakur AK, Chellgren VM, Thiagarajan G, Williams AD, et al. (2006) Oligo-proline effects on polyglutamine conformation and aggregation. J Mol Biol 355: 524–535.
18. Kim MW, Chelliah Y, Kim SW, Otwinowski Z, Bezprozvanny I (2009) Secondary structure of huntingtin amino-terminal region. Structure 17: 1205–1212.
19. Thakur AK, Jayaraman M, Mishra R, Thakur M, Chellgren VM, et al. (2009) Polyglutamine disruption of the huntingtin exon 1 n terminus triggers a complex aggregation mechanism. Nat Struct Mol Biol 16: 380–389.
20. Darnell GD, Orgel JP, Pahl R, Meredith SC (2007) Flanking polyproline sequences inhibit beta-sheet structure in polyglutamine segments by inducing ppii-like helix structure. J Mol Biol 374: 688–704.
21. Darnell GD, Derryberry J, Kurutz JW, Meredith SC (2009) Mechanism of cis-inhibition of polyq fibrillation by polyp: Ppii oligomers and the hydrophobic effect. Biophys J 97: 2295–2305.
22. Wood SJ, Wetzel R, Martin JD, Hurle MR (1995) Prolines and amyloidogenicity in fragments of the alzheimers peptide beta/a4. Biochem 34: 724–730.
23. Thakur AK, Wetzel R (2002) Mutational analysis of the structural organization of polyglutamine aggregates. Proc Natl Acad Sci U S A 99: 17014–17019.
24. Dougan L, Li J, Badilla CL, Berne BJ, Fernandez JM (2009) Single homopolypeptide chains collapse into mechanically rigid conformations. Proc Nat Acad Sci U S A 106: 12605–12610.
25. Starikov EB, Lehrach H, Wanker EE (1999) Folding of oligoglutamines: a theoretical approach based upon thermodynamics and molecular mechanics. J Biomol Struct Dyn 17: 409–427.

26. Burke MG, Woscholski R, Yaliraki SN (2003) Differential hydrophobicity drives self-assembly in Huntington's disease. Proc Natl Acad Sci U S A 100: 13928–13933.
27. Barton S, Jacak R, Khare SD, Ding F, Dokholyan NV (2007) The length dependence of the polyq-mediated protein aggregation. J Biol Chem 282: 25487–25492.
28. Marchut AJ, Hall CK (2007) Effects of chain length on the aggregation of model polyglutamine peptides: Molecular dynamics simulations. Prot: Struct Func Bioinf 66: 96–109.
29. Lakhani VV, Ding F, Dokholyan NV (2010) Polyglutamine induced misfolding of huntingtin exon1 is modulated by the flanking sequences. PLoS Comput Biol 6: e1000772.
30. Laghaei R, Mousseau N (2010) Spontaneous formation of polyglutamine nanotubes with molecular dynamics simulations. J Chem Phys 132: 165102.
31. Digambaranath JL, Campbell TV, Chung A, McPhail MJ, Stevenson KE, et al. (2011) An accurate model of polyglutamine. Prot: Struct Funct Bioinf 79: 1427–1440.
32. Daggett V (2006) α-sheet: The toxic conformer in amyloid diseases? Acc Chem Res 39: 594–602.
33. Armen RS, Bernard BM, Day R, Alonso DOV, Daggett V (2005) Characterization of a possible amyloidogenic precursor in glutamine-repeat neurodegenerative diseases. Proc Natl Acad Sci U S A 102: 13433–13438.
34. Babin V, Roland C, Sagui C (2011) The α-sheet: A missing-in-action secondary structure? Prot: Struct Funct Bioinf 79: 937–946.
35. Wang X, Vitalis A, Wyczalkowski MA, Pappu RV (2006) Characterizing the conformational ensemble of monomeric polyglutamine. Prot: Struct Funct Bioinf 63: 297–311.
36. Vitalis A, Wang X, Pappu RV (2007) Quantitative characterization of intrinsic disorder in polyglutamine: Insights from analysis based on polymer theories. Biophys J 93: 1923–1937.
37. Vitalis A, Wang X, Pappu RV (2008) Atomistic simulations of the effects of polyglutamine chain length and solvent quality on conformational equilibria and spontaneous homodimerization. J Mol Biol 384: 279–297.
38. Wang Y, Voth GA (2010) Molecular dynamics simulations of polyglutamine aggregation using solvent-free multiscale coarse-grained models. J Phys Chem B 114: 8735–8743.
39. Ramachandran GN, Ramakrishnan C, Sasisekharan V (1963) Stereochemistry of polypeptide chain configurations. J Mol Biol 7: 95–99.
40. Edwards AWF (1963) The measure of association in a 2×2 table. J Royal Stat Soc Series A (General) 126: 109–114.
41. Moradi M, Babin V, Sagui C, Roland C (2011) A statistical analysis of the PPII propensity of amino acid guests in proline-rich peptides. Biophys J 100: 1083–1093.
42. Moradi M, Babin V, Sagui C, Roland C (2011) PPII propensity of multiple-guest amino acids in a proline-rich environment. J Phys Chem B 115: 8645–8656.
43. Babin V, Sagui C (2010) Conformational free energies of methyl-β-l-iduronic and methyl-β-d-glucronic acids in water. J Chem Phys 132: 104108.
44. Moradi M, Babin V, Roland C, Sagui C (2010) A classical molecular dynamics investigation of the free energy and structure of short polyproline conformers. J Chem Phys 133: 125104.
45. Geyer CJ (1991) Markov chain monte carlo maximum likelihood. In: Computing Science and Statistics: The 23rd symposium on the interface. Fairfax: Interface Foundation of North America. pp 156–163.
46. Sugita Y, Okamoto Y (1999) Replica-exchange molecular dynamics method for protein folding. Chem Phys Lett 314: 141.
47. Moradi M, Babin V, Roland C, Darden T, Sagui C (2009) Conformations and free energy landscapes of polyproline peptides. Proc Natl Aca Sci U S A 106: 20746.
48. Frenkel D, Smit B (2002) Understanding Molecular Simulation Comput Sci Ser Acad Press.
49. Moradi M, Lee JG, Babin V, Roland C, Sagui C (2010) Free energy and structure of polyproline peptides: an ab initio and classical molecualr dynamics investigation. Int J Quant Chem 110: 2865–2879.
50. Babin V, Roland C, Sagui C (2008) Adaptively biased molecular dynamics for free energy calculations. J Chem Phys 128: 134101.
51. Babin V, Karpusenka V, Moradi M, Roland C, Sagui C (2009) Adaptively biased molecular dynamics: An umbrella sampling method with a time-dependent potential. Int J Quant Chem 109: 3666–3678.
52. Case DA, Darden TA, Cheatham III TE, Simmerling CL, Wang J, et al. (2008) "AMBER 10". San Francisco: University of California.
53. Hornak V, Abel R, Okur A, Strockbine B, Roitberg A, et al. (2006) Comparison of multiple amber force fields and development of improved protein backbone parameters. Proteins 65: 712–725.
54. Onufriev A, Bashford D, Case DA (2000) Modification of the generalized Born model suitable for macromolecules. J Phys Chem B 104: 3712–3720.
55. Onufriev A, Bashford D, Case DA (2004) Exploring protein native states and large-scale conformational changes with a modified generalized Born model. Proteins 55: 383–394.
56. Weiser J, Shenkin PS, Still WC (1999) Approximate Atomic Surfaces from Linear Combinations of Pairwise Overlaps (LCPO). J Comp Chem 20: 217–230.
57. Zimmerman SS, Pottle MS, N'emethy G, Scheraga HA (1977) Conformational analysis of the 20 naturally occurring amino acid residues using ecepp. Macromolecules 10: 1–9.
58. Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22: 2577–2637.
59. Joosten RP, Te Beek TAH, Krieger E, Hekkelman ML, Hooft RWW, et al. (2011) A series of pdb related databases for everyday needs. Nucleic Acids Res 39: D411–D419.
60. Frishman D, Argos P (1995) Knowledge-based secondary structure assignment. Prot: Struct Funct Bioinf 23: 566–579.
61. Humphrey W, Dalke A, Schulten K (1996) VMD – Visual Molecular Dynamics. J Mol Graph 14: 33–38.
62. Levenberg K (1944) A method for the solution of certain non-linear problems in least squares. Q App Math 2: 164–168.
63. Rose GD, Fleming PJ, Banavar JR, Maritan A (2006) A backbone-based theory of protein folding. Proc Natl Acad Sci U S A 103: 16623–16633.
64. Fitzkee NC, Rose GD (2004) Reassessing random-coil statistics in unfolded proteins. Proc Natl Acad Sci U S A 101: 12497–12502.
65. Flory PJ (1953) Principles of Polymer Chemistry. Ithaca, NY: Cornell Univ. Press.