PATHOGENESIS OF EXPANDED POLYGLUTAMINE REPEATS IN NEURODEGENERATIVE DISEASES: GAINING INSIGHT INTO PROTEIN FOLDING, HYDROGEN BONDING, AND WATER-ACCESSIBILITY BY ADVANCED INFORMATICS AND SIMULATIONS

by

Jingran Wen

A dissertation submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Biomedical Informatics

The University of Utah

May 2017

The University of Utah Graduate School

STATEMENT OF DISSERTATION APPROVAL

The dissertation of **Jingran Wen**

has been approved by the following supervisory committee members:

Julio Cesar Facelli , Chair **11/18/2016**
Date Approved

Daniel R. Scoles , Member **11/18/2016**
Date Approved

Anita Orendt , Member **11/18/2016**
Date Approved

Karen Eilbeck , Member **11/18/2016**
Date Approved

Catherine Janes Staes , Member **11/18/2016**
Date Approved

and by **Wendy W. Chapman** , Chair of

the Department of **Biomedical Informatics**

and by David B. Kieda, Dean of The Graduate School.

ABSTRACT

The unstable expansion of the polyglutamine (polyQ) tract is a critical factor in the pathogenic pathway of at least ten neurodegenerative diseases, including Huntington's disease, spinal and bulbar muscular atrophy (SBMA), dentatorubral-pallidoluysian atrophy (DRPLA), and seven spinocerebellar ataxias, all of which are termed as polyglutamine diseases. One less understood but common feature of polyQ diseases is polyQ protein aggregation. This dissertation explores the protein folding, hydrogen bonding, and water accessibility changes which are induced by the enlargement of the polyQ tract using advanced informatics and computational methods, including protein 3D structure modeling and molecular dynamics simulations. This dissertation also demonstrates that these state-of-the-art computational and informatics methods are powerful tools to provide useful insights into protein aggregation in polyQ diseases.

The enlargement of polyQ segments affects both local and global structures of polyQ proteins as well as their water-accessibility, hydrogen bond patterns, and other structural characteristics. Results from both isolated polyQ and polyQ segments in the context of ataxin-2 and ataxin-3 show that the polyQ tracts increasingly prefer self-interaction as the lengths of the tracts increase, indicating an increased tendency toward aggregation among larger polyQ tracts. These results provide new insights into possible pathogenic mechanisms of polyQ diseases based solely on the increased propensity toward polyQ aggregation and suggest that the modulation of solvent-polyQ interaction may be a possible

therapeutic strategy for treating polyQ diseases.

The analysis pipeline designed and used in this study is an effective way to study the molecular mechanism of polyQ diseases, and can be generalized to study other diseases associated with the protein conformation changes, such as Parkinson’s disease, Alzheimer’s disease, and cancer.

TABLE OF CONTENTS

Chapters

LIST OF TABLES

Tables

LIST OF FIGURES

Figures

ACKNOWLEDGEMENTS

I would like to thank Dr. Julio C. Facelli, my dissertation committee chair, for his guidance and support which made my research a thoughtful and rewarding journey. I would like to thank my dissertation committee members Drs. Daniel R. Scoles, Anita M. Orendt, Karen Eilbeck, and Catherine J. Staes for their input and support since I started this exciting research project. I would like to thank Mr. Richard A. Fay and Ms. Carole M. Fay for their generous Graduate Fellowship which have partially supported my research and study in the third year of my Ph.D. training. I would like to thank my fellow graduate students and colleagues in the Department of Biomedical Informatics, especially those in the Facelli group, for their support, feedback, and friendship. I would like to thank the staff in the Center for High Performance Computing at the University of Utah for their technology support so that all my computational and simulation work was able to run smoothly on the high performance computing servers. I would like to thank Marie Groberg, the writing fellow at the University of Utah Graduate Writing Center, for the guidance and suggestions on part of this dissertation manuscript.

Special thanks to my dear family and friends, for all the needed encouragement and support.

CHAPTER 1

INTRODUCTION AND BACKGROUND

The unstable expansion of the cytosine-adenine-guanine (CAG) repeat in the coding regions of several genes is associated with at least ten neurodegenerative diseases. These diseases are termed polyglutamine (polyQ) diseases as the CAG repeats are translated into polyglutamine in the related proteins. Although the genes related to each polyQ disease were discovered in the early 20th century the pathogenesis mechanism of these diseases is still not well understood. As a result, there is no curative treatment available. The polyQ tract expansion is the only common feature shared by the ten polyQ diseases and has long been considered a key factor in their pathogenic pathway. The expanded polyQ tract can trigger protein misfolding and aggregation which can result in neuron dysfunction. Using advanced informatics and computational methods, this dissertation studies the folding, hydrogen bonding, and water accessibility properties of the polyQ tract – within and without the context of polyQ proteins – to understand, at the molecular level, the pathogenesis of expanded polyQ in polyQ diseases.

Polyglutamine Diseases

PolyQ diseases are a family of neurodegenerative disorders associated with the expansion of CAG repeats in a specific gene [1, 2], in which the CAG repeats are translated into a polyQ tract in the related proteins. To date, at least ten disorders have been described;

these diseases are Huntington's disease (HD) [3], dentatorubral-pallidoluysian atrophy (DRPLA) [4], spinal and bulbar muscular atrophy (SBMA) [5], and spinocerebellar ataxias (SCA) type 1, 2, 3, 6, 7, 8, 17 [6-9]. The polyQ genes express in both patients and normal individuals, but only individuals with a CAG repeat number larger than certain thresholds develop the disease [1]. In addition, the CAG repeat expansion in the ATXN8 gene is related to SCA8 and the pathogenesis of SCA8 also involves the noncoding gene ATXN8OS, with a CTG repeat [10].

PolyQ diseases are relatively rare, and the prevalence of each disorder can vary with geographic location and ethnic background. The estimated prevalence of HD is 2.7 per 100,000 worldwide. Meta-analysis shows a higher prevalence of HD in Europe, North America, and Australia (5.7 per 100,000) than it does in Asia (0.40 per 100,000) [11]. The prevalence of SCA, as a whole, is similar to HD and is estimated to be 2-3 per 100,000 [12], but the frequency can vary with different ethnic groups [13]. SCA3 is the most common spinocerebellar ataxia followed by SCA2, SCA1 and SCA8 [12].

PolyQ diseases share common pathogenic characteristics despite the different affected genes. They are all neurodegenerative diseases characterized by a progressive degeneration of neurons [4]. The number of CAG repeats is related to how quickly the diseases progress, with longer repeats exhibiting earlier and more severe symptoms than those seen in individual with smaller number of repeats. All polyQ diseases, except for SBMA, are autosomal dominant inherited diseases, which means one copy of abnormal CAG repeat genes in each cell is sufficient to cause the disease. Most polyQ disease patients are heterogeneous genotype with only one abnormal copy in the cell [14]. Evidence shows that patients with two abnormal copies experience more severe symptoms than those who carry only one mutant copy [15]. SBMA is X-linked because the CAG-repeat gene, Androgen Receptor (AR), is located in the X chromosome [5]. In males, who have only one copy of the X chromosome, the mutation in this chromosome can cause SBMA. Yet, the specific mechanism leading to neuronal dysfunctions related to polyQ diseases are not well understood [6].

PolyQ disease includes a group of neurodegenerative diseases that display clinical and neuropathological heterogeneity, but a common feature shared by these disorders is the degeneration of a population of neurons in the central nervous system (CNS) [16-18]. The group of disorders has a broad impact on a person's functional abilities in motor [13, 18], and some subtypes might involve cognitive and psychiatric disorders [13]. For some polyQ disease subtypes, the patients may show symptoms of other neurodegenerative diseases, such as Parkinson’s diseases [16, 17] and Amyotrophic Lateral Sclerosis (ALS) [19]. Ataxia is a predominant clinical feature in SCA patients [18], and speech and swallowing are often affected in these patients [18, 20].

Medical diagnoses of polyQ diseases can be made following the appearance of clinical symptoms specific to the diseases. Physical exam, family history, spinal tap, and magnetic resonance imaging (MRI) scanning of the brain and spine are usually included in the diagnosis procedures. Genetic tests detecting the number of CAG repeats in the defective genes can be used to confirm the diagnosis, as similar symptoms are shared among polyQ diseases [13].

PolyQ diseases are fatal and devastating diseases [21, 22] with no curative and/or disease-modifying treatment available [13, 18, 23]. After the onset of the disease, patients’ functional abilities gradually worsen over time [24]. The rate of disease progression and the duration of the disease may vary, but the time from disease onset to death is often about 10 to 30 years [18]. The longer the repeat, the faster the disease progresses [25]. Juvenile Huntington's disease, early-onset form of Huntington's disease that begins in childhood or adolescence, usually results in death within 10 years as the CAG segment in HD gene can repeat more than 60 times in these patients [26]. Eventually, patients with polyQ diseases need help with daily living activities and care [27, 28]. To date, there are no curative treatments for polyQ diseases [13, 18, 23]. Medications are available to help manage the symptoms of the disease, but cannot prevent the physical, mental, and behavioral decline associated with the conditions [29]. An accurate diagnosis of a specific subtype of polyQ disease, especially through a genetic test, can provide great value for making a treatment and care plan [30].

Polyglutamine Disease Related Genes and Proteins

The genetic causes of polyQ diseases were discovered in the 1990's. The expansion of CAG repeats in AR genes was identified as a possible cause of SBMA in 1991, making SBMA the first genetic disorder identified as part of the polyQ neurodegenerative disease group [5]. In 1993, a novel gene with the CAG trinucleotide repeat pattern, later called the HTT gene, was discovered within the HD chromosome, and the unstable expansion of CAG repeats was linked to Huntington's disease [3]. In the same year, the CAG repeat expansion was discovered in the ATXN1 gene, which was directly related to SCA1 [31]. The CAG repeat expansion in six other genes, ATXN3 [6], ATN1 [4], ATXN2 [32], CACNA1A [33], ATXN7, and TBP [34], was identified in the 1990s and associated with these specific types of neurodegenerative diseases.

Polyglutamine genes are scattered throughout different locations in the human genome

(Table 1.1). Apart from the CAG repeat, they share little in gene size, gene sequence, and function, as do the related proteins. In regards to gene size, the huntingtin protein, which is associated with Huntington's disease, can be as long as 3,144 amino acids, whereas the ataxin-3 protein and the TATA-box binding protein, which are associated with SCA3 and SCA17, respectively, are only around 300 amino acids in length. As regards gene sequence, the polyQ tracts lie in different regions of the polyQ proteins, but, in most cases, they are located close to the protein terminus. In addition to the differences in gene size and sequences, the functions of polyglutamine genes and related proteins also vary. The AR gene functions as a steroid-hormone activated transcription factor [35]. Ataxin-1, the protein translated from ATXN1, is reported to be involved in transcriptional repression and it is a component in the Notch signaling pathway [36]. Ataxin-2, the protein translated from ATXN2 gene, is part of the endocytic receptor cycling and affects EGF receptor trafficking [37]. Ataxin-3, the protein encoded by ATXN3, contains ubiquitin interaction motifs (UIM) and the Josehpin domain, and is involved in de-ubiquitination activity [38]. CACNA1A, the gene responsible for SCA6, is involved in voltage-gated calcium channels and mediates a number of calcium-dependent processes [39]. The function of huntingtin protein is still unclear [40, 41], but evidence shows that normal huntingtin is important for neuronal function [42]. Most of the polyQ proteins related to polyQ diseases can bind to a wide range of molecules, which bindings have been identified in a great number of studies. However, the exact function of the polyQ genes and proteins are still poorly understood.

PolyQ Length Dependent Pathogenesis

The number of CAG/polyQ repeats is a key factor for the progression of polyQ diseases [43, 44]. Symptoms occur when the number of consecutive repeats is longer than a critical

Table 1.1 Polyglutamine diseases related genes and proteins

Disease	Gene	Locus	Longest transcript	Location polyQ	Normal repeat length	Pathogenic repeat length
SCA1	ATXN1	6p23	815	N-terminus	6-39	41-83
SCA2	ATXN2	12q24.1	1313	N-terminus	13-31	>=32
SCA3	ATXN3	14q21	364	C-terminus	12-43	60-89
SCA6	CACNA1A	19p13	2512	C-terminus	<18	20-33
SCA7	ATXN7	3p21.1-p12	982	N-terminus	<19	36-460
SCA17	TBP	6q27	339	N-terminus	25-42	49-66
HD	HTT	4p16.3	3144	N-terminus	6-34	>40
DRPLA	ATN1	12p13.31	1190	N-terminus	6-35	>48
SBMA	AR	Xq12	920	N-terminus	< =36	>38

value [1, 17] (Table 1.1). Although it varies, the threshold of most of these diseases is between 32 to 40 repeats [45], except for SCA6 which has a small threshold of around 20 repeats [33]. A larger number of repeats is usually associated with an earlier onset of symptoms [46]. The expanded CAG segments are unstable [47] which can lead to an increase in repeats in successive generations. [48]. This phenomenon is known as genetic anticipation, and leads to an increase in disease severity and an early age of onset [48, 12, 49, 50]. The average age of onset is in the adult years among patients with polyQ diseases [49, 39, 51], but individuals with extremely long repeats may show symptoms at very young age [49, 39, 51, 48]. For instance, SCA2 patients with 32 or 33 repeats tend to show symptoms of SCA2 in late adulthood, whereas patients with more than 45 repeats usually have signs and symptoms by their teenage years [18].

There are still some unanswered but interesting questions about polyQ length dependent features in polyQ disease, such as why do symptoms only occur in patients carrying the protein in which the length of polyQ tract is above a threshold and why do longer lengths above these thresholds lead to an early age of onset and more severe symptoms? Although *in vivo* experimental studies yield evidence that short-repeated polyQ tracts may also aggregate [52], longer polyQ tracts provoke an earlier appearance of disease and faster progression [53]. One possible explanation is that the aggregation propensity of long polyQ tracts is higher than that of the short ones [53]. Therefore, people with short/normal repeats may not display symptoms during their lifetime. In some of the polyQ diseases, including Huntington's disease, SCA1, SCA3, SCA7, SCA17, and SBMA, there exists the reduced-penetrance form of polyQ tracts, and individuals carrying the reduced-penetrance gene allele may or may not be affected. Take Huntington's disease for example, the smallest repeats related to clinical symptoms is reported to vary from 36 to 40, while the upper range of normal repeats is reported from 30 to 39 repeats. The polyQ gene with 36 to 39 repeats is reported as the reduced-penetrance allele.

Protein Misfolding and Aggregation

The most important feature shared by polyQ proteins is their propensity to form oligomers and aggregates. The intranuclear inclusion bodies are found in the affected neurons in almost all types of the polyQ diseases [54, 55, 49, 56-60], except for SCA2 [61, 62] and SCA6 [63], in which the aggregations are found in the cytoplasm. However, the molecular mechanisms of polyQ protein aggregation remain unclear [2]. *It is also unclear if aggregation is a cause of the polyQ diseases or a consequence of them,* as neurological symptoms may show up before protein aggregation can be detected in several transgenic models [64]. Evidence also indicates that large visible inclusions may be protective by recruiting and enhancing the degradation of toxic polyQ proteins [65], whilst the oligomers or microaggregates might be the toxic formations [66].

Evidence also indicates that the expanded polyQ tracts alter protein conformation [47], which could allow the mutant protein to recruit normal cellular proteins through a series of abnormal interactions. Some interactions can be enhanced, whereas others may be lost or remain unchanged [47]. For example, in the SCA1 transgenic mice model, ataxin-1 with expanded polyQ can alter the transcription of several genes, including genes involved in signal transduction and calcium homeostasis [67].

In addition to the polyQ proteins, other proteins may be included in the aggregation to make large inclusions. A variety of glutamine-rich transcriptional regulators are known to localize into polyQ inclusions, both in cellular and animal models. Therefore, the aggregation could initiate cellular dysfunction by sequestering the respective proteins, both mutant polyQ proteins and other proteins, from their normal localization, causing imbalance of proteins in cells, and compromising their functions [2].

Proteolysis cleavage is confirmed in several polyQ proteins, including ataxin-3 [68], huntingtin [69], atrophin-1 [70], and ataxin-7 [71]. These proteins can be degenerated by proteases into fragments. As shown in Table 1.1, polyQ repeats are, in general, located in either the N-terminus or the C-terminus of polyQ proteins, which location facilitates the release of polyQ fragments. Fragments containing polyQ tracts might display misfolded structures and cause the micro-aggregation of oligomers [1]. These aggregated oligomers may cause the dysfunction of related proteins and show toxicity properties to the cell environment, which have a potential role in polyQ disease pathogenesis.

As aggregation is a hallmark of polyQ disease progression [72], the knowledge of aggregation mechanisms, at the molecular level, can provide new avenues to explore common therapeutic methods for treating these disorders. Also, as protein misfolding and aggregation are common features in other neurodegenerative disorders, such as Alzheimer’s disease, Parkinson’s disease, and ALS [73, 74], the knowledge of polyQ misfolding and aggregation identified here could also shed light on the pathogenic mechanism of other neurodegenerative diseases.

Hydrophobicity and Water Accessibility

Known intrinsic properties that can effect protein aggregation include charge, hydrophobic/hydrophilic patterns, and secondary structures [75]. Among these factors, hydrophobicity plays a significant role in setting the conditions governing protein aggregation [75, 76]. The more hydrophobic a sequence is the more readily it aggregates.

As a polar region, polyQ segments have the propensity to aggregate, and in aqueous milieus, an individual polyQ peptide prefers collapsed structures that minimize interactions with a surrounding solvent [77]. Perutz [78] proposes that expansion of glutamine repeats beyond a certain length may lead to a phase change from random coils to hydrogen-bonded self-associated confirmations. Consistent with this hypothesis, both experimental and computational studies, show that polyQ tracts can change the conformation to hydrogen-bonded structures during the formation of aggregation [79-83]. However, these hypotheses have not been verified by detailed 3D protein structure studies in the context of the full-length polyQ proteins or by using full atomic explicit solvent molecular dynamics simulations.

Conformational Study of PolyQ Containing Proteins

The elongation of the polyQ tract can lead to misfolding of polyQ proteins, which affects the normal functions of these proteins [55]. Therefore, a knowledge of protein structure changes as the function of the lengthening of the polyQ tract can bring insights into the molecular mechanism of pathogenesis in polyQ diseases.

However, due to the large protein size and aggregation property of polyQ proteins, it is difficult to obtain high-quality 3D structures of full-length polyQ proteins through experimental studies, such as X-ray and nuclear magnetic resonance spectroscopy (NMR). Until now, few 3D structures of polyQ protein are available in the Research Collaboratory for Structural Bioinformatics (RCSB) Protein Data Bank (PDB) [84], and none of these structures represent full-length polyQ proteins. Even fewer contain the polyQ tract in the structures. The only known structures of polyQ protein segments with polyQ tract are the crystal structures of N-terminal huntingtin protein with 17 [85] and 36 glutamine repeats

[86]. There are several structures of polyQ protein segments with no polyQ tract, such as solution structures of the Josephin domain in ataxin-3 [87] and solution structures of the UIM domains of ataxin-3 complexed with ubiquitin [88]. With state-of-the-art experimental methods, such as circular dichroism (CD) and fluorescence spectroscopy, it is possible to measure changes in the surface exposure of polyQ proteins induced by polyQ tract expansion[89], but these techniques cannot provide information with enough detail to address the structural changes at the residue and atomic levels.

By resolving the structure of a short polyQ monomer of 15Qs, Perutz puts forward the hypothesis that expanded polyQ tracts form misfolded structures which can lead to protein aggregation [90]. In Perutz's theory, the elongated polyQ tract may form a "polar zipper" conformation, a structure with β-strands and an extensive hydrogen-bonded network [90]. The polar zipper structure is latter interpreted as the cross-beta structure. Perutz's polyQ structure model led to a series of computational studies on the thermodynamic and kinetic characters of polyQ tracts. These computational studies yield several plausible structures for polyQ tracts [91, 80]. PolyQ with certain structures therefore can initialize the aggregation which leads to the formation of amyloid fibrils [53, 92]. Yet, the precise mechanism involved in polyQ aggregation is still largely unknown.

Protein 3D Structure Prediction

Protein 3D structure prediction is a set of computational techniques with the capability to infer the 3D models of a protein using its amino acid sequence [93]. Protein 3D structure prediction plays a critical role in biomedical research, as it is an essential tool to predict structures of biomedical molecules for which no experimental structures are yet obtainable. This technique is widely used for exploring protein folding and identifying potential protein binding sites that can be used for *in silico* drug design [94, 95].

The previous solved structures, or template, can be used in protein structure prediction. Based on whether templates are used or not, protein structure prediction methods can be divided into two categories: comparative protein modeling and *ab initio* modeling [96]. In comparative protein modeling, templates are used as the starting point. Comparative protein modeling can be split into two groups: homology modeling and fold recognition. Homology modeling is based on the hypothesis that proteins sharing sequence similarity may also share structural similarity. The sequences of templates are the homologies of the protein sequence to be predicted (the target). Homology modeling can generate high-quality models if sequence homologies exist in the template database. However, homology modeling is not suitable for proteins for which homological templates are not available. Also, protein structures are more conserved in evolution than protein sequences [97], therefore proteins with similar structures may not be homologous in sequences [98].

Fold recognition, also known as threading, is designed to model protein 3D structures based on structural similarity. Threading is a technique to match the query protein sequence directly on to the known 3D structures [96, 99], which aims to recognize fold similarity even when there is no evolutional relationship between the target and the structure templates. The threading method differs from the homology modeling as threading is used for proteins which do not have solved structures of their homologous proteins, whereas homology modeling is used for proteins which have their homologous protein structures solved. The protein threading method consists of several general steps: 1) the construction of a structure template database, or the selection of structure templates from a protein structure database, such as RCSB PDB database; 2) the design of a scoring function to measure the similarity or fitness between the target and the templates; 3) threading alignment to align the target sequence with each of the structure templates by optimizing the designed scoring function; and 4) the threading prediction. This last step determines the most probable structures from the structure pools generated in the previous steps. Finally, the structure models are constructed by placing the backbone atoms of the target sequence at their aligned backbone positions in the selected structural templates.

Some preconditions are necessary for the success of fold recognition, such as the availability of good structural templates in the template database and the ability to recognize good templates. In regard to the former precondition, structure databases, such as RCSB PDB, are good repositories for structure templates. In regards to the latter, to date, some threading methods, such as threading assembly refinement (TASSER), can successfully build high-quality full-length protein models with an average root-mean-square-deviation (RMSD) of 2.25 Å. I-TASSER is a protein structure prediction package which implements the TASSER threading method. It is ranked the first of template-based prediction methods in several Critical Assessment of Protein Structure Prediction (CASP) competitions [96] and widely recognized as one of the best methods for protein structure prediction.

Although template-based modeling is an essential approach for protein structure prediction [100], there are proteins with no sufficient sequence homologies and structural analogies. In this situation, the current template-based modeling lack the ability to successfully find appropriate templates. Here, the structure prediction has to be done from the first principle, or *ab initio*. The *ab initio* methods try to predict protein 3D structures based on physical laws rather than the previous solved structures. According to the physical laws, molecules adopt conformations where their energy achieves the global minima on the potential energy surface. It is difficult to search the global minima of the complex potential energy at the all-atom level for large proteins with more than 100 amino acids. The conformation sampling needs substantial computational resources, especially for proteins with large sizes. This limits the *ab initio* methods to prediction of proteins with relatively small sizes. S. Ołdziej et al. designed a hierarchical physics-based protein structure prediction methods which searches the global minima in a united atom force field. When compared the predicted structures of a 102-residue target to the native structure, the RMSD value is 7.4 Å [101].

Currently, the boundary between template-based methods and *ab initio* methods is blurred. These two types of methods can be combined to improve prediction performance. For example, the Rosetta *ab initio* protein structure modeling application uses small structure segments, usually with the length of 3 or 9 amino acids, to generate initial structures [102]. I-TASSER uses the *ab initio* methods to do the modeling when structural templates are not available [96].

With the improvement of computational capacity, 3D structure prediction algorithms can now predict protein structures at the atomic level, leading to practical applications to understand pathogenesis at this level of resolution. For example, I-TASSER [96, 93] has been successfully applied in structural and functional modeling of proteins related to studies of aging, cancer, diabetes, and other diseases [94, 103, 104]. These complementary computational studies bring new insights into pathogenesis and potential treatments for these diseases. The advances in protein structure prediction are systematically reviewed at regular blind tests [105], such as the CASP competition which is a community-wide competition to test state-of-the-art protein structure prediction methods.

Molecular Dynamics Simulation

Molecular dynamics (MD) simulations are computer simulations that intend to model the physical movement of atoms and molecules. During the simulation, the atoms and molecules interact in the simulation system for a period of time, thus allowing the analysis of the atom movement and bonds between the atoms. MD simulation is an excellent tool for the study of the thermodynamic and kinetic properties of biomolecules and many publications report both their success and limitations [106, 107].

In MD simulations, mathematical functions are used to describe the potential energy of the system particles, the forms and parameters of which are called the force field. “All-atom” force fields provide parameters for every type of atom in a system. 3D structure prediction finds the static structures with the lowest energy, whereas MD simulations can reveal the process of the folding and motion of the molecules.

Using MD simulations, several researches have studied the stability, folding, and aggregation properties of polyQ tracts [83, 108-110]. A brief summary of the simulation work on polyQ is listed in the “Introduction” section in Chapter 5 of this dissertation. In this dissertation, MD simulations are applied to study secondary structure ensemble, water accessibility, and other structural propensities of polyQ segments to investigate the mechanisms of polyQ protein aggregation.

Specific Aims of This Dissertation

In this dissertation, state-of-the-art informatics and computational simulation techniques are used to demonstrate that they can be applied to enhance the understanding of the pathogenesis mechanisms of polyQ diseases. The effects of polyQ enlargement on protein folding, conformational stability, and water accessibility of polyQ proteins are studied to explore the pathogenesis of polyQ diseases. Three specific aims are addressed in this dissertation:

1) To compare and validate the predicted structures in polyQ disease relevant proteins (Chapter 2);

2) To study the effect of the polyQ lengthening on protein folding, hydrogen bonding, and water accessibility of full-length polyQ proteins using protein structure prediction methods (Chapter 3 and Chapter 5);

3) To study the effect of solvation on the folding stability and solvent-polyQ interaction of the polyQ segment using atomistic MD models (Chapter 4).

References

1. Shao J, Diamond MI. Polyglutamine diseases: emerging concepts in pathogenesis and therapy. Hum Mol Genet. 2007;16(R2):R115-R23.

2. Michalik A, Van Broeckhoven C. Pathogenesis of polyglutamine disorders: aggregation revisited. Hum Mol Genet. 2003;12 Spec No 2:R173-86.

3. Group THsDcR. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. Cell. 1993;72(6):971-83.

4. Yazawa I, Nukina N, Hashida H, Goto J, Yamada M, Kanazawa I. Abnormal gene product identified in hereditary dentatorubral-pallidoluysian atrophy (DRPLA) brain. Nat Genet. 1995;10(1):99-103.

5. La Spada AR, Wilson EM, Lubahn DB, Harding AE, Fischbeck KH. Androgen receptor gene mutations in X-linked spinal and bulbar muscular atrophy. Nature. 1991;352(6330):77-9.

6. Kawaguchi Y, Okamoto T, Taniwaki M, Aizawa M, Inoue M, Katayama S et al. CAG expansions in a novel gene for Machado-Joseph disease at chromosome 14q32.1. Nat Genet. 1994;8(3):221-8.

7. Imbert G, Saudou F, Yvert G, Devys D, Trottier Y, Garnier JM et al. Cloning of

the gene for spinocerebellar ataxia 2 reveals a locus with high sensitivity to expanded CAG/glutamine repeats. Nat Genet. 1996;14(3):285-91.

8. David G, Abbas N, Stevanin G, Durr A, Yvert G, Cancel G et al. Cloning of the SCA7 gene reveals a highly unstable CAG repeat expansion. Nat Genet. 1997;17(1):65-70.

9. Zoghbi HY, Jodice C, Sandkuijl LA, Kwiatkowski TJ, Jr., McCall AE, Huntoon SA et al. The gene for autosomal dominant spinocerebellar ataxia (SCA1) maps telomeric to the HLA complex and is closely linked to the D6S89 locus in three large kindreds. Am J Hum Genet. 1991;49(1):23-30.

10. Moseley ML, Zu T, Ikeda Y, Gao W, Mosemiller AK, Daughters RS et al. Bidirectional expression of CUG and CAG expansion transcripts and intranuclear polyglutamine inclusions in spinocerebellar ataxia type 8. Nat Genet. 2006;38(7):758-69.

11. Pringsheim T, Wiltshire K, Day L, Dykeman J, Steeves T, Jette N. The incidence and prevalence of Huntington's disease: a systematic review and meta-analysis. Mov Disord. 2012;27(9):1083-91.

12. Whaley N, Fujioka S, Wszolek Z. Autosomal dominant cerebellar ataxia type I: a review of the phenotypic and genotypic characteristics. Orphanet J Rare Dis. 2011;6(1):33.

13. Fan HC, Ho LI, Chi CS, Chen SJ, Peng GS, Chan TM et al. Polyglutamine (PolyQ) diseases: genetics to treatments. Cell Transplant. 2014;23(4-5):441-58.

14. Gusella JF, MacDonald ME. Molecular genetics: unmasking polyglutamine triggers in neurodegenerative disease. Nat Rev Neurosci. 2000;1(2):109-15.

15. Squitieri F, Gellera C, Cannella M, Mariotti C, Cislaghi G, Rubinsztein DC et al. Homozygosity for CAG mutation in Huntington disease is associated with a more severe clinical course. Brain. 2003;126(4):946-55.

16. Bettencourt C, Lima M. Machado-Joseph Disease: from first descriptions to new perspectives. Orphanet J Rare Dis. 2011;6:35.

17. Magana JJ, Velazquez-Perez L, Cisneros B. Spinocerebellar ataxia type 2: clinical presentation, molecular mechanisms, and therapeutic perspectives. Mol Neurobiol. 2013;47(1):90-104.

18. Matilla-Duenas A, Corral-Juan M, Volpini V, Sanchez I. The spinocerebellar ataxias: clinical aspects and molecular genetics. Adv Exp Med Biol. 2012;724:351-74.

19. Tazen S, Figueroa K, Kwan JY, Goldman J, Hunt A, Sampson J et al. Amyotrophic lateral sclerosis and spinocerebellar ataxia type 2 in a family with full CAG repeat expansions of ATXN2. JAMA Neurol. 2013;70(10):1302-4.

20. Matilla-Duenas A, Sanchez I, Corral-Juan M, Davalos A, Alvarez R, Latorre P.

Cellular and molecular pathways triggering neurodegeneration in the spinocerebellar ataxias. Cerebellum. 2010;9(2):148-66.

21. Zhang K, Yi F, Liu G-H, Belmonte JCI. Huntington's disease: dancing in a dish. Cell Res. 2012;22(12):1627-30.

22. Selkoe DJ. Folding proteins in fatal ways. Nature. 2003;426(6968):900-4.

23. Ross CA, Tabrizi SJ. Huntington's disease: from molecular pathogenesis to clinical treatment. Lancet Neurol. 2011;10(1):83-98.

24. Kimura Y, Kakizuka A. Polyglutamine diseases and molecular chaperones. IUBMB life. 2003;55(6):337-45.

25. Andrew SE, Goldberg YP, Kremer B, Telenius H, Theilmann J, Adam S et al. The relationship between trinucleotide (CAG) repeat length and clinical features of Huntington's disease. Nat Genet. 1993;4(4):398-403.

26. Quarrell OWJ, Nance MA, Nopoulos P, Paulsen JS, Smith JA, Squitieri F. Managing juvenile Huntington’s disease. Neurodegener Dis Manag. 2013; doi:10.2217/nmt.13.18.

27. Skirton H, Williams JK, Jackson Barnette J, Paulsen JS. Huntington disease: families’ experiences of healthcare services. J Adv Nurs. 2010;66(3):500-10.

28. Zielonka D, Mielcarek M, Landwehrmeyer GB. Update on Huntington's disease: advances in care and emerging therapeutic options. Parkinsonism Relat Disord. 2015;21(3):169-78.

29. Imarisio S, Carmichael J, Korolchuk V, Chen CW, Saiki S, Rose C et al. Huntington's disease: from pathology and genetics to potential therapies. Biochem J. 2008;412(2):191-209.

30. Powell A, Chandrasekharan S, Cook-Deegan R. Spinocerebellar ataxia: patient and health professional perspectives on whether and how patents affect access to clinical genetic testing. Genet Med. 2010;12 Suppl 4:S83-S110.

31. Banfi S, Chung MY, Kwiatkowski TJ, Jr., Ranum LP, McCall AE, Chinault AC et al. Mapping and cloning of the critical region for the spinocerebellar ataxia type 1 gene (SCA1) in a yeast artificial chromosome contig spanning 1.2 Mb. Genomics. 1993;18(3):627-35.

32. Pulst SM, Nechiporuk A, Nechiporuk T, Gispert S, Chen XN, Lopes-Cendes I et al. Moderate expansion of a normally biallelic trinucleotide repeat in spinocerebellar ataxia type 2. Nat Genet. 1996;14(3):269-76.

33. Zhuchenko O, Bailey J, Bonnen P, Ashizawa T, Stockton DW, Amos C et al. Autosomal dominant cerebellar ataxia (SCA6) associated with small polyglutamine

expansions in the alpha 1A-voltage-dependent calcium channel. Nat Genet. 1997;15(1):62-9.

34. Koide R, Kobayashi S, Shimohata T, Ikeuchi T, Maruyama M, Saito M et al. A neurological disease caused by an expanded CAG trinucleotide repeat in the TATA-binding protein gene: a new polyglutamine disease? Hum Mol Gen. 1999;8(11):2047-53.

35. Estébanez-Perpiñá E, Moore JMR, Mar E, Delgado-Rodrigues E, Nguyen P, Baxter JD et al. The Molecular Mechanisms of Coactivator Utilization in Ligand-dependent Transactivation by the Androgen Receptor. J Biol Chem. 2005;280(9):8060-8.

36. Tong X, Gui H, Jin F, Heck BW, Lin P, Ma J et al. Ataxin-1 and Brother of ataxin-1 are components of the Notch signalling pathway. EMBO Rep. 2011;12(5):428-35.

37. Nonis D, Schmidt MH, van de Loo S, Eich F, Dikic I, Nowock J et al. Ataxin-2 associates with the endocytosis complex and affects EGF receptor trafficking. Cell Signal. 2008;20(10):1725-39.

38. Mao Y, Senic-Matuglia F, Di Fiore PP, Polo S, Hodsdon ME, De Camilli P. Deubiquitinating function of ataxin-3: insights from the solution structure of the Josephin domain. Proc Natl Acad Sci U S A. 2005;102(36):12700-5.

39. Solodkin A, Gomez CM. Chapter 29 - Spinocerebellar ataxia type 6. In: Sankara HS, Alexandra D, editors. Handbook of Clinical Neurology. Elsevier; 2012. p. 461-73.

40. Zheng Q, Joinnides M. Hunting for the function of Huntingtin. Dis Model Mech. 2009;2(5-6):199-200.

41. Cisbani G, Cicchetti F. An in vitro perspective on the molecular mechanisms underlying mutant huntingtin protein toxicity. Cell Death Dis. 2012;3:e382.

42. Cattaneo E, Zuccato C, Tartari M. Normal huntingtin function: an alternative approach to Huntington's disease. Nat Rev Neurosci. 2005;6(12):919-30.

43. Margulis BA, Vigont V, Lazarev VF, Kaznacheyeva EV, Guzhova IV. Pharmacological protein targets in polyglutamine diseases: mutant polypeptides and their interactors. FEBS Lett. 2013;587(13):1997-2007.

44. Riley BE, Orr HT. Polyglutamine neurodegenerative diseases and regulation of transcription: assembling the puzzle. Genes Dev. 2006;20(16):2183-92.

45. Zhou ZL, Zhao JH, Liu HL, Wu JW, Liu KT, Chuang CK et al. The possible structural models for polyglutamine aggregation: a molecular dynamics simulations study. J Biomol Struct Dyn. 2011;28(5):743-58.

46. Walters RH, Murphy RM. Examining polyglutamine peptide length: a connection between collapsed conformations and increased aggregation. J Mol Biol. 2009;393(4):978-92.

47. Gatchel JR, Zoghbi HY. Diseases of Unstable Repeat Expansion: Mechanisms and Common Principles. Nat Rev Genet. 2005;6(10):743-55.

48. Walker FO. Huntington's disease. Lancet. 2007;369(9557):218-28.

49. Zoghbi HY, Orr HT. Pathogenic mechanisms of a polyglutamine-mediated neurodegenerative disease, spinocerebellar ataxia type 1. J Biol Chem. 2009;284(12):7425-9.

50. Giunti P, Sabbadini G, Sweeney MG, Davis MB, Veneziano L, Mantuano E et al. The role of the SCA2 trinucleotide repeat expansion in 89 autosomal dominant cerebellar ataxia families. Frequency, clinical and genetic correlates. Brain. 1998;121 (Pt 3):459-67.

51. Martin JJ. Chapter 30 - Spinocerebellar ataxia type 7. In: Sankara HS, Alexandra D, editors. Handbook of Clinical Neurology. Elsevier; 2012. p. 475-91.

52. Chen S, Berthelier V, Yang W, Wetzel R. Polyglutamine aggregation behavior in vitro supports a recruitment mechanism of cytotoxicity. J Mol Biol . 2001;311(1):173-82.

53. Chen S, Ferrone FA, Wetzel R. Huntington's disease age-of-onset linked to polyglutamine aggregation nucleation. Proc Natl Acad Sci U S A. 2002;99(18):11884-9.

54. DiFiglia M, Sapp E, Chase KO, Davies SW, Bates GP, Vonsattel JP et al. Aggregation of huntingtin in neuronal intranuclear inclusions and dystrophic neurites in brain. Science. 1997;277(5334):1990-3.

55. Matilla-Duenas A, Ashizawa T, Brice A, Magri S, McFarland KN, Pandolfo M et al. Consensus paper: pathological mechanisms underlying neurodegeneration in spinocerebellar ataxias. Cerebellum. 2014;13(2):269-302.

56. Paulson HL, Perez MK, Trottier Y, Trojanowski JQ, Subramony SH, Das SS et al. Intranuclear inclusions of expanded polyglutamine protein in spinocerebellar ataxia type 3. Neuron. 1997;19(2):333-44.

57. Stenoien DL, Cummings CJ, Adams HP, Mancini MG, Patel K, DeMartino GN et al. Polyglutamine-expanded androgen receptors form aggregates that sequester heat shock proteins, proteasome components and SRC-1, and are suppressed by the HDJ-2 chaperone. Hum Mol Gen. 1999;8(5):731-41.

58. Hayashi Y, Kakita A, Yamada M, Koide R, Igarashi S, Takano H et al. Hereditary dentatorubral-pallidoluysian atrophy: detection of widespread ubiquitinated neuronal and glial intranuclear inclusions in the brain. Acta Neuropathol. 1998;96(6):547-52.

59. Holmberg M, Duyckaerts C, Durr A, Cancel G, Gourfinkel-An I, Damier P et al. Spinocerebellar ataxia type 7 (SCA7): a neurodegenerative disorder with neuronal intranuclear inclusions. Hum Mol Gen. 1998;7(5):913-8.

60. Nakamura K, Jeong SY, Uchihara T, Anno M, Nagashima K, Nagashima T et al.

SCA17, a novel autosomal dominant cerebellar ataxia caused by an expanded polyglutamine in TATA-binding protein. Hum Mol Gen. 2001;10(14):1441-8.

61. Huynh DP, Del Bigio MR, Ho DH, Pulst SM. Expression of ataxin-2 in brains from normal individuals and patients with Alzheimer's disease and spinocerebellar ataxia 2. Ann Neurol. 1999;45(2):232-41.

62. Huynh DP, Figueroa K, Hoang N, Pulst SM. Nuclear localization or inclusion body formation of ataxin-2 are not necessary for SCA2 pathogenesis in mouse or human. Nat Genet. 2000;26(1):44-50.

63. Ishikawa K, Fujigasaki H, Saegusa H, Ohwada K, Fujita T, Iwamoto H et al. Abundant expression and cytoplasmic aggregations of α1A voltage-dependent calcium channel protein associated with neurodegeneration in spinocerebellar ataxia type 6. Hum Mol Gen. 1999;8(7):1185-93.

64. Yoo SY, Pennesi ME, Weeber EJ, Xu B, Atkinson R, Chen S et al. SCA7 knockin mice model human SCA7 and reveal gradual accumulation of mutant ataxin-7 in neurons and abnormalities in short-term plasticity. Neuron. 2003;37(3):383-401.

65. Taylor JP, Tanaka F, Robitschek J, Sandoval CM, Taye A, Markovic-Plese S et al. Aggresomes protect cells by enhancing the degradation of toxic polyglutamine-containing protein. Hum Mol Gen. 2003;12(7):749-57.

66. Sanchez I, Mahlke C, Yuan J. Pivotal role of oligomerization in expanded polyglutamine neurodegenerative disorders. Nature. 2003;421(6921):373-9.

67. Lin X, Antalffy B, Kang D, Orr HT, Zoghbi HY. Polyglutamine expansion down-regulates specific neuronal genes before pathologic changes in SCA1. Nat Neurosci. 2000;3(2):157-63.

68. Berke SJ, Schmied FA, Brunt ER, Ellerby LM, Paulson HL. Caspase-mediated proteolysis of the polyglutamine disease protein ataxin-3. J Neurochem. 2004;89(4):908-18.

69. Sieradzan KA, Mechan AO, Jones L, Wanker EE, Nukina N, Mann DM. Huntington's disease intranuclear inclusions contain truncated, ubiquitinated huntingtin protein. Exp Neurol. 1999;156(1):92-9.

70. Schilling G, Wood JD, Duan K, Slunt HH, Gonzales V, Yamada M et al. Nuclear accumulation of truncated atrophin-1 fragments in a transgenic mouse model of DRPLA. Neuron. 1999;24(1):275-86.

71. Garden GA, Libby RT, Fu YH, Kinoshita Y, Huang J, Possin DE et al. Polyglutamine-expanded ataxin-7 promotes non-cell-autonomous purkinje cell degeneration and displays proteolytic cleavage in ataxic transgenic mice. J Neurosci. 2002;22(12):4897-905.

72. Todd TW, Lim J. Aggregation formation in the polyglutamine diseases: protection at a cost? Mol Cells. 2013;36(3):185-94.

73. Scherzinger E, Sittler A, Schweiger K, Heiser V, Lurz R, Hasenbank R et al. Self-assembly of polyglutamine-containing huntingtin fragments into amyloid-like fibrils: implications for Huntington's disease pathology. Proc Natl Acad Sci U S A. 1999;96(8):4604-9.

74. Soto C. Unfolding the role of protein misfolding in neurodegenerative diseases. Nat Rev Neurosci. 2003;4(1):49-60.

75. Zbilut JP, Colosimo A, Conti F, Colafranceschi M, Manetti C, Valerio M et al. Protein aggregation/folding: the role of deterministic singularities of sequence hydrophobicity as determined by nonlinear signal analysis of acylphosphatase and Abeta(1-40). Biophys J. 2003;85(6):3544-57.

76. Chiti F, Stefani M, Taddei N, Ramponi G, Dobson CM. Rationalization of the effects of mutations on peptide and protein aggregation rates. Nature. 2003;424(6950):805-8.

77. Vitalis A, Wang X, Pappu RV. Atomistic simulations of the effects of polyglutamine chain length and solvent quality on conformational equilibria and spontaneous homodimerization. J Mol Biol. 2008;384(1):279-97.

78. Perutz MF. Glutamine repeats and inherited neurodegenerative diseases: molecular aspects. Curr Opin Struct Biol. 1996;6(6):848-58.

79. Khare SD, Ding F, Gwanmesia KN, Dokholyan NV. Molecular origin of polyglutamine aggregation in neurodegenerative diseases. PLoS Comput Biol. 2005;1(3):230-5.

80. Ogawa H, Nakano M, Watanabe H, Starikov EB, Rothstein SM, Tanaka S. Molecular dynamics simulation study on the structural stabilities of polyglutamine peptides. Comput Biol Chem. 2008;32(2):102-10.

81. Marchut AJ, Hall CK. Side-chain interactions determine amyloid formation by model polyglutamine peptides in molecular dynamics simulations. Biophys J. 2006;90(12):4574-84.

82. Marchut AJ, Hall CK. Effects of chain length on the aggregation of model polyglutamine peptides: molecular dynamics simulations. Proteins. 2007;66(1):96-109.

83. Nakano M, Watanabe H, Rothstein SM, Tanaka S. Comparative characterization of short monomeric polyglutamine peptides by replica exchange molecular dynamics simulation. J Phys Chem B. 2010;114(20):7056-61.

84. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H et al. The Protein Data Bank. Nucleic Acids Res. 2000;28(1):235-42.

85. Kim MW, Chelliah Y, Kim SW, Otwinowski Z, Bezprozvanny I. Secondary structure of Huntingtin amino-terminal region. Structure. 2009;17(9):1205-12.

86. Kim M. Beta conformation of polyglutamine track revealed by a crystal structure of Huntingtin N-terminal region with insertion of three histidine residues. Prion. 2013;7(3):221-8.

87. Nicastro G, Menon RP, Masino L, Knowles PP, McDonald NQ, Pastore A. The solution structure of the Josephin domain of ataxin-3: structural determinants for molecular recognition. Proc Natl Acad Sci U S A. 2005;102(30):10493-8.

88. Song A-X, Zhou C-J, Peng Y, Gao X-C, Zhou Z-R, Fu Q-S et al. Structural transformation of the tandem ubiquitin-interacting motifs in ataxin-3 and their cooperative interactions with ubiquitin chains. PLoS One. 2010;5(10):e13202.

89. Tobelmann MD, Murphy RM. Location trumps length: polyglutamine-mediated changes in folding and aggregation of a host protein. Biophys J. 2011;100(11):2773-82.

90. Perutz MF. Glutamine repeats and neurodegenerative diseases: molecular aspects. Trends Biochem Sci. 1999;24(2):58-63.

91. Stork M, Giese A, Kretzschmar HA, Tavan P. Molecular dynamics simulations indicate a possible role of parallel beta-helices in seeded aggregation of poly-Gln. Biophys J. 2005;88(4):2442-51.

92. Kar K, Jayaraman M, Sahoo B, Kodali R, Wetzel R. Critical nucleus size for disease-related polyglutamine aggregation is repeat-length dependent. Nat Struct Mol Biol. 2011;18(3):328-36.

93. Zhang Y. I-TASSER server for protein 3D structure prediction. BMC Bioinformatics. 2008;9:40.

94. de Carvalho MD, De Mesquita JF. Structural modeling and in silico analysis of human superoxide dismutase 2. PLoS One. 2013;8(6):e65558.

95. Maurer-Stroh S, Debulpaep M, Kuemmerer N, Lopez de la Paz M, Martins IC, Reumers J et al. Exploring the sequence determinants of amyloid structure using position-specific scoring matrices. Nat Methods. 2010;7(3):237-42.

96. Roy A, Kucukural A, Zhang Y. I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protocols. 2010;5(4):725-38.

97. Illergard K, Ardell DH, Elofsson A. Structure is three to ten times more conserved than sequence--a study of structural response in protein cores. Proteins. 2009;77(3):499-508.

98. Grishin NV. Fold change in evolution of protein structures. J Struct Biol. 2001;134(2-3):167-85.

99. Zhang Y. Progress and challenges in protein structure prediction. Curr Opin Struct Biol. 2008;18(3):342-8.

100. Huang YJ, Mao B, Aramini JM, Montelione GT. Assessment of template-based protein structure predictions in CASP10. Proteins. 2014;82 Suppl 2:43-56.

101. Oldziej S, Czaplewski C, Liwo A, Chinchio M, Nanias M, Vila JA et al. Physics-based protein-structure prediction using a hierarchical protocol based on the UNRES force field: assessment in two blind tests. Proc Natl Acad Sci U S A. 2005;102(21):7547-52.

102. Leaver-Fay A, Tyka M, Lewis SM, Lange OF, Thompson J, Jacak R et al. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. Methods Enzymol. 2011;487:545-74.

103. Noureldein MH. In silico discovery of a perilipin 1 inhibitor to be used as a new treatment for obesity. Eur Rev Med Pharmacol Sci. 2014;18(4):457-60.

104. Anbarasu K, Jayanthi S. Structural modeling and molecular dynamics studies on the human LMTK3 domain and the mechanism of ATP binding. Mol Biosyst. 2014;10(5):1139-45.

105. Moult J, Fidelis K, Kryshtafovych A, Schwede T, Tramontano A. Critical assessment of methods of protein structure prediction (CASP) - round x. Proteins. 2014;82:1-6.

106. Adcock SA, McCammon JA. Molecular dynamics: survey of methods for simulating the activity of proteins. Chem Rev. 2006;106(5):1589-615.

107. Durrant J, McCammon JA. Molecular dynamics simulations and drug discovery. BMC Biol. 2011;9(1):71.

108. Wang Y, Voth GA. Molecular dynamics simulations of polyglutamine aggregation using solvent-free multiscale coarse-grained models. J Phys Chem B. 2010;114(26):8735-43.

109. Cote S, Wei G, Mousseau N. All-atom stability and oligomerization simulations of polyglutamine nanotubes with and without the 17-amino-acid N-terminal fragment of the Huntingtin protein. J Phys Chem B. 2012;116(40):12168-79.

110. Miettinen M, Knecht V, Monticelli L, Ignatova Z. Assessing polyglutamine conformation in the nucleating event by molecular dynamics simulations. J Phys Chem B. 2012;116(34):10259-65.

CHAPTER 2

STRUCTURE PREDICTION OF POLYGLUTAMINE DISEASE PROTEINS: COMPARISON OF METHODS [a]

Jingran Wen[1], Daniel R Scoles[2], Julio C Facelli[1]

[1]Department of Biomedical Informatics, and [2]Department of Neurology,

University of Utah, Salt Lake City, Utah, U.S.A

Abstract

Background

The expansion of polyglutamine (polyQ) repeats in several unrelated proteins is associated with at least ten neurodegenerative diseases. The length of the polyQ regions plays an important role in the progression of the diseases. The number of glutamines (Q) is inversely related to the onset age of these polyglutamine diseases, and the expansion of polyQ repeats has been associated with protein misfolding. However, very little is known about the structural changes induced by the expansion of the repeats. Computational methods can provide an alternative to determine the structure of these polyQ proteins, but it is important to evaluate their performance before large scale prediction work is done.

[a] Reprinted with adaption from BMC Bioinformatics, 2014,15 Suppl 7: S11, doi: 10.1186/1471-2105-15-S7-S11, with permission from Biomedical Central Ltd.. Copyright: Wen et al.; licensee BioMed Central Ltd. 2014

Results

In this paper, two popular protein structure prediction programs, I-TASSER and Rosetta, have been used to predict the structure of the N-terminal fragment of a protein associated with Huntington’s disease with 17 glutamines. Results show that both programs have the ability to find the native structures, but I-TASSER performs better for the overall task.

Conclusions

Both I-TASSER and Rosetta can be used for structure prediction of proteins with polyQ repeats. Knowledge of polyQ structure may significantly contribute to development of therapeutic strategies for polyQ diseases.

Background

Knowledge of protein structure can be critical for devising therapeutic strategies for diseases in which protein dysfunction contributes to pathogenesis. For the polyglutamine (polyQ) diseases, pathogenic polyQ expansions typically cause gains of toxic functions associated with protein misfolding or aberrant interactions with RNAs or other proteins [1]. At least ten neurodegenerative disorders are caused by polyQ expansions, including Huntington’s disease (HD), dentatorubral and pallidoluysian atrophy (DRPLA), spinal and bulbar muscular atrophy (SBMA), and the polyQ spinocerebellar ataxias [2] (SCA1, SCA2, SCA3, SCA6, SCA7, SCA8, and SCA17) [3-5]. The proteins involved in these diseases have no significant sequence, compositional or structural homologies [6, 7] and numerous studies and observations have established that the length of the polyglutamine repeats plays a critical role in the progress and pathogenesis of these diseases [5, 8].

Analysis from patients’ data reveals that the expansion of polyglutamine repeats beyond certain pathological threshold causes the disease phenotype (Table 1.1) [9-12]. Also the number of the glutamines in the polyglutamine region is inversely correlated with age of onset [9, 13-17]. For instance for SCA2, people with 32 or 33 repeats tend to first experience symptoms of SCA2 in late adulthood, while people with more than 45 repeats usually have symptoms by their teens [2].

One possible mechanism for these diseases pathology is the assembly of unfolded protein monomers into β-sheet amyloid fibers [18]. Both *in vivo* and *in vitro* studies have shown that the polyQ expansion may lead to protein misfolding [19] and may cause a structure transition to form parallel β-helix and β-sheet folds [20]. Protein misfolding and aggregation has been shown to depend on the polyQ length and the concentration of the protein [21-23]. As shown in [24] the polyQ tract will form β-sheet structures when the number of the Qs increases resulting in an increase of the chance of aggregation. Therefore the understanding of the effect of the lengthening of the polyQ repeat segment on protein folding can provide new insights and perhaps therapies for these diseases.

Although the association of the lengthening of the polyQ repeats with the related polyglutamine diseases has been known for almost 20 years [25, 26], high-resolution structural analysis of these proteins in their native context has eluded researchers [27] and only very limited experimental information exists. Kim has crystallized multiple structures of the N-terminal segment of huntingtin protein with 17 and 36 glutamines repeats [28, 29], finding that the polyQ regions exhibit conformational flexibility with α-helix, random coil, and extended loops [28, 29]. These structures are the only crystal structures of polyQ segments available in the RCSB PDB database. Computational modeling can provide valuable insights into this problem [23, 30, 31], but to our knowledge no comprehensive studies have been reported comparing the 3D structures predicted for these segments with the limited experimental data available.

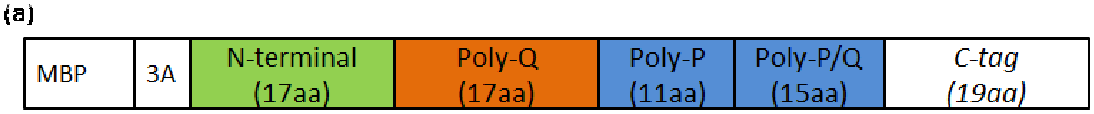
The accuracy of the structures obtained using 3D structure prediction programs is improving rapidly, and some of the commonly available programs have shown excellent performance in the CASP competition [32]. However, all the 3D structure prediction programs are trained with a variety of proteins and their performance is usually evaluated on a general dataset [33]. There is no literature evidence reporting the performance of these programs on proteins containing polyQ tracts. So it is necessary for us to evaluate the performance of these programs before we use them to predict the structure of polyglutamine disease proteins at large scale.

In this paper we present our results of the evaluation of the prediction performance of two efficient and popular 3D structure prediction programs, I-TASSER and Rosetta, on the N-terminal end of huntingtin protein with 17 glutamines (HTT17Q-EX1).

Methods

PolyQ Segments

We searched the RCSB PDB database [34] for structures with more than 10 consecutive glutamines in their sequences on November 2012. A total of 11 structures were retrieved, including 7 of the first exon of the huntingtin protein with 17 glutamines (HTT17Q-EX1) [28] and 4 of the first exon of huntingtin protein with 36 glutamines (HTT36Q-EX1) [29]. Figure 2.1(a) shows the sequence construction for the X-Ray diffraction experiment on HTT17Q-EX1 which was expressed and crystallized as a maltose-binding (MDP) fusion protein [28]. The same methods were used to get the crystal structure of HTT36Q-EX1,



(b)

>HTT17Q-EX1
MATLEKLMKAFESLKSFQQQQQQQQQQQQQQQQQPPPPPPPPPPPQLPQPPPQAQPLLPQ

Figure 2.1 The sequence construction of HTT17Q-EX1. (a) sequence structure of the PDB records; (b) sequence used for structure prediction.

but the resolution of the HTT36Q-EX1 is of such poor quality that only HTT17Q-EX1 structures were used in this study.

PDB identification numbers of the 7 HTT17Q-EX1 crystal structures used here are 3IO4 [PDB: 3IO4], 3IO6 [PDB: 3IO6], 3IOT [PDB: 3IOT], 3IOU [PDB: 3IOU], 3IOR [PDB: 3IOR], 3IOV [PDB: 3IOV] and 3IOW [PDB: 3IOW]. Each crystal includes a trimer of MDP-HTT17Q-EX1, so a total of 21 structures of HTT17Q-EX1 were considered. Figure 2.1 (b) shows the sequence of the HTT17Q-EX1 used as the input of the 3D structure prediction.

Protein 3D Structure Prediction

Two protein structure prediction programs were used in this study, I-TASSER and Rosetta. Both I-TASSER and Rosetta have been used by thousands users and they are among the few programs which can handle large proteins with more than 1000 residues [35, 36].

I-TASSER is the 3D structure prediction program based on multiple-threading alignments and iterative template fragment assembly simulations [37]. I-TASSER is a fully automated method and was used without further modifications, but we have verified that none of the templates corresponding to the structures 3IO4 [PDB: 3IO4], 3IO6 [PDB: 3IO6], 3IOT [PDB: 3IOT], 3IOU [PDB: 3IOU], 3IOR [PDB: 3IOR], 3IOV [PDB: 3IOV] and 3IOW [PDB: 3IOW] was included in the knowledge data used in the version of I-TASSER used here. Rosetta is a flexible software suite for macromolecular modeling, which includes tools for structure prediction and design [38]. Rosetta *ab initio* module was used in this study. For Rosetta, the quota protocol fragment picking was used to generate 3-mers and 9-mers fragments, which took into account the secondary structure predictions by PsiPred [39], Jufo9D Server [40] and SAM-T08 [41] as the quota pools. The weight given to the each quota pool was assigned following reference [42] and 200 fragments were picked from the total of 700 candidates available from both 3-mers and 9-mers fragments. The default parameters were used for Rosetta *ab initio* modelling with the number of output structures set as 5000, the default parameters also were used for Rosetta cluster module.

We installed I-TASSER Version 2.1 and Rosetta Version 3.4 in a cluster at the Center for High Performance Computing (CHPC) of University of Utah, where all computations were performed. As a fully automated program, the number of decoys to screen and the number of simulation jobs in I-TASSER are fixed, whereas Rosetta is much more flexible and users can define the output number of structures and the number of parallel simulation jobs, making it much more adaptable to the hardware architecture used. So it is difficult to compare the computational cost of the two programs. However, for the modelling tasks with the parameters used in our simulation, the total CPU time for I-TASSER to finish one HTT17Q-EX1 (60 amino acid residues) prediction was, in average, 24.58 hours using one core in a 2.4 GHz dual-core Opteron processor, whereas the average total CPU time for Rosetta to finish one HTT17Q-EX1 prediction with 5000 prediction structures was about 50.91 hours in the same computing environment.

3D Structure Alignment

To assess 3D structure similarity, TM-align was used for structure comparison and alignment [43]. The TM-score calculated by TM-align, which lies in (0,1] interval, is considered a good measure of the similarity of two structures [44]. A TM-score of less than 0.17 indicates a random alignment, whereas TM-score greater than 0.5 indicates that the two structures are generally in the same fold [44].

Similarity Measurement

Besides the TM-score, exact structure overlap (ESO) and exact structure overlap of polyQs (ESOP) were also used to measure the similarity of two structures. The words 'exact' here means the aligned residues are within certain threshold, 5Å in this study, and that they are the same residue in the HTT17Q-EX1 sequence. For example, if a serine (SER) in the 16th position of the predicted structure of HTT17Q-EX1is aligned, within the distance threshold, with the serine (SER) in the 16th position of PDB experimental structure, the 16SER-16SER is an exact match. ESO and ESOP is derived from the Structure Overlap (SO) which is a standardized score to compare the structure alignments and measure the local similarity of two structures [45]. The SO score is calculated as:

$$SO = 100 \times \frac{L(A)}{min(Lm, Le)} \tag{2.1}$$

where *L(A)* is the structure alignment length; the *Lm* and *Le* are the length of the predicted model and the experimental structure, respectively.

We have modified Equation (1) to meet the aim of more strict structure comparison, and get the ESO score:

$$ESO = 100 \times \frac{L(EA)}{min(Lm, Le)} \tag{2.2}$$

where *L(EA)* is the length of exact match; *Lm* and *Le* is the length of predicted model and the length of the PDB experimental structure respectively.

The structure of polyQ region may play a more important role than other positions. In this study, the ESOP score is calculated to evaluate the structure similarity of the polyQ regions. The ESOP is a special version of ESO, and it is calculated as:

$$ESOP = 100 \times \frac{L(EAQ)}{min(LQm,LQe)} \qquad (2.3)$$

where *L(EAQ)* is the length of the exact match of Qs; *LQm* and *LQe* are the length of polyQ in predicted model and PDB experimental structure respectively.

Secondary Structure Calculation

The secondary structure of the predicted models and the PDB experimental structures were calculated using the DSSP algorithm, which is an algorithm to standardize secondary structure assignment [46]. Secondary structures assigned by DSSP are 8 conformational states, including α-helix, β-bridge, strand, 3-helix, 5-helix, turn, bend, and random coil.

The results of DSSP are the secondary structures represented by one letter for each position. In order to get a better view of the results, "WebLogo 3" [47] was used to plot the secondary structure logo at each position. The overall height of the stack indicates the secondary structure conservation at that position, and the height of the symbols within the stack indicates the relative frequency of each secondary structure type at that position.

3D Structure Visualization

The 3D structure and the 3D structure superposition were visualized in the UCSF Chimera software, a free program for molecular graphics and analysis [48].

Statistic

To depict the data distribution of the parameters calculated here, the (mean value ± standard deviation) is listed for data with normal distribution, whereas for data that do not follow the normal distribution, the 25% quantile and 75% quantile values are listed.

The Student t test was applied for data with normal distribution and the Wilcoxon ranked test was performed on other data sets to assess significance. The significant level was set at 0.05. All the statistic work was done in the R environment which is a free software environment for statistical computing and graphics [49].

Results

Predicted Models

As evidence shows that the polyQ region can adopt different structures [28, 29] in the proteins of interest for polyQ diseases, it is not appropriate to seek the 'best structure' of this region, but it is more appropriate to look for ensembles of structures (generated by multiple independent runs) which can show overall trends and represent the variety of structures observed by experimental methods.

Following this reasoning, both Rosetta and I-TASSER were run 10 times using different random seeds for each run of 3D structure prediction of the HTT17Q-EX1 sequence shown in Figure 2.1(b). For each run we kept the five best models, so a total of 50 I-TASSER models and 50 Rosetta models were retained for analysis.

Each structure prediction program will return some parameters to estimate the accuracy of the models. For I-TASSER, the C-score, which lies in the (-5,2) range, is calculated for each model [35]. The C-scores of the best 50 I-TASSER models, listed in Table 2.1, range from -2.62 to -4.72.

The clustering algorithm from Rosetta was used to identify the most frequently sampled conformations. For each run we selected the five structures with the lowest energy from the structures encountered in the five different clusters in which the number of structures was greater than 10 on each. The energies of the total 50 Rosetta structures, listed in Table 2.2, they range from 16.06 to 20.13.

Table 2.1 C-scores for the best I-TASSER models

Model	# 1	# 2	# 3	# 4	# 5
Run 1	-2.91	-3.69	-3.33	-3.62	-4.72
Run 2	-2.84	-3.71	-3.31	-3.5	-4.42
Run 3	-2.81	-3.76	-3.48	-3.89	-3.74
Run 4	-2.62	-3.69	-3.32	-3.76	-3.49
Run 5	-3.02	-3.21	-3.91	-4.11	-3.42
Run 6	-2.67	-3.76	-3.48	-3.62	-4.37
Run 7	-2.77	-3.42	-3.96	-3.3	-3.51
Run 8	-3.09	-3.45	-4.09	-4.22	-4.27
Run 9	-2.73	-3.61	-3.38	-3.76	-4.42
Run 10	-2.62	-3.49	-3.75	-4.33	-4.01

Table 2.2 Energy for the best Rosetta models

Model	# 1	# 2	# 3	# 4	# 5
Run 1	16.061	16.349	18.609	19.656	19.956
Run 2	17.881	18.309	18.373	18.386	19.215
Run 3	16.943	17.598	17.639	18.306	19.436
Run 4	18.414	18.662	18.691	18.812	19.076
Run 5	16.74	18.004	18.192	18.3	19.015
Run 6	18.353	18.388	18.572	18.766	18.96
Run 7	17.435	18.897	19.571	19.603	19.617
Run 8	18.128	19.111	19.521	19.643	19.707
Run 9	17.317	17.586	17.655	17.916	18.69
Run 10	19.329	19.899	19.928	20.104	20.13

Secondary Structure

For better visualization, WebLogo [47] was used to display secondary structure patterns. The WebLogo of the secondary structures of the experimental PDB structures and the best I-TASSER and Rosetta models are shown in Figure 2.2. For easy description, we divided the sequence into three regions: the 17-residue head region including residues 1 to 17; the polyQ region including residues 18 to 34 and C-terminal region including residues 35 to 60. As discussed in the original publication for the 21 PDB structures most crystals show α-helix in the head region, which is always well resolved, with only a few structures showing turns at the beginning and end of the head region. Both the I-TASSER and Rosetta best models reproduce the observed trends showing a majority of helix structures in the head region, but the I-TASSER structures show better agreement with the experimental findings showing a preference for α-helix, while the Rosetta structures show a mix of α-helix and 3-helix. The secondary structure, for the resolved structures, in the polyQ region is more diverse showing a number of structures with α-helix, random coils, and turns. The Pro-enriched C-terminal region is dominated, at least for the resolved structures in this region, by coil structures. Unfortunately, as depicted in Figure 2.2(a), the number of well resolved structures rapidly decreases beyond the head region making comparison with the experiments less reliable. None-the-less the overall experimental trends are reproduced by both I-TASSER and Rosetta, but it appears that the I-TASSER structures show more loops than the experimental data.

Overall I-TASSER appears to be superior reproducing quite well the stable α-helix structure of the N-terminal regions and showing increased diversity of structures in the polyQ region and a predominance of coil structures in the C-terminal region.

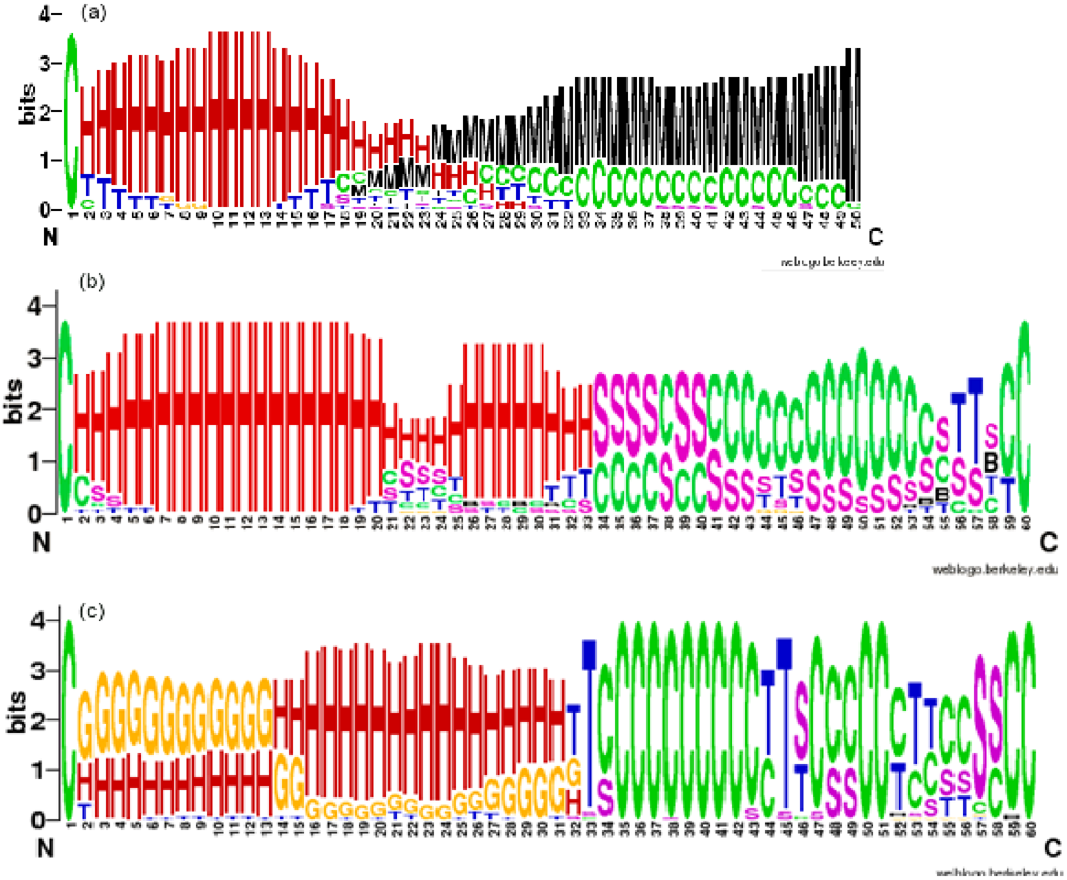


Figure 2.2 Secondary structure WebLogo. (a) PDB structures; (b) I-TASSER models; (c) Rosetta models. In (a) M represents the number of structures with missing values due to lack of resolution in the experimental data. The codes for secondary structure are as follows: H: α-helix; B: β-bridge; E: Strand; G: Helix-3; I: Helix-5; T: Turn; S: Bend; C: Coil; M: Missing data.

Reproducibility of I-TASSER and Rosetta Results

In order to test the sensitivity of I-TASSER and Rosetta with the selection of the seeds used in the calculations, we have calculated the structure similarity using the TM-score between models obtained using the same prediction program. A total of 1225 TM-scores were generated comparing pairwise the best 50 I-TASSER and 50 Rosetta models, respectively.

TM-scores between any two models from I-TASSER range from 0.2781 to 0.7163, with an average of 0.4086 and a standard deviation 0.0692. Whereas the TM-scores between any two Rosetta models range from 0.2865 to 0.8236, with an average of 0.4979 and a standard deviation 0.0892. The difference between TM-scores of I-TASSER and Rosetta is statistically significant (t-test, $p<0.001$, Figure 2.3). The number of TM-scores greater than 0.5 is two times greater for Rosetta/Rosetta pairs than for I-TASSER/I-TASSER pairs, i.e., 561 pairs in Rosetta and 126 pairs in I-TASSER have scores larger than 0.5.

When comparing only the best models of each run, the TM-scores range from 0.4539 to 0.6813 for I-TASSER (Table 2.3) and from 0.2872 to 0.6879 for Rosetta (Table 2.4). Therefore the best models of each run from I-TASSER are more similar among themselves than those from Rosetta, i.e., 33 pairs of the 45 structure pairs have TM-scores greater than 0.5 for I-TASSER, whereas for Rosetta, only 18 pairs of best models have TM-scores greater than 0.5.

The sensitivity to the selected random seeds was also evaluated at the run level. TM-scores were calculated for the structures of any 5 models in one run compared with any 5 models of other runs. The number of pairs with TM-score greater than 0.5 between

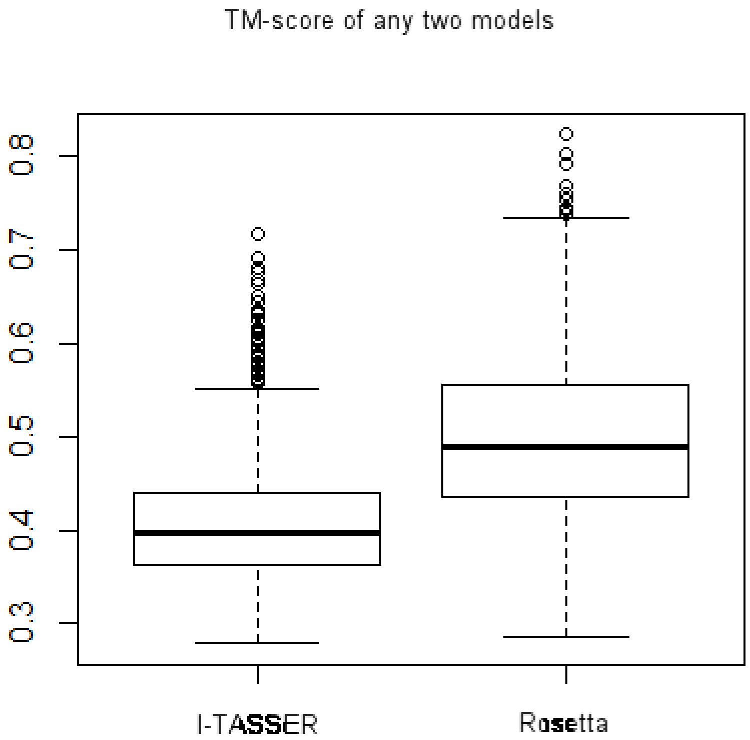


Figure 2.3 Distribution of the TM-scores of any two models from I-TASSER and Rosetta respectively.

Table 2.3 TM-scores between the best models from I-TASSER

run	# 2	# 3	# 4	# 5	# 6	# 7	# 8	# 9	# 10
# 1	0.48	0.48	0.63	0.48	0.54	0.57	0.49	0.51	0.46
# 2		0.56	0.54	0.55	0.48	0.54	0.45	0.58	0.46
# 3			0.54	0.63	0.53	0.51	0.54	0.56	0.51
# 4				0.50	0.52	0.66	0.49	0.56	0.49
# 5					0.60	0.51	0.58	0.59	0.54
# 6						0.51	0.68	0.54	0.49
# 7							0.50	0.57	0.45
# 8								0.50	0.52
# 9									0.50

Table 2.4 TM-scores between the best models from Rosetta

run	# 2	# 3	# 4	# 5	# 6	# 7	# 8	# 9	# 10
# 1	0.30	0.43	0.37	0.42	0.42	0.53	0.40	0.38	0.43
# 2		0.30	0.34	0.31	0.39	0.28	0.34	0.32	0.38
# 3			0.52	0.53	0.53	0.65	0.47	0.57	0.50
# 4				0.48	0.53	0.47	0.68	0.49	0.43
# 5					0.63	0.52	0.48	0.58	0.61
# 6						0.52	0.47	0.64	0.59
# 7							0.41	0.56	0.55
# 8								0.46	0.42
# 9									0.60

any two experiments is shown in Table 2.5 for I-TASSER and Table 2.6 for Rosetta. For I-TASSER, the number of pairs with TM-score greater than 0.5 ranges from 0 to 6. There are 6 pairs with TM-scores greater than 0.5 between Run 4 and Run 7, however, no pairs with TM-scores greater than 0.5 between Run 1 and Run 8. For Rosetta, the number of pairs with TM-score greater than 0.5 at run level ranges from 5 to 20. 20 of 25 pairs are with TM-scores greater than 0.5 between Run 3 and Run 7, which is the best. The smallest number of pairs for Rosetta is 5, which shows in 3 pairs, Run 1 and Run 6, Run 6 and Run 8, Run 5 and Run 8.

These results show that our ensemble approach to predict the structure of proteins associated with polyQ diseases appears to be appropriate. Using multiple seeds it is possible to obtain an ensemble of structures that show reasonable diversity, but still retain the main features. We believe that this approach is quite promising because it can incorporate in future analysis the diverse structure of which appears to be an emerging observation from the limited experimental data on these proteins.

Validity Evaluation of I-TASSER and Rosetta

As depicted in Figure 2.2(a) not all of the 21 PDB structures have been resolved in the polyQ region, which is our main interest. For instance, the longest well resolved polyQ region is the B chain of the 3IOW [PDB: 3IOW] structure in which all the 17 Qs structures are resolved, whereas for the A chain of the 3IOT [PDB: 3IOT] structure only one Q has been resolved. Also, there are numerous gaps in several structures as some of the residues are not resolved. Taking this into account and in order to make an accurate comparison with the experimental ones in the region of interest, only PDB structures in which at least 9 (more than half the total number) of consecutive Qs in the polyQ regions show well

Table 2.5 Number of pairs with TM-score greater than 0.5 between any two runs of I-TASSER

run	# 2	# 3	# 4	# 5	# 6	# 7	# 8	# 9	# 10
# 1	2	3	3	1	3	3	0	5	1
# 2		2	2	2	2	3	2	2	1
# 3			3	3	2	3	2	1	4
# 4				3	2	6	3	5	1
# 5					3	5	3	2	4
# 6						4	1	5	2
# 7							4	5	2
# 8								3	2
# 9									2

Table 2.6 Number of pairs with TM-score greater than 0.5 between any two runs of Rosetta

run	# 1	# 2	# 3	# 4	# 5	# 6	# 7	# 8	# 9	# 10
# 1	4	6	10	11	8	5	11	12	8	9
# 2		2	11	8	7	6	13	11	11	10
# 3			6	15	9	13	20	16	18	16
# 4				3	9	7	14	13	16	14
# 5					4	8	15	5	16	15
# 6						1	8	5	13	9
# 7							5	15	16	14
# 8								8	7	13
# 9									9	17
#10										6

resolved structures were used for the evaluation of the results produced with I-TASSER and Rosetta. There are ten PDB structures that meet this criteria: the B chain of 3IO4 [PDB: 3IO4] (3io4_b), the C chain of 3IO4 [PDB: 3IO4] (3io4_c), the B chain of 3IO6 [PDB: 3IO6] (3io6_b), the C chain of 3IO6 [PDB: 3IO6] (3io6_c), the C chain of 3IOR [PDB: 3IOR] (3ior_c), the B chain of 3IOT [PDB: 3IOT] (3iot_b), the C chain of 3IOU [PDB: 3IOU] (3iou_c), the B chain of 3IOV [PDB: 3IOV] (3iov_b), the C chain of 3IOV [PDB: 3IOV] (3iov_c), and the B chain of 3IOW [PDB: 3IOW] (3iow_b). The number of consecutive Qs in each structure is shown in Table 2.7.

The best 50 I-TASSER and 50 Rosetta models were compared with these 10 PDB structures using the TM-align program. TM-scores, root-mean-square deviation (RMSD), aligned number of residues, sequence identity and the structure superposition were obtained from TM-align [43]; the number of exact matches and the number of exact matched Qs were extracted from the structure alignment and finally the exact structure overlap (ESO) and exact structure overlap of Qs (ESOP) were calculated using Equation (2.2) and Equation (2.3) given in the methods section. The values of each similarity parameter considered here are shown in Table 2.8 along with the p-values assessing the significance of the difference between the I-TASSER and Rosetta results.

The average TM-score of I-TASSER/PDB superposition pairs is 0.50 and the average TM-score of Rosetta/PDB pairs is 0.45, reflecting the fact that 253, of the 500, I-TASSER/PDB pairs have TM-scores greater than 0.5 while only 87 pairs of the Rosetta/PDB pairs have TM-scores greater than 0.5. The average RMSD of I-TASSER/PDB pairs (1.53 Å) is also smaller than that of Rosetta/PDB pairs (1.74 Å). Other TM-align parameters depicted in Table 2.8 also show that I-TASSER performs better

Table 2.7 Numbers of Qs in the PDB structures

PDB structure	Number of Qs
3io4_b	10
3io4_c	11
3io6_b	14
3io6_c	10
3ior_c	13
3iot_b	12
3iou_c	14
3iov_b	11
3iov_c	15
3iow_b	17

Table 2.8 Distribution of structure superposition parameters between predicted models and PDB structures

	I-TASSER	Rosetta	p-value
TM-score	0.50±0.06	0.45±0.06	<0.0001
RMSD (Å)	1.53±0.34	1.74±0.34	<0.0001
Aligned number	24.05±2.14	25.56±2.41	<0.0001
Sequence Identity [a]	(0.30,0.71)	(0.38,0.52)	<0.0001
Exact Match (<5.0 Å) [a]	(0,16)	(0,0)	<0.0001
Exact Qs Match(<5.0 Å) [a]	(0,1)	(0,0)	<0.0001
Total Qs Match(<5.0 Å) [a]	(5,8)	(8,11)	<0.0001
Exact Match (other) [a]	(0,0)	(0,0)	<0.0001
Exact Qs Match(other) [a]	(0,0)	(0,0)	<0.0001
Total Qs Match(other) [a]	(0,1)	(0,1)	<0.0001
Exact Match (all) [a]	(6,25)	(0,0)	<0.0001
Exact Qs Match (all) [a]	(0,1)	(0,0)	<0.0001
Total Qs Match (all) [a]	(6,8)	(9,12)	<0.0001
ESOP [a]	(0,9.09)	(0,0)	<0.0001
ESO [a]	(0, 53.13)	(0,0)	<0.0001

[a] The values between brackets represent the value of the property for the best structure superposition at the first and third quartile, respectively, of their distributions.

than Rosetta in this test.

The structure overlap scores, ESOP and ESO, for I-TASSER models are also better than those for Rosetta models. For instance more than 75% of the Rosetta models have no exact match in the polyQ region nor for the entire sequence, whereas the 75% quantile of the ESO and ESOP scores for I-TASSER are 53.13 and 9.09, respectively. The statistical tests have shown that these differences are significant (Table 2.8).

Fifty of the I-TASSER/PDB structure superpositions have ESOP values greater than or equal to 50, which means that 50 pairs have more than 50% of Qs in the polyQ region with exact match. These 50 pairs include 9 of the 10 PDB structures, so 9 of the 10 structures have corresponding I-TASSER models with very good matches in the polyQ regions. In contrast only 5 of these 10 structures have corresponding Rosetta/PDB structure superposition matches when the same criteria are used.

The best matches between the predicted structures by I-TASSER and Rosetta, respectively, and one of the PDB structures considered here are depicted in Figure 2.4. The I-TASSER structure best match is with the B chain of 3IO6 [PDB: 3IO6]; the match has a TM-score of 0.56 and the ESOP score of 100. The best two matches for Rosetta structures show matches with the C chain of 3IOU [PDB: 3IOU] and the B chain of 3IOW [PDB: 3IOW]. Their TM-scores are 0.5074 and 0.5057, respectively, and the ESOP score of 100.

Discussion

This study evaluated two software tools for predicting, from amino acid sequences, the 3D structures of the polyQ regions of proteins related to polyglutamine diseases. Pathogenic neurodegenerative polyQ proteins were used as a model, for relevance to developing structure-specific therapeutics based on normal vs. polyQ expanded protein

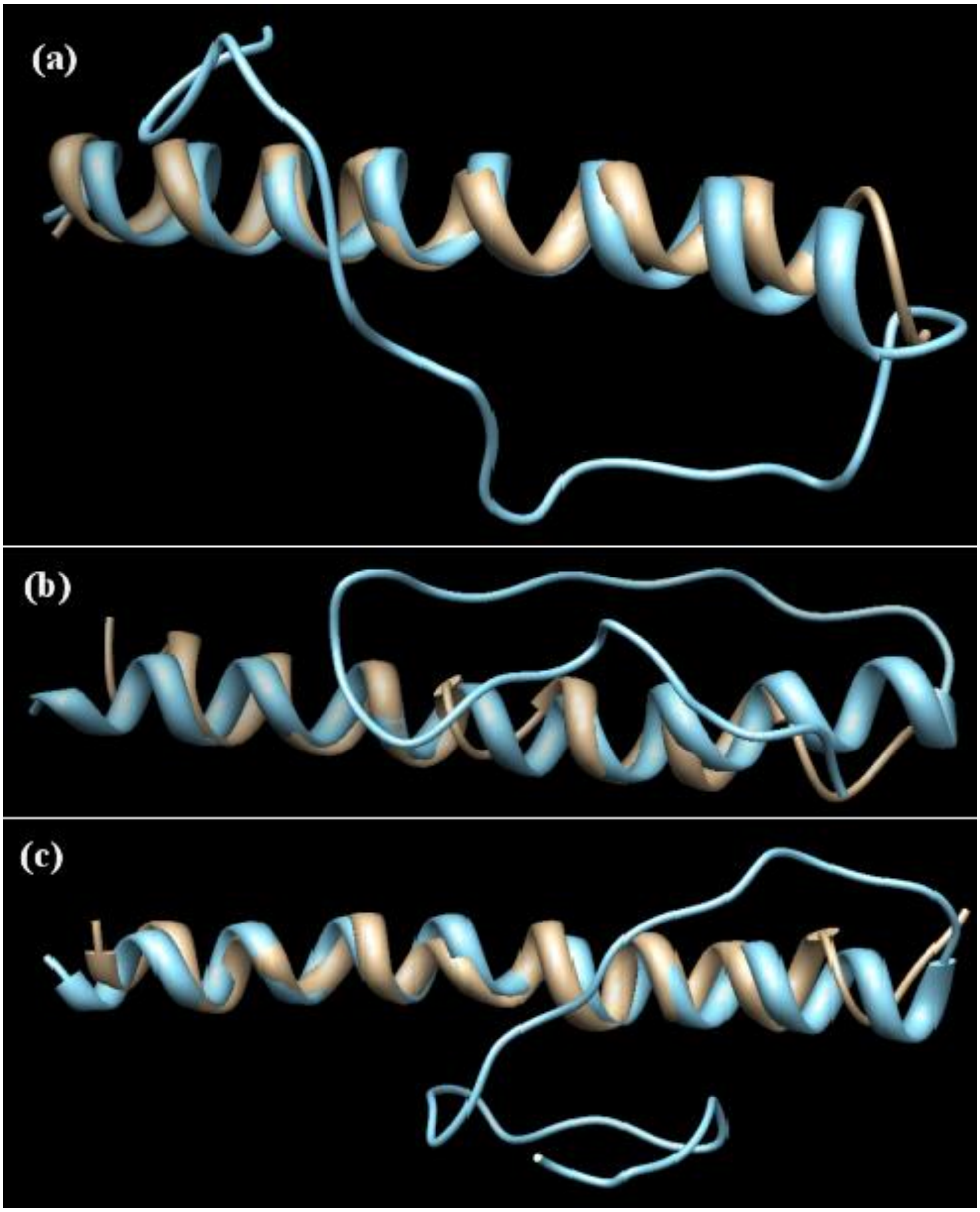


Figure 2.4 Structure superposition of predicted models and PDB structures. Structure superposition of predicted models and PDB structures with TM-score>0.5 and ESOP=100. (a) I-TASSER third model in the tenth run with 3io6_b; (b) Rosetta forth model in the first run with 3iou_c; (c) Rosetta third model in the fifth run with 3iow_b. Tan: PDB structure; Sky blue: predicted models. The N-terminal end of each structure is shown on the left.

structures. Two highly recognized and efficient 3D structure prediction programs, I-TASSER and Rosetta, were evaluated to assess their performance for structure prediction using segments of the huntingtin protein harboring polyQ repeats. Both I-TASSER and Rosetta produced good results.

When tested for structure stability under changes of the initial random seed, Rosetta shows less variability than I-TASSER. This means that if we run Rosetta and I-TASSER several times respectively, it is possible that we will get less variance in the results from Rosetta than from I-TASSER. Nonetheless, both programs produce a reasonable ensemble of structures with sufficient diversity and without extreme deviations. Several studies have illustrated that the polyQ repeat regions of these proteins are highly disordered with structure flexibility [31], but this has not been quantified experimentally. Therefore it is challenging to discriminate among these two approaches using these criteria. In consequence we must conclude that both I-TASSER and Rosetta are suitable for the task on predicting ensemble structures of protein containing polyQ segments.

The accuracy of the prediction program is a very important factor that we evaluated here. In this study, the structure similarity between the predicted models and the PDB experiment structures available was used to evaluate the validity of the prediction programs. The root-mean-square deviation (RMSD) score is the most often-used parameter to calculate the structure similarity, but a drawback of its use is that a relatively small local variation can result in a high RMSD [44]. TM-score weights the close atom pairs stronger than the distant matches, and it is more sensitive to the topology fold than the RMSD [44]. Besides the global similarity measured by TM-score, more restricted scores on the exact match of two structures were also calculated. The exact structure overlap (ESO), derived from the structure overlap (SO) score [45], was introduced and instead of calculating the number of aligned pairs it counts the exact match pairs, which not only counts aligned residues but also residues that lie in the same positions in both the sequences of predicted model and PDB structure. The exact structure overlap of polyQ repeat (ESOP) is the special version of ESO, which is used to measure the prediction accuracy in the polyQ region. Considering the TM-score, ESO and ESOP together gives a more comprehensive view of similarity between the predicted model and the PDB experimental structures from both a global and a local aspect. The ESO score and ESOP score can be used for similarity comparison tasks, especially if there are regions which play more critical roles than others.

Rosetta models have a larger number of aligned residues on average than I-TASSER, but the average RMSD values and TM-scores are much higher (lower) than that of I-TASSER. So when the Rosetta models are aligned with the PDB structures, the distance between the models and the experimental structures is large, which is not a good sign for good structural matches. On the contrary, I-TASSER models aligned better with PDB structures not only with better RMSD and TM-scores, but also better ESO and ESOP scores. This can also be seen from the secondary structures patterns. When considering specific structure pairs, both I-TASSER and Rosetta have predicted models which can match the PDB structures with good global (TM>0.5) and local (ESO>=50 and ESOP>=50) structures. So both Rosetta and I-TASSER have the ability to get the native models, but for the overall performance, I-TASSER appears to be better than Rosetta.

As several models are returned by the structure prediction programs, it is important to have criteria to select the best models. However, the model with the lowest energy in the prediction program may not be the best model for reproducing the polyQ regions. For instance for Rosetta, the two predicted models with TM-score greater than 0.5 and ESOP of 100 (Figure 2.4(b) and 4(c)) are not the models with the lowest energy in that Rosetta run. This is true also for the I-TASSER model with TM-score greater than 0.5 and ESOP of 100 (Figure 2.4(a)). In fact, of the 29 good models which have TM-score greater than 0.5 and ESOP score greater than 50, only one model is ranked as the best by I-TASSER.

Conclusions

Both I-TASSER and Rosetta can be used for *in silico* studies of the structures of proteins with polyQ repeats related to neurodegenerative diseases. However, I-TASSER shows better performance than Rosetta when considering the overall agreement between results produced using these two prediction models with the limited experimental results available for comparison.

Both I-TASSER and Rosetta are computationally efficient as both applications can be easily parallelized by executing numerous jobs each with a unique random seed.

In our future studies we will attempt to predict the change of the structure as function of the number of Qs in the polyQ repeat segment for all the proteins involved in polyQ neurological diseases. Ideally we could use both these two programs to predict structures of the polyQ disease related proteins. This could provide a quasi “crowdsourcing” mechanism to cross check the results, but may prove computationally too expensive (see Methods). Therefore the results presented here suggest that studies should be, at least initially, performed using I-TASSER.

Authors' Contributions

JCF and JW designed the study; JW did the research work; JW, DRS and JCF discussed the results within the clinical and biochemical framework. All authors read and approved the final manuscript.

Acknowledgements

Molecular graphics were performed with the UCSF Chimera package. Chimera is developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco (supported by NIGMS P41-GM103311). Computer time was provided by an allocation to JCF at the CHPC of the University of Utah. JCF has been partially supported by the University of Utah Center for Clinical and Translational Science under NCATS Grants UL1RR025764 and 3UL1RR025764-02S2.

References

1. Wetzel R. Physical chemistry of polyglutamine: intriguing tales of a monotonous sequence. J Mol Biol. 2012;421(4-5):466-90.

2. Matilla-Duenas A, Corral-Juan M, Volpini V, Sanchez I. The spinocerebellar ataxias: clinical aspects and molecular genetics. Adv Exp Med Biol. 2012;724:351-74.

3. Zoghbi HY, Orr HT. Glutamine repeats and neurodegeneration. Annu Rev Neurosci. 2000;23:217-47.

4. Moseley ML, Zu T, Ikeda Y, Gao W, Mosemiller AK, Daughters RS et al. Bidirectional expression of CUG and CAG expansion transcripts and intranuclear polyglutamine inclusions in spinocerebellar ataxia type 8. Nat Genet. 2006;38(7):758-69.

5. Michalik A, Van Broeckhoven C. Pathogenesis of polyglutamine disorders: aggregation revisited. Hum Mol Gen. 2003;12 Spec No 2:R173-86.

6. Albrecht M, Golatta M, Wullner U, Lengauer T. Structural and functional analysis of ataxin-2 and ataxin-3. Eur J Biochem. 2004;271(15):3155-70.

7. Williams AJ, Paulson HL. Polyglutamine neurodegeneration: protein misfolding revisited. Trends Neurosci. 2008;31(10):521-8.

8. Magana JJ, Velazquez-Perez L, Cisneros B. Spinocerebellar ataxia type 2: clinical presentation, molecular mechanisms, and therapeutic perspectives. Mol Neurobiol. 2013;47(1):90-104.

9. Walters RH, Murphy RM. Examining polyglutamine peptide length: a connection between collapsed conformations and increased aggregation. J Mol Biol. 2009;393(4):978-92.

10. Garden GA, La Spada AR. Molecular pathogenesis and cellular pathology of spinocerebellar ataxia type 7 neurodegeneration. Cerebellum. 2008;7(2):138-49.

11. Costa Mdo C, Paulson HL. Toward understanding Machado-Joseph disease. Prog Neurobiol. 2012;97(2):239-57.

12. Imarisio S, Carmichael J, Korolchuk V, Chen CW, Saiki S, Rose C et al. Huntington's disease: from pathology and genetics to potential therapies. Biochem J. 2008;412(2):191-209.

13. Pulst SM, Nechiporuk A, Nechiporuk T, Gispert S, Chen XN, Lopes-Cendes I et al. Moderate expansion of a normally biallelic trinucleotide repeat in spinocerebellar ataxia type 2. Nat Genet. 1996;14(3):269-76.

14. Zoghbi HY, Jodice C, Sandkuijl LA, Kwiatkowski TJ, Jr., McCall AE, Huntoon SA et al. The gene for autosomal dominant spinocerebellar ataxia (SCA1) maps telomeric to the HLA complex and is closely linked to the D6S89 locus in three large kindreds. Am J Hum Genet. 1991;49(1):23-30.

15. Ikeda Y, Dalton JC, Day JW, Ranum LPW. Spinocerebellar ataxia type 8. In: Pagon RA, Adam MP, Bird TD, Dolan CR, Fong CT, Stephens K, editors. GeneReviews. Seattle WA: University of Washington. 2001. https://www.ncbi.nlm.nih.gov/books/NBK1268/. Accessed 1 Nov 2016.

16. Nozaki K, Onodera O, Takano H, Tsuji S. Amino acid sequences flanking polyglutamine stretches influence their potential for aggregate formation. Neuroreport. 2001;12(15):3357-64.

17. Pulst SM, Santos N, Wang D, Yang H, Huynh D, Velazquez L et al. Spinocerebellar ataxia type 2: polyQ repeat variation in the CACNA1A calcium channel modifies age of onset. Brain. 2005;128(Pt 10):2297-303.

18. Finke JM, Cheung MS, Onuchic JN. A structural model of polyglutamine determined from a host-guest method combining experiments and landscape theory. Biophys J. 2004;87(3):1900-18.

19. Li X, Li H, Li X-J. Intracellular degradation of misfolded proteins in polyglutamine neurodegenerative diseases. Brain Res Rev. 2008;59(1):245-52.

20. Cote S, Wei G, Mousseau N. All-atom stability and oligomerization simulations of

polyglutamine nanotubes with and without the 17-amino-acid N-terminal fragment of the Huntingtin protein. J Phys Chem B. 2012;116(40):12168-79.

21. Kubota H, Kitamura A, Nagata K. Analyzing the aggregation of polyglutamine-expansion proteins and its modulation by molecular chaperones. Methods. 2011;53(3):267-74.

22. Perney NM, Braddick L, Jurna M, Garbacik ET, Offerhaus HL, Serpell LC et al. Polyglutamine aggregate structure in vitro and in vivo; new avenues for coherent anti-Stokes Raman scattering microscopy. PLoS One. 2012;7(7):e40536.

23. Wang Y, Voth GA. Molecular dynamics simulations of polyglutamine aggregation using solvent-free multiscale coarse-grained models. J Phys Chem B. 2010;114(26):8735-43.

24. Lakhani VV, Ding F, Dokholyan NV. Polyglutamine induced misfolding of huntingtin exon1 is modulated by the flanking sequences. PLoS Comput Biol. 2010;6(4):e1000772.

25. Kawaguchi Y, Okamoto T, Taniwaki M, Aizawa M, Inoue M, Katayama S et al. CAG expansions in a novel gene for Machado-Joseph disease at chromosome 14q32.1. Nat Genet. 1994;8(3):221-8.

26. Imbert G, Saudou F, Yvert G, Devys D, Trottier Y, Garnier JM et al. Cloning of the gene for spinocerebellar ataxia 2 reveals a locus with high sensitivity to expanded CAG/glutamine repeats. Nat Genet. 1996;14(3):285-91.

27. Miller J, Rutenber E, Muchowski PJ. Polyglutamine dances the conformational cha-cha-cha. Structure. 2009;17(9):1151-3.

28. Kim MW, Chelliah Y, Kim SW, Otwinowski Z, Bezprozvanny I. Secondary structure of Huntingtin amino-terminal region. Structure. 2009;17(9):1205-12.

29. Kim M. Beta conformation of polyglutamine track revealed by a crystal structure of Huntingtin N-terminal region with insertion of three histidine residues. Prion. 2013;7(3):221-8.

30. Esposito L, Paladino A, Pedone C, Vitagliano L. Insights into structure, stability, and toxicity of monomeric and aggregated polyglutamine models from molecular dynamics simulations. Biophys J. 2008;94(10):4031-40.

31. Miettinen MS, Knecht V, Monticelli L, Ignatova Z. Assessing polyglutamine conformation in the nucleating event by molecular dynamics simulations. J Phys Chem B. 2012;116(34):10259-65.

32. Runthala A. Protein structure prediction: challenging targets for CASP10. J Biomol Struct Dyn. 2012;30(5):607-15.

33. Moult J, Fidelis K, Kryshtafovych A, Rost B, Tramontano A. Critical assessment of methods of protein structure prediction - round VIII. Proteins. 2009;77 Suppl 9:1-4.

34. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H et al. The Protein Data Bank. Nucleic Acids Res. 2000;28(1):235-42.

35. Zhang Y. I-TASSER server for protein 3D structure prediction. BMC Bioinformatics. 2008;9:40.

36. Raman S, Vernon R, Thompson J, Tyka M, Sadreyev R, Pei J et al. Structure prediction for CASP8 with all-atom refinement using Rosetta. Proteins. 2009;77 Suppl 9:89-99.

37. Roy A, Kucukural A, Zhang Y. I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protocols. 2010;5(4):725-38.

38. Leaver-Fay A, Tyka M, Lewis SM, Lange OF, Thompson J, Jacak R et al. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. Methods Enzymol. 2011;487:545-74.

39. The PSIPRED protein sequence analysis workbench. The PSIPRED Team, London. 2012. http://bioinf.cs.ucl.ac.uk/psipred/. Accessed 1 Dec 2012.

40. Jufo9D Server. Meiler Lab, Nashville. 2012. http://www.meilerlab.org/index.php/servers/show?s_id=5. Accessed 1 Dec 2012.

41. Karplus K. SAM-T08, HMM-based protein structure prediction. Nucleic Acids Res. 2009;37(Web Server issue):W492-7.

42. Gront D, Kulp DW, Vernon RM, Strauss CE, Baker D. Generalized fragment picking in Rosetta: design, protocols and applications. PLoS One. 2011;6(8):e23294.

43. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. 2005;33(7):2302-9.

44. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. Proteins. 2004;57(4):702-10.

45. Slater AW, Castellanos JI, Sippl MJ, Melo F. Towards the development of standardized methods for comparison, ranking and evaluation of structure alignments. Bioinformatics. 2013;29(1):47-53.

46. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983;22(12):2577-637.

47. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. Genome Res. 2004;14(6):1188-90.

48. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC et al. UCSF Chimera--a visualization system for exploratory research and analysis. J Comput Chem. 2004;25(13):1605-12.

49. R Development Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing. 2011. http://www.R-project.org/. Accessed 1 Nov 2016.

CHAPTER 3

EFFECTS OF THE ENLARGEMENT OF POLYGLUTAMINE SEGMENTS ON THE STRUCTURE AND FOLDING OF ATAXIN-2 AND ATAXIN-3 PROTEINS

Jingran Wen[1], Daniel R Scoles[2], Julio C Facelli[1]

[1]Department of Biomedical Informatics, and [2]Department of Neurology,

University of Utah, Salt Lake City, Utah, U.S.A

J Biomol Struct Dyn. 2017;35(3):504-19. Effects of the enlargement of polyglutamine segments on the structure and folding of ataxin-2 and ataxin-3 proteins. Wen J, Scoles DR, Facelli JC.

Abstract

Spinocerebellar ataxia type 2 (SCA2) and type 3 (SCA3) are two common autosomal-dominant inherited ataxia syndromes, both of which are related to the unstable expansion of tri-nucleotide CAG repeats in the coding region of the related ATXN2 and ATXN3 genes, respectively. The poly-glutamine (polyQ) tract encoded by the CAG repeats has long been recognized as an important factor in disease pathogenesis and progress. In this study, using the I-TASSER method for 3D structure prediction, we investigated the effect of polyQ tract enlargement on the structure and folding of ataxin-2 and ataxin-3 proteins. Our results show good agreement with the known experimental structures of the Josephin and UIM domains providing credence to the simulation results presented here, which show that the enlargement of the polyQ region not only affects the local structure of these regions but also affects the structures of functional domains as well as the whole protein. The changes observed in the predicted models of the UIM domains in ataxin-3 when the polyQ track is enlarged provide new insights into possible pathogenic mechanisms.

CHAPTER 4

MOLECULAR DYNAMICS ANALYSIS OF THE AGGREGATION PROPENSITY OF POLYGLUTAMINE SEGMENTS [a]

Jingran Wen[1], Daniel R Scoles[2], Julio C Facelli[1]

[1]Department of Biomedical Informatics, and [2]Department of Neurology,

University of Utah, Salt Lake City, Utah, U.S.A

Abstract

Protein misfolding and aggregation is a pathogenic feature shared among at least ten polyglutamine (polyQ) neurodegenerative diseases. While solvent-solution interaction is a key factor driving protein folding and aggregation, the solvation properties of expanded polyQ tracts are not well understood. By using GPU-enabled all-atom molecular dynamics simulations of polyQ monomers in an explicit solvent environment, this study shows that solvent-polyQ interactions decrease as the lengths of polyQ tracts increase. This study finds a predominance of long-distance interactions between residues far apart in polyQ sequences in longer polyQ segments, that leads to significant conformational differences. This study also indicates that large loops, comprised of parallel β-structures, appear in long polyQ tracts and present new aggregation building blocks with aggregation driven by long-

[a] Adapted from the manuscript submitted to PLoS One.

distance intra-polyQ interactions. Finally, consistent with previous observations using coarse-grained simulations, this study demonstrates that there is a gain in the aggregation propensity with increased polyQ length, and that this gain is correlated with decreasing ability of solvent-polyQ interaction. These results suggest the modulation of solvent-polyQ interactions as a possible therapeutic strategy for treating polyQ diseases.

Introduction

The polyglutamine (polyQ) diseases are caused by unstable expansions of CAG repeats resulting in proteins with expanded polyQ tracts. The polyQ diseases include Huntington's disease (HD), the spinocerebellar ataxias (SCAs 1, 2, 3, 6, 8, 7, 17), dentatorubral-pallidoluysian atrophy (DRPLA), and spinal and bulbar muscular atrophy (SBMA) [1-6]. Pathogenesis in these diseases is associated with abnormal polyQ protein folding [7-9] and resultant neuronal inclusion body formation [10, 8, 11-14]. While polyQ protein folding, stability, and aggregation have been well described for the polyQ diseases [15, 16], the molecular mechanisms leading to protein misfolding and aggregation, at the atomic level, are not well understood.

Computational simulations, using a variety of approaches, are used by several publications to study polyQ aggregation. In order to provide context to the relevance of the studies presented here, this section briefly discusses how computational simulations are applied in the study of structure and aggregation of polyQ. Laghaei and Mousseau [17] performed Replica-Exchange Molecular Dynamics (REMD) simulations of polyQ monomers and dimers with 30Q to 50Q repeats in length, with an implicit solvent force filed, and observed that polyQ dimers with more than 40 repeats adopt antiparallel β-sheets, and triangular and circular β-helical structures. Nakano et al. [18], using the same simulation methods, explored the conformation ensemble of short monomeric polyQ tracts with 15 glutamines. They find that Q15 monomers can assume multiple configurations and that the formation of oligomers is stabilized by the β-turns and hydrogen bonds between the main chains. The same authors performed REMD [19] on short polyQ monomers and dimers with lengths ranging from 3Q to 15Q; their results demonstrate that polyQ dimers strongly favor the formation of antiparallel β-sheet structures. Chiang et al.[20] studied the aggregation of short polyQ segments with 15 residues using REMD with explicit solvent. Their results show that polyQ dimers mainly form helix and coil structures when they are far apart, but as the interpeptide distance decreases, an inter-peptide β-sheet is formed. Zhou et al. [21] studied the early stage of polyQ aggregation using initial β-helical structures of various shapes and sizes, which include left-handed circular, right-handed rectangular, and left- and right-handed triangular. Their results show that the stability of the β-helical structures increases as the number of rungs increases, and that the 3-rung left-handed triangular and right-handed rectangular β-helical models are stable. Using the all-atom REMD, Hayre et al. [22] find that the left-handed β-helical conformations are stable for polyQ tracts with 61 residues. To explore the stability of α-sheet configuration of polyQ, Babin et al. [23] performed Molecular Dynamics (MD) simulations in an explicit solvent environment on short polyQ dimers with 8 glutamines and find that the α-sheet configuration is a stable, metastable, or at least long-lived secondary structure. Miettinen et al.[24] assessed the stability of a 40Q polyQ segment with six different initial conformations, including steric zipper, β-nanotube, β-pseudohelix, β-sheetstack, β-sheet, and α-helix. Using atomistic MD simulations in explicit solvent, they find that β-hairpin-based (β-sheet and β-sheetstack) and α-helical conformations are kinetically stable enough to serve as a template to initiate polyQ fibrillation. The same group also studied the stability of polyQ dimers with the six initial conformations and finds that β-hairpin-containing conformers can form very stable dimers when their side chains are interdigitated, whereas dimers of α-helix, steric zipper, β-nanotube, and β-pseudohelix conformers are too short-lived to initiate aggregation [25]. Using both 2D IR spectroscopy and MD simulation, Buchanan et al. determined that stacked β-hairpins are the dominant structure of polyQ fibrils [26]. Wang et al. [27] conducted solvent-free multiscale coarse-grained models on polyQ segments with lengths ranging from 8Q to 56Q, and find that the degree of aggregation increases with the length and concentration of polyQ chain. Deng et al. [28] studied the polyQ aggregation on Q32 using solvent-free multiscale coarse-gained MD simulations, and their results show that polyQ aggregation is sensitive to concentration and temperature changes. Lakhani et al. show that the β-rich region in the exon one of huntingtin protein misfolds with increasing polyQ lengths (Q23-Q47) [29]. Similarly, the work of Morida et al. finds the presence of a larger population of aggregation-forming structures in 40Q polyQ segments than in those below the pathogenesis threshold [30]. Stork et al. [31] studied the stability of several types of β-helix of polyQ, polyA, and polyS using MD simulations, and find that the water-filled channels inside the β-helix can destabilize the β-helix structures. This study demonstrates that protein-solvent interaction may compete efficiently with the intramolecular hydrogen bonds, affecting conformation stability and aggregation. However, few studies employ full explicit solvent MD simulations to evaluate solvent effects on polyQ aggregation as a function of the polyQ segment length.

Although multinanosecond, all-atom and explicit solvent MD simulations can be a powerful means to further study solvent effects on the folding and aggregation properties of polyQ proteins, the high computational cost of all-atom and explicit solvent MD has been a major deterrence for this type of studies. The use of GPU acceleration hardware for MD simulations can efficiently speed up these calculations [32] and makes it possible to do atomic MD simulation with an explicit water environment in a reasonable time period [33]. Amber, one of the most popular MD programs, is able to use NVIDIA GPUs to massively accelerate MD for both implicit [34] and explicit solvent simulations [35, 36] with a dramatic performance improvement. Here we report the study of solvent effects on solution properties, folding, and aggregation propensity of simple polyQ sequences, which can be considered an initial and computable model for the study of solvent effects on the aggregation propensity of polyQ disease related proteins at the atomistic level [37]. Of note is that, even with the advances provided by the GPU technology, all-atom MD studies of full polyQ proteins in explicit solvent are still nonfeasible. Taking advantage of the GPU speedups, we are able to perform MD runs of these polyQ models with repeat lengths in both the normal and the pathological ranges, using explicit solvent and simulation lengths of 105 ns, within reasonable computational wall times.

In this paper, we report results of 105 ns simulation of polyQ segments of different repeat lengths, in the range observed for polyQ diseases, using full atomistic MD simulations with explicit solvent.

Methods

The polyQ tract is the only common region observed in the otherwise very dissimilar polyQ proteins which are associated with polyglutamine diseases, and in all cases, the polyQ expansion causes the disease. The threshold length of the polyQ segment that triggers these diseases is around 35 to 40 residues, except in SCA6 which has a shorter threshold of around 19 repeats [38-40]. Therefore, it is of interest to study the solvation behavior of polyQ segments shorter than 20 and longer than 40 repeats to find common features on how solvent interactions may affect the folding of such diverse set of proteins.

We performed MD simulations for polyQ monomers with 18 repeats (Q18), 46 repeats (Q46) and 32 repeats (Q32). These correspond to lengths below the lowest known disease threshold, above the highest known normal threshold and the average repeat length of these two, respectively. The extended structure of polyQ was used as the starting structure of the MD simulations. In order to avoid complications due to charged termini [41], the polyQ sequences were capped with an acetyl group in the N-terminus and a N-methylamide group in the C-terminus, i.e., the structures considered here are [acetyl-$(Gln)_n$-N-methylamide], where n = 18, 32, and 46 denotes the number of glutamines. xLEaP [42] was used to build the initial configurations, and the Amber force field, AMBER ff99SB force field[43], was used with a TIP3P water box to provide an explicit simulation of the solvent. A local minimization of the polyQ monomers was done in vacuum before the water box was added. The TIP3P water was included in a truncated octahedral box added to the polyQ monomer with a buffering distance of 9.0 Å between the edges of the box and the polyQ monomer. A second minimization was performed on the solvated system using a nonbonded cutoff distance of 9 Å to minimize the energy of the whole system. The whole system was then heated from 0 K to 310 K and equilibrated for 50 ps, followed by molecular dynamics simulations for 105 ns at the temperature of 310K and constant pressure of 1 atm. The temperature was maintained through the Berendsen thermostat with a coupling time of 0.1 ps. Isotropic position scaling was used to maintain the pressure and a relaxation time of 1 ps was used. The velocity was updated every 2 fs, and results were recorded every 1 ps.

For each polyQ monomer, six independent runs were performed and the results presented here are the average for these runs. All the MD simulations were done using the Amber 14 molecular simulation package [44] which supports a GPU accelerated PMEMD module, that implements the Particle Mesh Ewald (PME) method for electrostatics [36]. All calculations were performed using the clusters at the Center for High Performance Computing (CHPC) at the University of Utah. Each computing node in the cluster has two Nvidia 2090 GPUs and 12 Intel Xeon (Westmere X5660) processors. After a preliminary study to optimize the efficiency of the GPU-accelerated computing nodes (results not shown), we performed one simulation per GPU to obtain the best throughput performance with the settings of our cluster.

The Cpptraj utility in the Amber 14 tool box [44] was used for most of the analysis. The MD trajectories were re-imaged back to the primary box, and to speed up the analysis, only 1/100 of the frames were processed that is 100 ps per frame in the new trajectory. The secondary structure, hydrogen bond, solvent bridge, radius of gyration, and solvent surface area were calculated using Cpptraj for each simulation trajectory. The Rg value of the polyQ segments was calculated for each frame of the 105 ns and these values were used to calculate the exponent factor b in the Rg ~ N^b. The log transform was done on each data point, and a linear regression was fit to get the exponent factor b.

For each polyQ length, the results of six independent simulations were averaged, such that all values reported here represent the average values over these six runs. Statistical analyses were performed using R [45], figures were plotted with ggplot2 package [46] and Gnuplot [47], and VMD was used for trajectory visualization [48].

Results

Overall GPU Performance

The systems considered here, including both polyQ monomer and water solvent (Table 4.1), are large enough to exhibit excellent scaling when using GPUs. The GPU version of the Amber PMEMD module on the GPU furnished nodes provides highly consistent speedups, with an average factor of 8.5 times speedup over the CPU times.

Secondary Structure

Our simulations show that polyQ monomers can adopt various secondary structures instead of fixed structures during the 105 ns simulations. This is consistent with previous results showing that polyQ monomers are disordered [37]. All monomers predominately adopt helical structures (Figure 4.1), but the types and proportions of each of the helical structures vary among monomers with different repeat lengths. Q18 monomers show the highest proportion of helical structures including 3-helix and α-helix, whereas Q32 monomers adopt the lowest proportion of helical structures on average. It is apparent from Figure 4.1 that the number of β-structures, especially parallel ones, increases as the length of the polyQ segment increases. This is an important structural change as it has been established that parallel β-structures are a precursor for initiating aggregation [49].

The stability of the β-structures, as a function of time, is also different for monomers of different lengths (Figures 4.2-4.4). The simulations show that the parallel β-structures in Q46 monomers are very stable and most of them can last for the entire simulations (Figure 4.4), whereas in Q18 and Q32 these structures are less stable, occurring only in 0.1% of the simulation time in Q18 (Figure 4.2) and around 1% of the time for Q32 (Figure 4.3).

Table 4.1 Comparison of AMBER CPU and GPU performance for simulations of polyQ monomer in explicit solvent with different number of repeats.

	number of atoms	CPU Performance (ns/day)	GPU performance (ns/day)
Q46	341,249	0.30±0.01	2.73±0.02
Q32	134,359	0.86±0.00	7.49±0.04
Q18	34,407	3.55±0.01	27.75±0.12

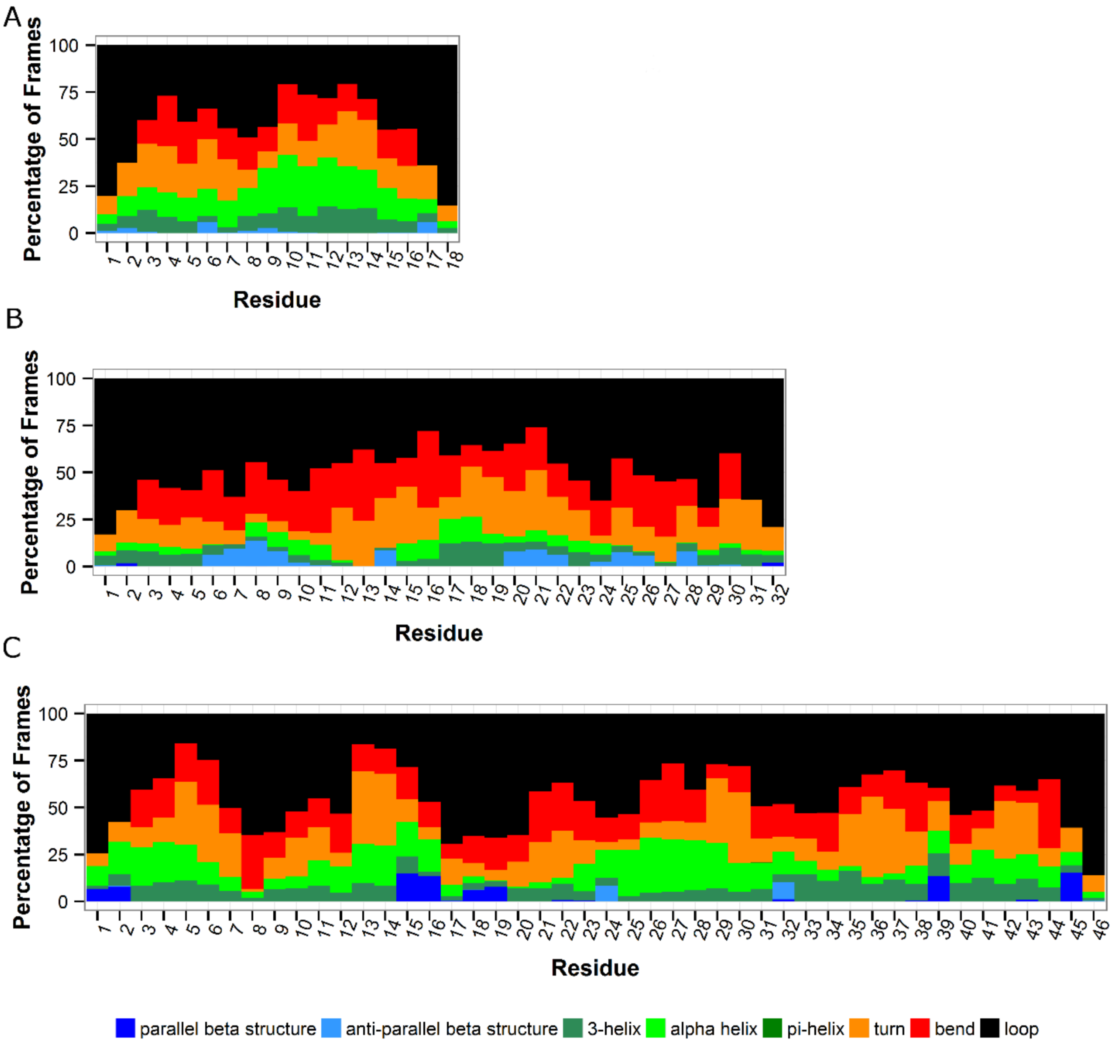


Figure 4.1 Secondary structure of polyQ fragments of different lengths. A Q18; B Q32; C Q46. Colours indicate different types of secondary structures. Blue: parallel β structure; Sky blue: antiparallel β structure; Dark green: 3-helix; Green: α-helix; Olive: pi-helix; Dark orange: turn; Red: bend; Black: loop. X-axis: residue index; Y-axis: percentage of frames in the 105 ns simulations, averaged over the six runs performed here.

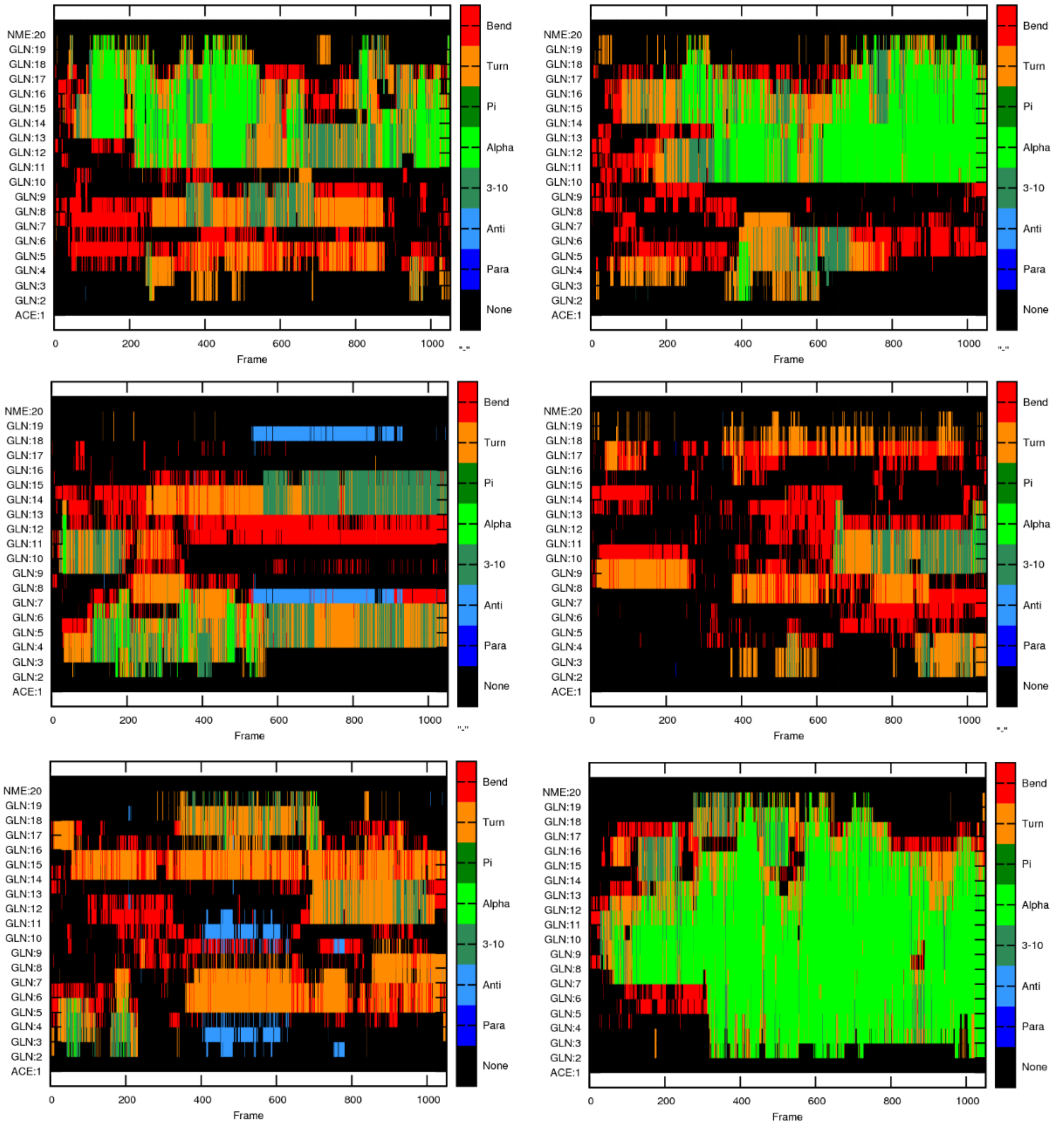


Figure 4.2 Secondary structure of Q18 monomers at different time frames for each of the six independent MD runs performed. X-axis: frame index with each frame representing 100 ps of simulation; Y-axis: residue index indicating the secondary structure as depicted at the right panel.

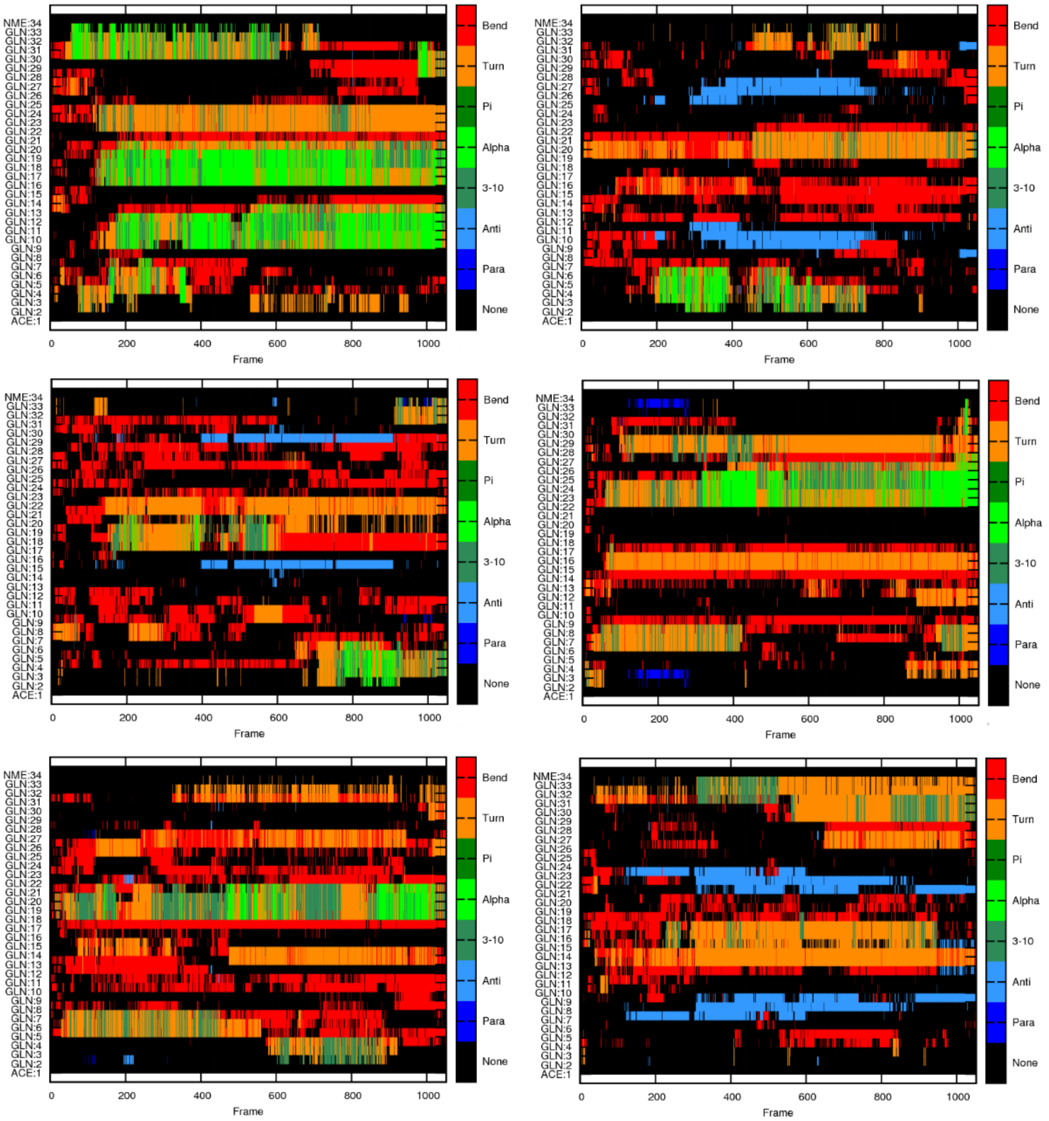


Figure 4.3 Secondary structure of Q32 monomers at different time frames for each of the six independent MD runs performed. X-axis: frame index with each frame representing 100 ps of simulation; Y-axis: residue index indicating the secondary structure as depicted at the right panel.

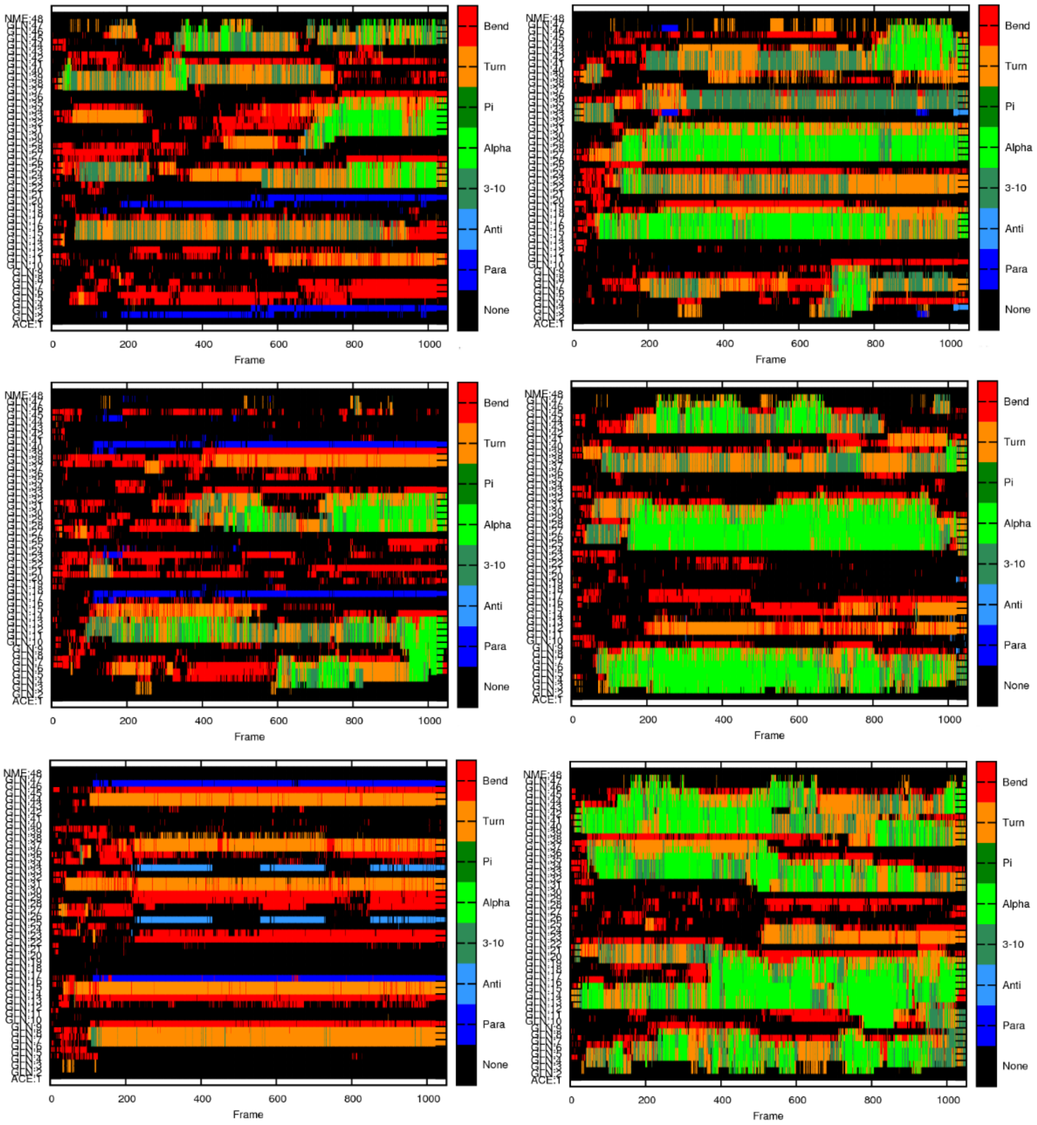


Figure 4.4 Secondary structure of Q46 monomers at different time frames for each of the six independent MD runs performed. X-axis: frame index with each frame representing 100 ps of simulation; Y-axis: residue index indicating the secondary structure as depicted at the right panel.

Hydrogen Bonding

Hydrogen bonding plays a critical role in polyQ folding and stability [50], therefore changes in hydrogen bond patterns with polyQ expansion may be a signal of changes in folding and aggregation propensity. Of particular interest is the balance between intra-polyQ and polyQ-solvent hydrogen bonds. In this study, we term the hydrogen bond as intra-polyQ if both donor and acceptor are from glutamine residues, and hydrogen bonds between glutamine residues and solvent water molecules are called solvent-polyQ hydrogen bonds. As the amide group in the sidechain of a glutamine can provide one hydrogen donor (hydrogen in NE2) and two hydrogen acceptors (NE2 and OE1), the intra-polyQ sidechain hydrogen bonds can be both backbone-sidechain as well as sidechain-sidechain hydrogen bonds. The hydrogen bonds are identified using the hbond command in Cpptraj program in the Amber 14 Toolbox. The distance cut-off is set at 3.5 Å, and the angle cut-off is set at 120°. Therefore, there can be more than one water molecule, surrounding one glutamine, which meet these criteria. So the number of hydrogen bonds reported in this study represents the dynamic count of the number of hydrogen bonds detected by Cpptraj over all six simulations.

Intra-polyQ hydrogen bonds. Using the procedure described above, the number of hydrogen bonds is counted for each individual frame in the 105 ns MD simulation for each MD run. The normalized count of hydrogen bonds per 100 Qs, which is the number of hydrogen bonds normalized by length of the polyQ segment multiplied by 100, is calculated as a measure of the relative ability of polyQ monomers to form hydrogen bonds.

While longer polyQ monomers adopt more intra-polyQ hydrogen bonds than shorter ones (Figure 4.5A, red), the normalized count of intra-polyQ hydrogen bonds per 100

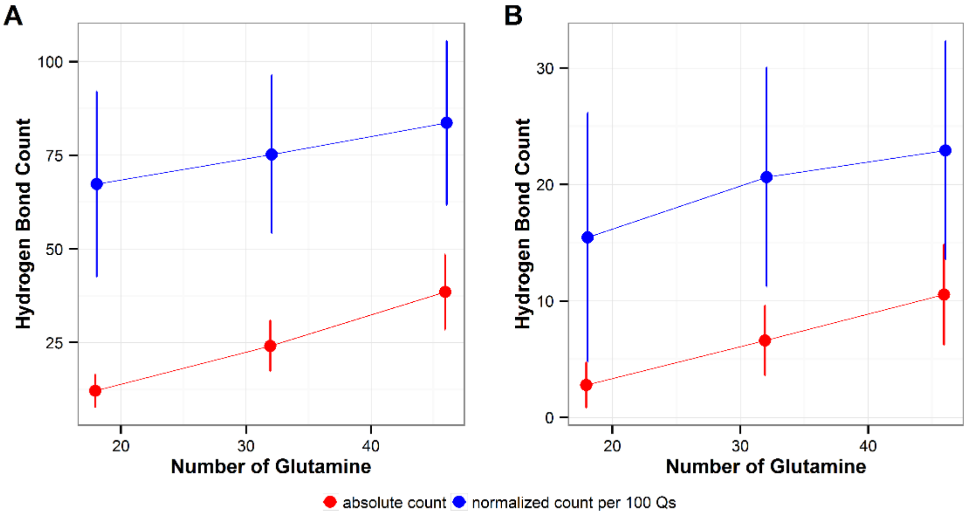


Figure 4.5 Intra-polyQ hydrogen bonding. A total number of intra-polyQ hydrogen bonds; B sidechain-sidechain hydrogen bonds. Ted: total count of hydrogen bonds; Blue: normalized count of hydrogen bonds per 100Qs.

glutamines increases as the monomer length increases (Figure 4.5A, blue). There is a linear relationship between normalized intra-hydrogen bonds and monomer length ($p<0.001$, $r=0.282$).

When considering intra-polyQ hydrogen bonds formed with the glutamine sidechains, the number of sidechain-sidechain hydrogen bonds increases with the number of glutamines in the polyQ tract ($p<0.001$, $r=0.70$) (Figure 4.5B, red). For polyQ monomers of the same length, the number of sidechain-sidechain hydrogen bonds is consistent among simulations, and independent of the secondary structure (Figure 4.6). When normalized by the number of glutamines in the polyQ tract, the normalized count of sidechain-sidechain hydrogen bonds per 100Qs also increases with polyQ length ($p<0.001$, $r=0.30$) (Figure 4.5B, blue). The normalized count of hydrogen bonds per 100 Qs formed by glutamine sidechains, including both sidechain-sidechain and sidechain-backbone hydrogen bonds, are similar in 32 Q and 46Q polyQs, but are fewer in the 18Q polyQ (Figure 4.7).

At the residue level, all polyQ tracts studied here show some common hydrogen bond patterns. The results of this study show that, in all of the repeat lengths studied here, the i^{th} residue prefers forming hydrogen bonds with residues in the position of $i+2$, $i+3$ or $i+4$ (Figure 4.8). We verified that both backbone-backbone and sidechain hydrogen bonds contribute to the patterns of $i+2$, $i+3$, and $i+4$, but that the backbone-backbone hydrogen bonds contribute more than sidechain ones. Some hydrogen-bonded residue pairs are ‘hot’ in all the polyQ segments studied here and this trend is independent of the lengths of the polyQ monomers. Residues 1 and 4 show hydrogen bond propensity in 4 out of the 6 MD simulation runs of 18Q, 32Q, and 46Q polyQ segments. In addition to these common patterns, the intra-polyQ hydrogen bonds also have length-dependent features. The long-

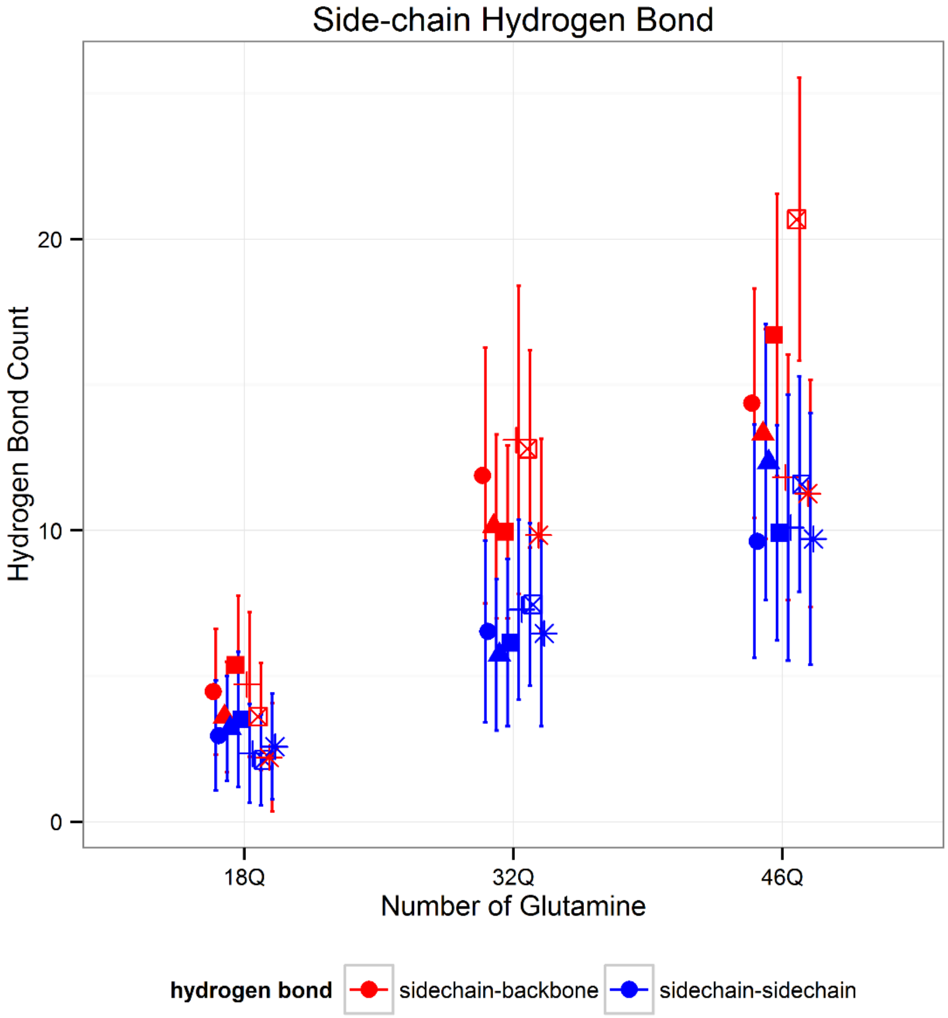


Figure 4.6 Sidechain hydrogen bond. Red: sidechain-backbone hydrogen bond; Blue: sidechain-sidechain hydrogen bond. Shapes indicate different experiments. From left to right, Q18, Q32, and Q46.

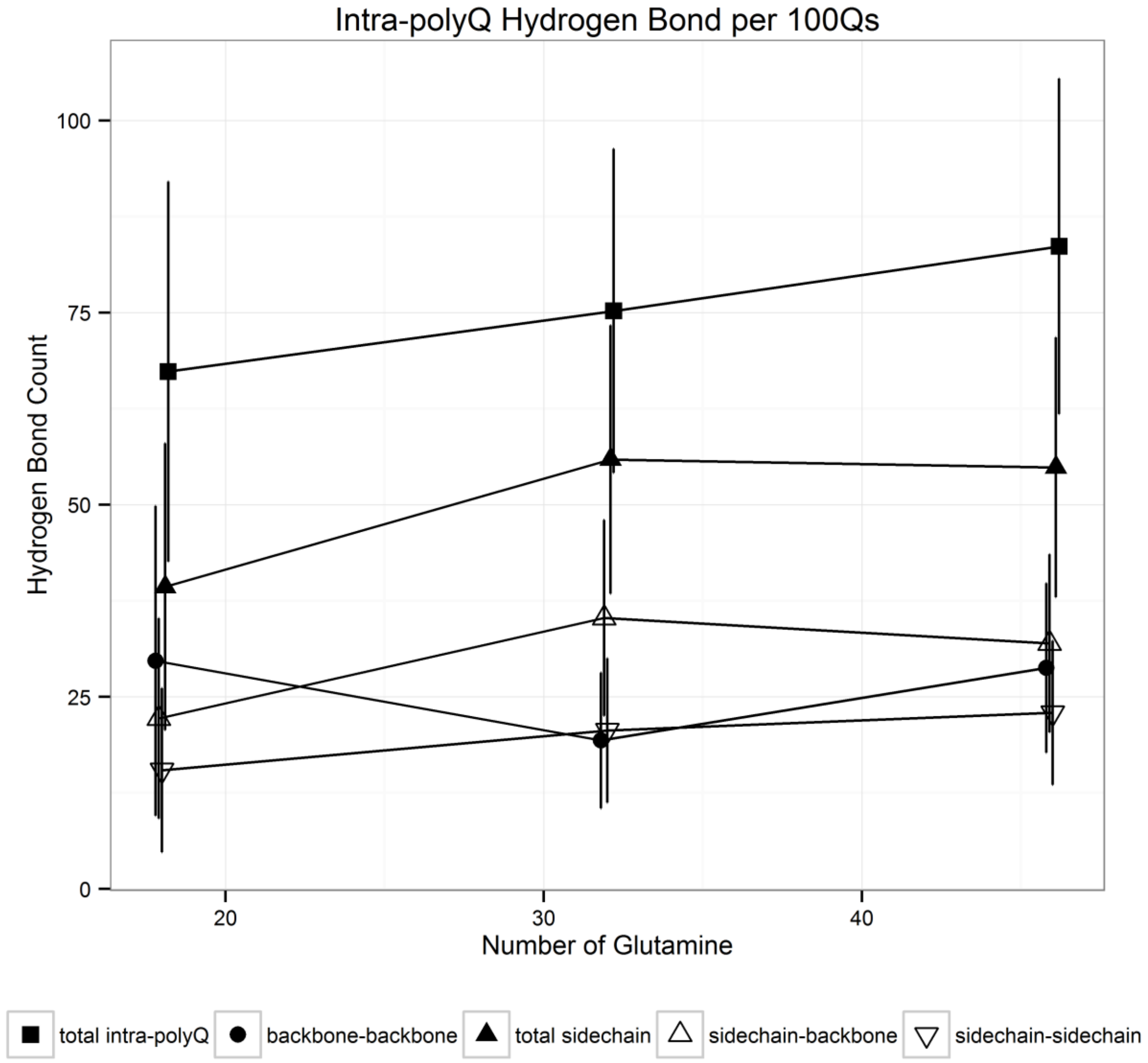


Figure 4.7 Number of intra-polyQ hydrogen bond normalized by the length of polyQ monomer. X-axis: the length of polyQ monomer; Y-axis: the number of hydrogen bond. Shapes indicate different types of hydrogen bonds.

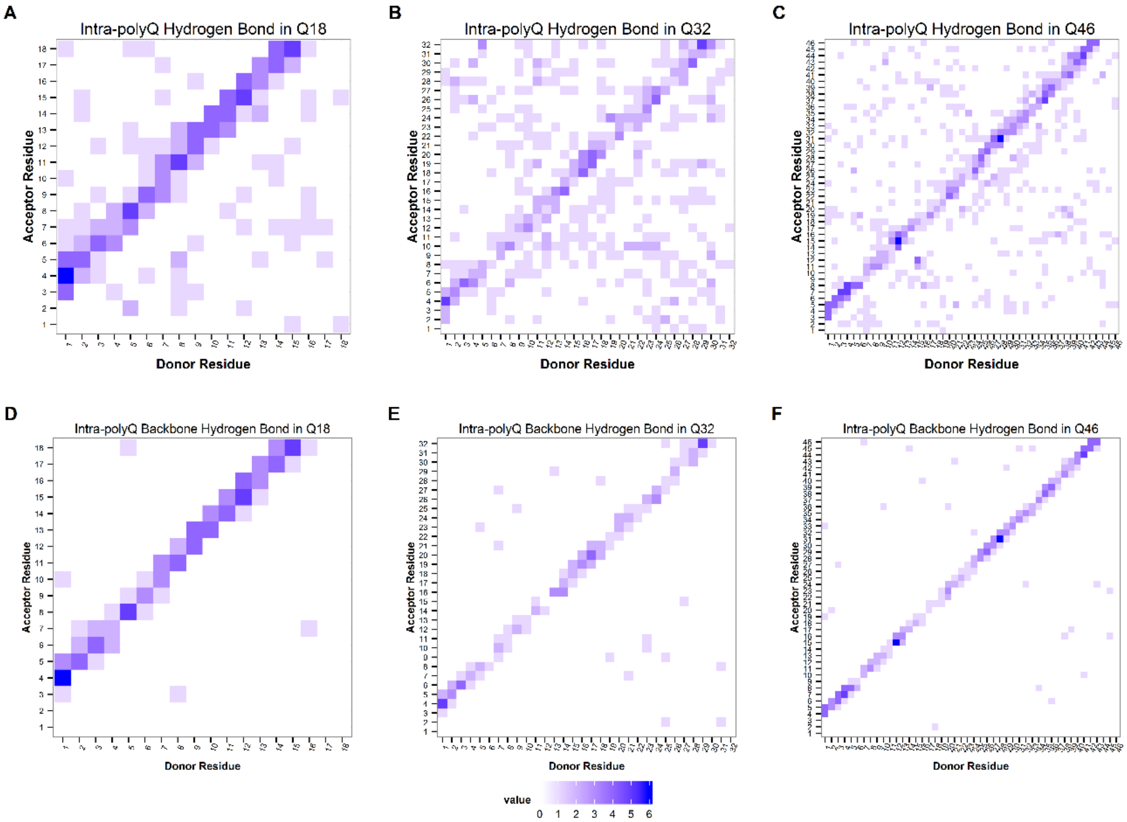


Figure.4.8 Intra-polyQ hydrogen bond among different experiments. A, B, and C represent the total intra-polyQ hydrogen bond; D, E, and F represent the backbone-backbone hydrogen bond. A and D Q18; B and E Q32; C and F Q46.

ranged hydrogen bonds considered here are the ones formed by two glutamines with a sequence distance longer than at least half of the length of the polyQ monomer. The percentage of long-ranged hydrogen bonds is greater in the longer polyQ tracts than that shown in the shorter ones. For example, when considering hydrogen bonds with a time frequency greater than 50%, all of the hydrogen bonds in Q18 are short-ranged ones (Figure 4.9), whereas 4.5% and 10.0% of the hydrogen bonds are long-ranged ones in Q32 and Q46, respectively. For Q32, the long-lived hydrogen bond can occur even between glutamines that are 15 residues apart in the polyQ sequence, and this distance can extend to 30 residues in 46Q polyQ monomers.

Solvent-polyQ hydrogen bonding. As expected, the number of solvent-polyQ hydrogen bonds, which are calculated using the criteria defined in the above section, increases as the length of polyQ monomers increases (Figure 4.10A). The slope of the increase is different among different types of hydrogen bonds, with sidechain solvent hydrogen bonds increasing the greatest (Figure 4.10A). However, when the total number of intra-polyQ hydrogen bond is normalized by the number of repeats in the polyQ segment, this normalized number of hydrogen bonds decreases as the polyQ length increases (Figure 4.10B), which is the reversed trend from what observed for the normalized number of intra-polyQ hydrogen bonds. When classified at the atomic level, the number of hydrogen bonds using each atom, shown in Figure 4.11, also increases with the length of polyQ (Figure 4.11A), with sidechain O-mediated hydrogen bonds increasing the greatest. However, when normalized by the polyQ segment length, the number of sidechain O-mediated hydrogen bonds decreases with the polyQ length, as did the backbone O-mediated hydrogen bonds (Figure 4.11B). The number of normalized

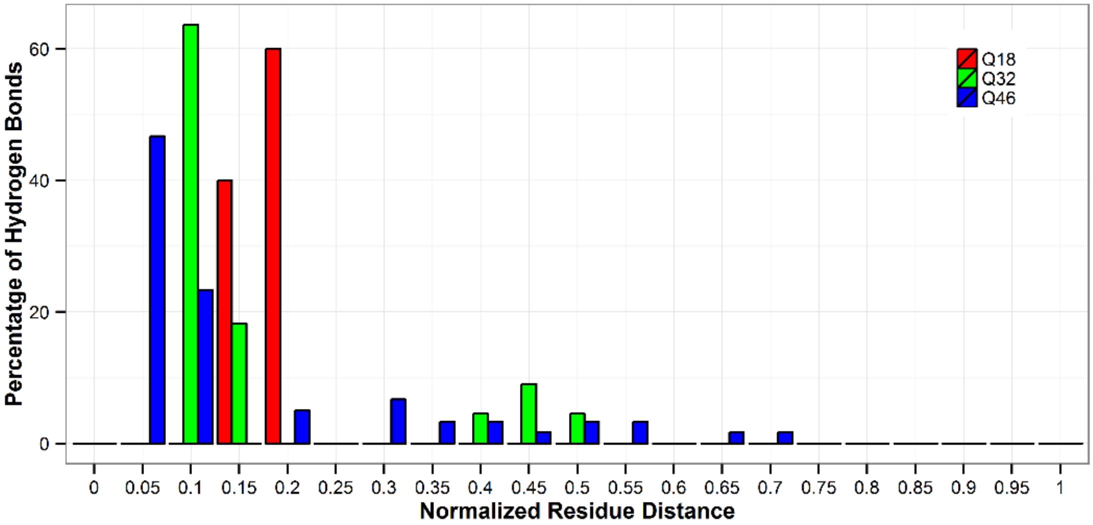


Figure 4.9 Distance distribution of observed hydrogen bonds with more than 50% frequency. The normalized distance is calculated as (|acceptor residue index-donor residue index|-1)/(number of repeat in polyQ-2). Red: Q18; Green: Q32; Blue: Q46.

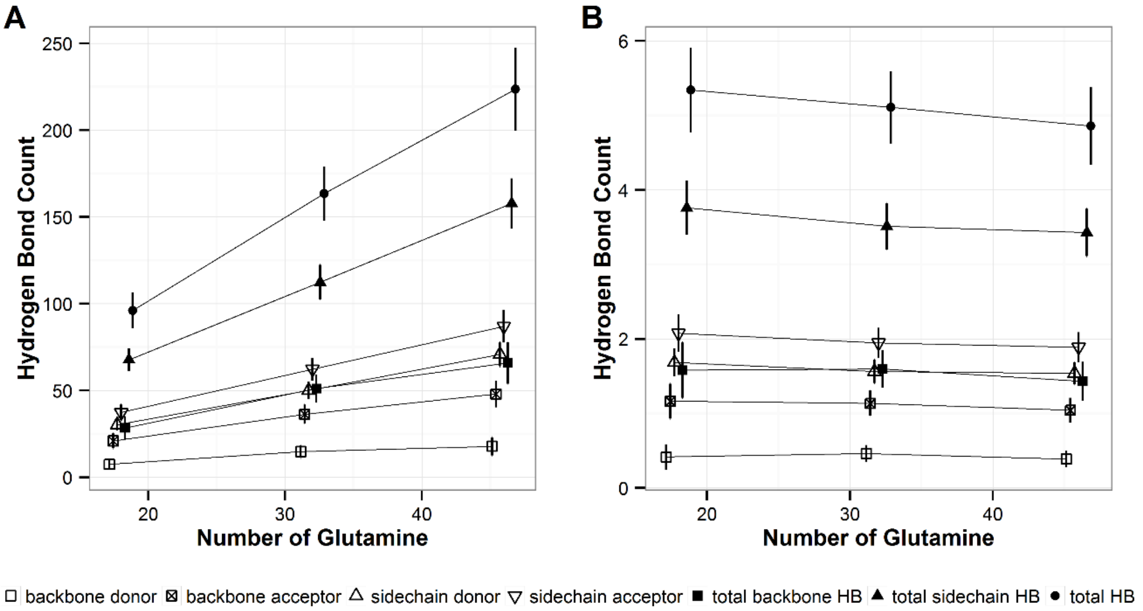


Figure 4.10 Solvent hydrogen bond count. A total count; B count normalized by polyQ length. Shapes indicate different hydrogen bond types.

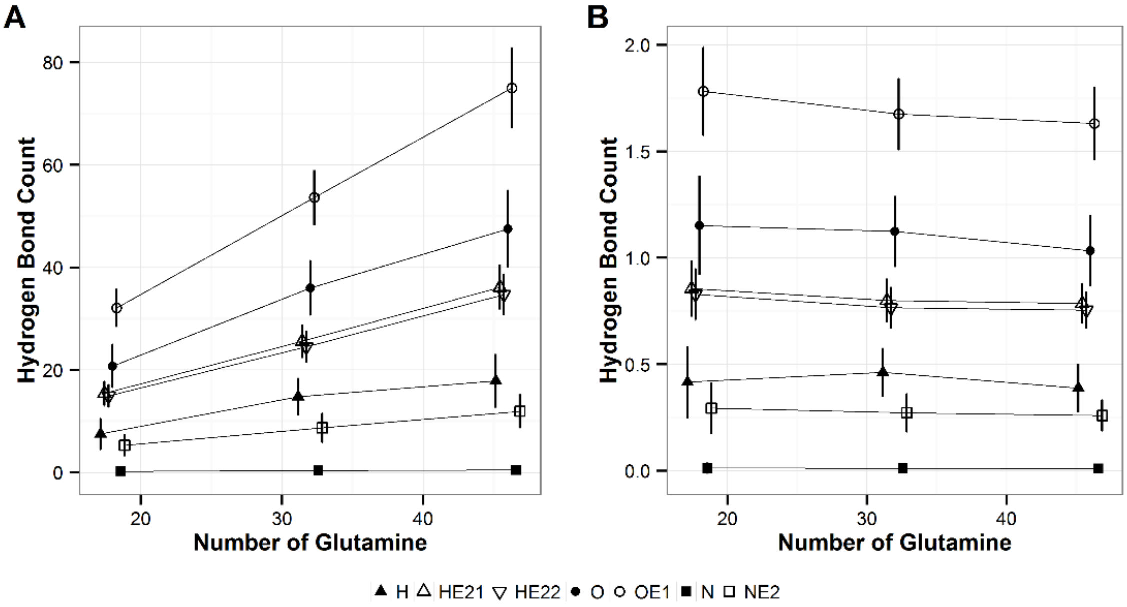


Figure 4.11 Solvent hydrogen bond count at the atomic level. A total count; B count normalized by polyQ length. Shapes indicate different types of hydrogen bond donor and acceptor.

hydrogen bonds formed by other atoms do not change substantially, and are similar among polyQs with different lengths (Figure 4.11B).

Intra-polyQ hydrogen bond vs. solvent-polyQ hydrogen bond. For each simulation time frame, the potential number of intra-polyQ hydrogen bonds and the potential number of solvent-polyQ hydrogen bonds are calculated and the values are plotted in Figure 4.12. As expected, there is strong positive linear correlation between the total number of hydrogen bonds of both types that increases with the length of polyQ tract (Figure 4.12A), but the direction of the correlation dramatically changed to a negative relationship when considering the normalized count per 100 Qs (Figure 4.12B). This negative relationship is a very strong indication that the relative proportion of intra-polyQ hydrogen bonds increases in detriment of solvent-polyQ ones for longer repeat (Figure 4.12B).

Solvent Bridges

Water solvent molecules can form bridges with glutamine residues in the polyQ tracts, and these bridges can affect folding and structure stability of polyQ tracts. Therefore, it may be expected that if polyQ tracts with different lengths have different solvent bridge patterns, their folding and structural integrity will also be affected. Figure 4.13 depicts the frequency of solvent-bridged glutamine pairs with the normalized residue distances. The bridges considered here are the ones that show in more than 100 time frames which correspond to at least 10% of the simulation time. We find that the frequencies of occurrence, for the observed bridges, range from 1% to 50%. Although the number of long-ranged bridges is small among all three polyQ lengths considered here (Figure 4.13), polyQs with 32Q and 46Q repeats form more long-ranged bridges than the polyQ monomers with 18Q repeats. More than 5% of these bridges are long-ranged ones in 32Q

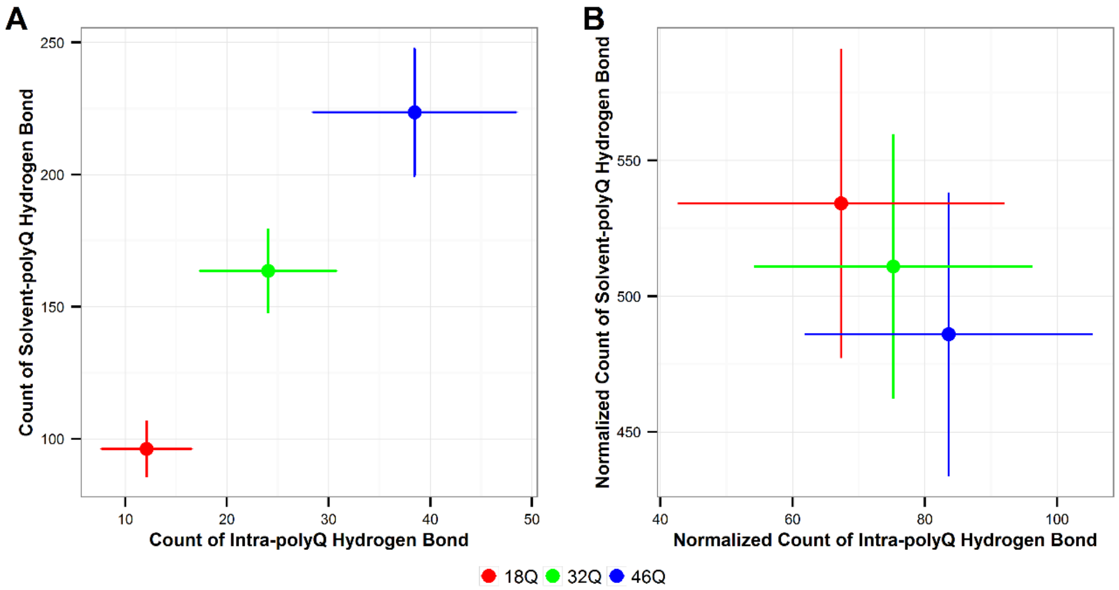


Figure 4.12 Number of intra-polyQ hydrogen bonds vs. the number of solvent-polyQ hydrogen bond. A total count; B count normalized by number of repeats. The error bars are the standard deviation from all the MD simulations for each polyQ length. Red: Q18; Green: Q32; Blue: Q46.

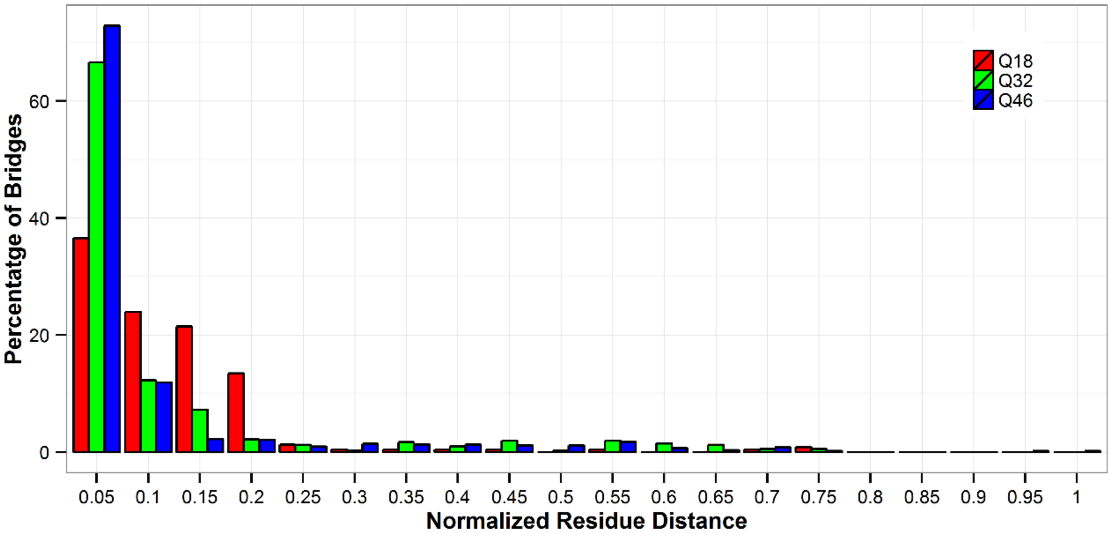


Figure 4.13 Distance distribution of observed solvent bridges lasting more than 100 frames in the simulation. The normalized distance is calculated as (|acceptor residue index-donor residue index|-1)/(number of repeat in polyQ-2). Red: Q18; Green: Q32; Blue: Q46.

and 46Q polyQs, whereas only ~1.6% of the bridges are in long range in the 18Q polyQs. These results are consistent with the above discussion on the hydrogen bond results, both of which show a substantial decrease of solvent interactions and likely, more compact structures as the length of the polyQ tracts increases.

Radius of Gyration (Rg)

The Rg is used to describe the compactness of a protein. It is defined as the root mean square distance of the collection of atoms from their common center of gravity.

For flexible polymers, the Rg value is proportional to N^b, where N is the length of the polymer [51] and b is characteristic of the solubility of the polymer. A good solvent is characterized by an exponent of ~ 0.59, as chain-solvent interactions are preferred, whereas a poor solvent has an exponent value of ~ 0.33, as the chain collapses to minimize contact with solvent [51] .Using the 105-ns simulation data, we find a value of b = 0.45 for the polyQ segments in water solution studied here (Figure 4.14). This indicates that there is a tendency to collapse among the polyQ structures as the length of the polyQ segment increases. Consistent with the results of previous sections in this paper, the results of Rg indicates that longer polyQ segments are less soluble, which is also consistent with an increase of their propensity to aggregation as the length of the polyQ increases.

Solvent Accessible Surface Area (SASA)

As expected, the total SASA of polyQ segments studied here increases as the number of polyQ length increases (Figure 4.15). Both total backbone and total sidechain SASA follow the same trend, but sidechain SASA increases faster than the backbone SASA. However, when SASA is normalized by the length of polyQ, this normalized SASA value

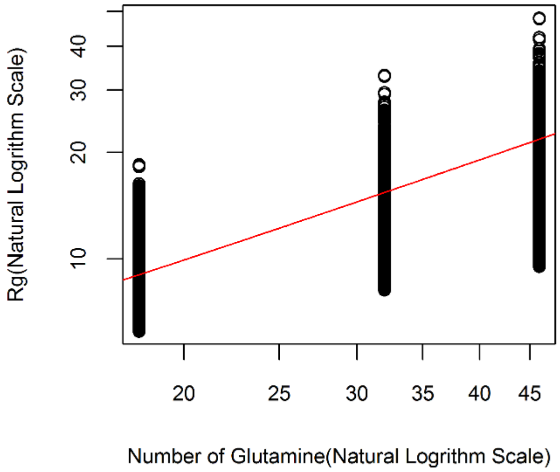


Figure 4.14 Scaling laws for polyQ monomers in water. Black dots are the Rg values of each frame, and the red line is the regression line of number of repeat and Rg, both in natural log-scale.

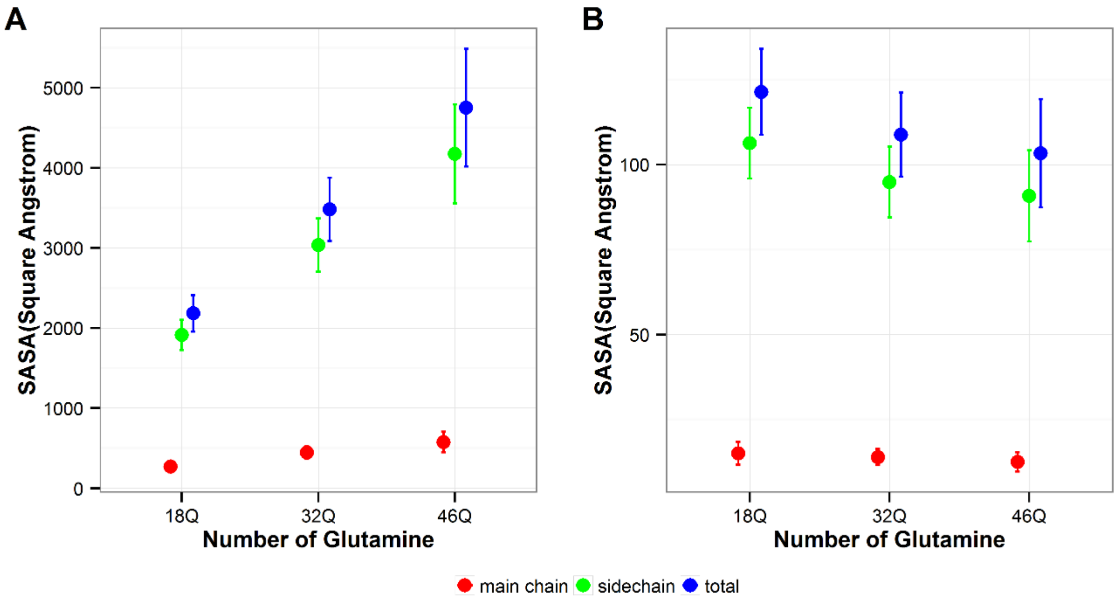


Figure 4.15 Solvent accessible surface area. A total SASA; B normalized SASA. Red: backbone SASA; Green: sidechain SASA; Blue: total SAS

decreases with polyQ length (Figure 4.15), which is the same trend from what observed for the Rg values. The normalized backbone SASA for all polyQ sequences studied here is on average smaller than 20 Å^2, therefore it is likely that polyQ backbone may be buried inside the structures rather than residing at the surface [52]. These results are also consistent with the results of previous sections, all of which indicate that the ability of polyQ monomers to interact with the solvent decreases as the length of the polyQ sequence increases.

Discussion

Consistent with the fact that the expansion of polyQ sequence beyond a certain threshold, specific for each polyQ disease, triggers pathogenesis [53], numerous observations have suggested that the polyQ tract by itself may play a central role in the pathogenesis of the ten known polyQ neurodegenerative diseases [13].

In this study, three different lengths of polyQ segments are considered, Q18, Q32, and Q46, to cover normal, intermediate and pathological ranges relevant for most of the polyQ diseases. The full atomistic MD simulations with explicit solvent presented here show that all polyQ segments mainly form random-coiled structures, which is consistent with previous literature studies [20]. But the results in this study also show an increasing propensity to form helical and β structures as the number of glutamines increases in the polyQ tract. The type of β-structures are different among polyQ monomers of different lengths. The β-structures in Q46 are dominated by parallel β-structures, whereas for Q18 and Q32, the majority are antiparallel β-structures. While Q18 and Q32 polyQ monomers can form parallel β-structures, these structures are not stable and cannot last till the end of the simulations (Figures 4.2 and 4.3). On the contrary, for Q46 the parallel β-structures, once formed, can last to the end of the simulations, which may be a clear evidence of the formation of a proto-structure conducive to aggregation.

In this study, the MD simulations of polyQ segments in water predict a b scaling factor for the Rg of 0.45 indicating that water is not a good solvent for polyQ [54, 51]. Consistent with the results of Vitalis et al. [51], this result indicates the decreased preference of solvent-polyQ interaction as the number of repeats increases in polyQ monomers. This observation is also consistent with all the results, performed here, of the changes in hydrogen bond patterns as the lengths of the polyQ sequences increase.

The results of the normalized SASA also support the idea that the preference of water-polyQ interaction decreases as the length of polyQ increases. Although the total SASA is larger for polyQ monomers with longer repeats, the SASA per residue decreases as the repeat number of polyQ tract increases, especially for the sidechain surface area (Figure 4.15).

This study also explores the preference of the intra-polyQ vs. solvent-polyQ hydrogen bond formation, and the results show that the normalized number of hydrogen bonds per residue increases for the former and decreases for the latter type of hydrogen bond, as the number of repeats increases (Figure 4.5 and Figure 4.10). Q18, Q32, and Q46 can potentially form long-ranged hydrogen bonds. Considering the hydrogen bonds that show in more than 50% of the simulation time, in Q18 the majority of them are short-ranged ones with residues that are only 2- and 3-residue apart. However, long-ranged hydrogen bonds do exist in a larger proportion in Q46 (Figure 4.9). Driven by the long-distance interaction, polyQ sequences with longer lengths can fold into more compact structures, which also indicates an increasing propensity to avoid solvent interactions.

All the results presented here consistently point towards an increased propensity to hydrophobicity as the polyQ segments become longer. This raises the hypothesis that the pathogenic cause of the polyglutamine diseases may be rooted in the increased hydrophobicity of their polyQ tracts, which hydrophobicity, by increasing protein aggregation, causes neural degeneration. While results of this study does not provide direct evidence of the role which the polyQ segments play in polyglutamine protein aggregation, in the neurodegenerative diseases considered here, given the fact that the only common element of these diseases is the enlargement of polyQ segments in their associated proteins, the results presented here provide impetus to further exploring the hypothesis listed above.

This study is not without limitations. Only polyQ monomers are studied and the inter-molecular interactions among polyQ monomers, which can contribute to aggregation [55], are not included in this study. Additionally, regions flanking the polyQ tract are not considered in this study. Results of our previous study have demonstrated that regions flanking polyQ tracts alter polyQ secondary structure models [56], consistent with findings that these flanking regions alter aggregation of polyQ proteins [57, 29, 58]. However, with the existing study settings, it is easy for us to study the solvation of polyQ tract with the sequence context of the polyQ proteins, both monomers and polymers in the future.

Conclusions

This paper studies the effect of solvation on the folding of polyQ segments with repeat lengths in the normal, intermediate, and pathological ranges using all-atom molecular dynamics simulations with an explicit water solvent environment. In accordance with the literature, the results of this study show that polyQ monomers can fold, but are disordered, when in water. GPU acceleration has effectively improved computational performance, which makes it possible to study the polyQ aggregation in all-atom explicit environment with reasonable time. The simulations show that, as the length of a polyQ monomer increases, the water solubility of the polyQ segments decreases, while the propensity to form more compact structures with intra-polyQ hydrogen bonds increases. The results of this study demonstrate gains in aggregation propensity with increased polyQ lengths that correlates with decreasing ability of solvent-polyQ interaction. These results are consistent with previous observations using coarse-grained simulations, and suggest that modulation of solvent-polyQ interaction may be a possible therapeutic strategy for treating polyQ diseases.

References

1. Yazawa I, Nukina N, Hashida H, Goto J, Yamada M, Kanazawa I. Abnormal gene product identified in hereditary dentatorubral-pallidoluysian atrophy (DRPLA) brain. Nat Genet. 1995;10(1):99-103.

2. La Spada AR, Wilson EM, Lubahn DB, Harding AE, Fischbeck KH. Androgen receptor gene mutations in X-linked spinal and bulbar muscular atrophy. Nature. 1991;352(6330):77-9.

3. Kawaguchi Y, Okamoto T, Taniwaki M, Aizawa M, Inoue M, Katayama S et al. CAG expansions in a novel gene for Machado-Joseph disease at chromosome 14q32.1. Nat Genet. 1994;8(3):221-8.

4. David G, Abbas N, Stevanin G, Durr A, Yvert G, Cancel G et al. Cloning of the SCA7 gene reveals a highly unstable CAG repeat expansion. Nat Genet. 1997;17(1):65-70.

5. Orr HT, Chung MY, Banfi S, Kwiatkowski TJ, Jr., Servadio A, Beaudet AL et al. Expansion of an unstable trinucleotide CAG repeat in spinocerebellar ataxia type 1. Nat Genet. 1993;4(3):221-6.

6. Pulst SM, Nechiporuk A, Nechiporuk T, Gispert S, Chen XN, Lopes-Cendes I et al. Moderate expansion of a normally biallelic trinucleotide repeat in spinocerebellar ataxia type 2. Nat Genet. 1996;14(3):269-76.

7. Tobelmann MD, Murphy RM. Location trumps length: polyglutamine-mediated changes in folding and aggregation of a host protein. Biophys J. 2011;100(11):2773-82.

8. Wetzel R. Misfolding and aggregation in Huntington disease and other expanded

polyglutamine repeat diseases. In: Ramirez-Alvarado M, Kelly JW, Dobson CM, editors. Protein misfolding diseases: current and emerging principles and therapies. Hoboken: John Wiley & Sons, Inc.; 2010. p. 305-24.

9. Scarafone N, Pain C, Fratamico A, Gaspard G, Yilmaz N, Filée P et al. Amyloid-like fibril formation by polyQ proteins: a critical balance between the polyQ length and the constraints imposed by the host protein. PLoS One. 2012;7(3):e31253.

10. Robertson AL, Bottomley SP. Towards the treatment of polyglutamine diseases: the modulatory role of protein context. Curr Med Chem. 2010;17(27):3058-68.

11. Shao J, Diamond MI. Polyglutamine diseases: emerging concepts in pathogenesis and therapy. Hum Mol Gen. 2007;16(R2):R115-R23.

12. Zoghbi HY, Orr HT. Pathogenic mechanisms of a polyglutamine-mediated neurodegenerative disease, spinocerebellar ataxia type 1. J Biol Chem. 2009;284(12):7425-9.

13. Michalik A, Van Broeckhoven C. Pathogenesis of polyglutamine disorders: aggregation revisited. Hum Mol Gen. 2003;12 Spec No 2:R173-86.

14. Matilla-Duenas A, Sanchez I, Corral-Juan M, Davalos A, Alvarez R, Latorre P. Cellular and molecular pathways triggering neurodegeneration in the spinocerebellar ataxias. Cerebellum. 2010;9(2):148-66.

15. Matilla-Duenas A, Ashizawa T, Brice A, Magri S, McFarland KN, Pandolfo M et al. Consensus paper: pathological mechanisms underlying neurodegeneration in spinocerebellar ataxias. Cerebellum. 2014;13(2):269-302.

16. Orr HT. Beyond the Qs in the polyglutamine diseases. Genes Dev. 2001;15(8):925-32.

17. Laghaei R, Mousseau N. Spontaneous formation of polyglutamine nanotubes with molecular dynamics simulations. J Chem Phys. 2010;132(16):165102.

18. Nakano M, Watanabe H, Rothstein SM, Tanaka S. Comparative characterization of short monomeric polyglutamine peptides by replica exchange molecular dynamics simulation. J Phys Chem B. 2010;114(20):7056-61.

19. Nakano M, Ebina K, Tanaka S. Study of the aggregation mechanism of polyglutamine peptides using replica exchange molecular dynamics simulations. J Mol Model. 2013;19(4):1627-39.

20. Chiang HL, Chen CJ, Okumura H, Hu CK. Transformation between alpha-helix and beta-sheet structures of one and two polyglutamine peptides in explicit water molecules by replica-exchange molecular dynamics simulations. J Comput Chem. 2014;35(19):1430-7.

21. Zhou ZL, Zhao JH, Liu HL, Wu JW, Liu KT, Chuang CK et al. The possible structural models for polyglutamine aggregation: a molecular dynamics simulations study. J Biomol Struct Dyn. 2011;28(5):743-58.

22. Hayre NR, Singh RR, Cox DL. Sequence-dependent stability test of a left-handed beta-helix motif. Biophys J. 2012;102(6):1443-52.

23. Babin V, Roland C, Sagui C. The alpha-sheet: a missing-in-action secondary structure? Proteins. 2011;79(3):937-46.

24. Miettinen M, Knecht V, Monticelli L, Ignatova Z. Assessing polyglutamine conformation in the nucleating event by molecular dynamics simulations. J Phys Chem B. 2012;116(34):10259-65.

25. Miettinen MS, Monticelli L, Nedumpully-Govindan P, Knecht V, Ignatova Z. Stable polyglutamine dimers can contain beta-hairpins with interdigitated side chains-but not alpha-helices, beta-nanotubes, beta-pseudohelices, or steric zippers. Biophys J. 2014;106(8):1721-8.

26. Buchanan LE, Carr JK, Fluitt AM, Hoganson AJ, Moran SD, de Pablo JJ et al. Structural motif of polyglutamine amyloid fibrils discerned with mixed-isotope infrared spectroscopy. Proc Natl Acad Sci U S A. 2014;111(16):5796-801.

27. Wang Y, Voth GA. Molecular dynamics simulations of polyglutamine aggregation using solvent-free multiscale coarse-grained models. J Phys Chem B. 2010;114(26):8735-43.

28. Deng L, Wang Y, Ou-Yang ZC. Concentration and temperature dependences of polyglutamine aggregation by multiscale coarse-graining molecular dynamics simulations. J Phys Chem B. 2012.

29. Lakhani VV, Ding F, Dokholyan NV. Polyglutamine induced misfolding of huntingtin exon1 is modulated by the flanking sequences. PLoS Comput Biol. 2010;6(4):e1000772.

30. Moradi M, Babin V, Roland C, Sagui C. Are long-range structural correlations behind the aggregration phenomena of polyglutamine diseases? PLoS Comput Biol. 2012;8(4):e1002501.

31. Stork M, Giese A, Kretzschmar HA, Tavan P. Molecular dynamics simulations indicate a possible role of parallel beta-helices in seeded aggregation of poly-Gln. Biophys J. 2005;88(4):2442-51.

32. Friedrichs MS, Eastman P, Vaidyanathan V, Houston M, Legrand S, Beberg AL et al. Accelerating molecular dynamic simulation on graphics processing units. J Comput Chem. 2009;30(6):864-72.

33. Harvey MJ, De Fabritiis G. An implementation of the smooth particle mesh ewald

method on GPU hardware. J Chem Theory Comput. 2009;5(9):2371-7.

34. Götz AW, Williamson MJ, Xu D, Poole D, Le Grand S, Walker RC. Routine microsecond molecular dynamics simulations with AMBER on GPUs. 1. generalized born. J Chem Theory Comput. 2012;8(5):1542-55.

35. Salomon-Ferrer R, Case DA, Walker RC. An overview of the Amber biomolecular simulation package. WIREs Comput Mol Sci. 2013;3(2):198-210.

36. Salomon-Ferrer R, Götz AW, Poole D, Le Grand S, Walker RC. Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. explicit solvent particle mesh ewald. J Chem Theory Comput. 2013;9(9):3878-88.

37. Wetzel R. Physical chemistry of polyglutamine: intriguing tales of a monotonous sequence. J Mol Biol. 2012;421(4-5):466-90.

38. Rub U, Schols L, Paulson H, Auburger G, Kermer P, Jen JC et al. Clinical features, neurogenetics and neuropathology of the polyglutamine spinocerebellar ataxias type 1, 2, 3, 6 and 7. Prog Neurobiol. 2013;104:38-66.

39. Riley BE, Orr HT. Polyglutamine neurodegenerative diseases and regulation of transcription: assembling the puzzle. Genes Dev. 2006;20(16):2183-92.

40. Mohan RD, Abmayr SM, Workman JL. The expanding role for chromatin and transcription in polyglutamine disease. Curr Opin Genet Dev. 2014;26:96-104.

41. Wang X, Vitalis A, Wyczalkowski MA, Pappu RV. Characterizing the conformational ensemble of monomeric polyglutamine. Proteins. 2006;63(2):297-311.

42. Schafmeister CEAF, Ross WS, Vladimir Romanovski. LEaP. University of California, San Francisco. 1995.

43. Hornak V, Abel R, Okur A, Strockbine B, Roitberg A, Simmerling C. Comparison of multiple Amber force fields and development of improved protein backbone parameters. Proteins. 2006;65(3):712-25.

44. Case DA, Berryman JT, Betz RM, Cerutti DS, T.E. Cheatham I, Darden TA et al. AMBER 2015. University of California, San Francisco. 2015.

45. R Development Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing. 2011. http://www.R-project.org/. Accessed 1 Nov 2016.

46. Wickham H. ggplot2: elegant graphics for data analysis. 2nd ed. New York: Springer; 2009.

47. Racine J. gnuplot 4.0: a portable interactive plotting utility. J Appl Econ. 2006;21(1):133-41.

48. Humphrey W, Dalke A, Schulten K. VMD: visual molecular dynamics. J Mol Graph Model. 1996;14(1):33-8.

49. Jahn TR, Radford SE. Folding versus aggregation: polypeptide conformations on competing pathways. Arch Biochem Biophys. 2008;469(1):100-17.

50. Rhys NH, Dougan L. The emerging role of hydrogen bond interactions in polyglutamine structure, stability and association. Soft Matter. 2013;9(8):2359-64.

51. Vitalis A, Wang X, Pappu RV. Quantitative characterization of intrinsic disorder in polyglutamine: insights from analysis based on polymer theories. Biophys J. 2007;93(6):1923-37.

52. Tsolis AC, Papandreou NC, Iconomidou VA, Hamodrakas SJ. A Consensus Method for the prediction of 'aggregation-prone' peptides in globular proteins. PLoS One. 2013;8(1):e54175.

53. Ignatova Z, Gierasch LM. Extended polyglutamine tracts cause aggregation and structural perturbation of an adjacent beta barrel protein. J Biol Chem. 2006;281(18):12959-67.

54. Chan HS, Dill KA. Polymer principles in protein structure and stability. Annu Rev Biophys Biophys Chem. 1991;20:447-90.

55. Mishra R, Thakur AK. Amyloid nanospheres from polyglutamine rich peptides: assemblage through an intermolecular salt bridge interaction. Org Biomol Chem. 2015;13(14):4155-9.

56. Wen J, Scoles DR, Facelli JC. Effects of the enlargement of polyglutamine segments on the structure and folding of ataxin-2 and ataxin-3 proteins. J Biomol Struct Dyn. 2016:1-16.

57. Williamson TE, Vitalis A, Crick SL, Pappu RV. Modulation of polyglutamine conformations and dimer formation by the N-terminus of huntingtin. J Mol Biol. 2010;396(5):1295-309.

58. Almeida B, Fernandes S, Abreu IA, Macedo-Ribeiro S. Trinucleotide repeats: a structural perspective. Front Neurol. 2013;4.

CHAPTER 5

HYDROGEN BONDING AND WATER ACCESSIBILITY OF THE POLYGLUTAMINE REGIONS IN ATAXIN-2 AND ATAXIN-3 AS FUNCTION OF THE POLYQ TRACT LENGTH

Abstract

The hydrogen bondin and solvent accessibility of the polyglutamine (polyQ) chain may play an important role in polyQ protein aggregation, which has been postulated as a key factor in the pathogenesis pathway of at least ten neurodegenerative diseases. In a recent paper, using molecular dynamics simulations of isolated polyQ segments in explicit solvent, our previous study predicts an increase in aggregation propensity, due to increasing intra-polyQ hydrogen bonding capacity, as the length of the polyQ chain increases. Here, this study reports the hydrogen bonding and solvent accessibility changes of polyQ segments of increasing length, within the sequence-context of full-length ataxin-2 and ataxin-3 3D structures predicted by I-TASSER. The results of this study show that, as the length of the polyQ region increases, there is a corresponding increase of intra-polyQ hydrogen bonding propensity. These results are consistent with our previous molecular dynamics simulations and support the hypothesis that polyQ hydrogen bonding self-interactions increase with the lengths of polyQ chains, suggesting this increased self-interaction capacity as a relevant cause of polyQ enlargement pathogenesis. These results also suggest that the modulation of solvent-polyQ interactions may be a possible strategy

for the treatment of polyQ diseases.

Introduction

The unstable expansion of the polyglutamine (polyQ) tract is associated with at least ten neurodegenerative diseases, including Huntington's disease [1], dentatorubral-pallidoluysian atrophy (DRPLA) [2], spinal and bulbar muscular atrophy (SBMA) [3], and seven spinocerebellar ataxias (SCA) type 1, 2, 3, 6, 7, 8, and 17 [4-7], all together, known as polyglutamine diseases [8]. The proteins underlying these diseases, termed as polyQ proteins, vary in sequence and function, whereas polyQ tract is the only common sequence element shared by the ten polyQ proteins [9]. The protein conformational misfolding and formation of intracellular aggregation are the hallmarks of polyQ diseases [10].

Solvent accessibility and hydrogen bonding are important factors for determining structural stability and aggregation in polyQ peptides. In a recent paper, using molecular dynamics (MD) simulations of isolated polyQ segments in explicit solvent, our previous study indicates that, as the length of the polyQ chain increases, there is an increase in aggregation propensity in polyQ tracts, due to increasing intra-polyQ hydrogen bonding capacity [Chapter 4]. The hydrogen bonding trend, as function of polyQ length, is also shown in other simulation studies of individual polyQ segments [11, 12]. Yet, very little is known about how these trends of the hydrogen bonding and solvent accessibility, in isolated polyQ segments, translate to the situation where the polyQ segment is embedded within a polyQ protein. The hydrogen bonding and solvent accessibility capacity of polyQ within the polyQ protein is a particularly important issue, as our previous computational studies show that the length of the polyQ tract can affect, not only the local, but also the global conformation of ataxin-2 and ataxin-3 [13].

Here, this study reports the changes in hydrogen bonding and solvent accessibility of polyQ segments with increasing polyQ lengths, within the sequence-context of full-length ataxin-2 and ataxin-3 proteins, using 3D structures predicted by I-TASSER [13]. Ataxin-2 and ataxin-3 are selected for this study because they are two of the best characterized polyQ containing proteins [14, 15], responsible for SCA2 and SCA3, respectively.

Methods

Protein 3D Structure Prediction

An ensemble of ataxin-2 and ataxin-3 structures, predicted using the I-TASSER protein structure prediction package [16], are used for the analysis presented here. Both ataxin-2 and ataxin-3 with polyQ regions of varying length are studied. For ataxin-2, the number of repeats of the polyQ segments considered here are 13, 22, and 31 as normal repeats, and 32, 37, and 79 as pathogenic ones [17]. For ataxin-3, the normal repeat lengths considered are 27 and 36, the intermediate ones are 48 and 56, and the pathogenic ones are 64 and 75 [18]. Sequences with a given number of glutamine repeats are named P-nQ, where P is either ataxin-2 or ataxin-3 and n is the number of repeats. The details of the structure prediction procedure and the predicted structures can be found in our previous publication [13].

Hydrogen Bond Determination

Hydrogen bond (HB) is inferred using UCFS Chimera [19] FindHBond module with a hydrogen donor-acceptor distance cutoff of 3.5 Å. Several types of HBs are identified here: HBs formed by residues within the polyQ region are defined as intra-polyQ HBs, whereas HBs formed between glutamines in the polyQ region and residues outside the polyQ region are called inter-polyQ HBs. The intra-polyQ HB can be further classified as backbone-backbone HB, backbone-sidechain HB and sidechain-sidechain HB according to the amide groups in glutamine which participate in the hydrogen bond.

Solvent Accessible Surface Area

Solvent-accessible surface area (SASA) is the surface area of a biomolecule that is accessible to a solvent, and in this study it is used as a parameter to quantify the solvent accessibility of the polyQ peptide. The Dictionary of Protein Secondary Structure (DSSP) program [20] is used to calculate the absolute SASA of each residue in the predicted structures. The commonly accepted precutoff of 20 Å^2 is used to determine whether the peptide/protein regions are on the protein surface or not [21].

Statistical Tests

The Pearson product-moment correlation coefficient test is used with the significant level set at 0.05 in this study. All the statistic computing and graphics are generated using R [22].

Results

The polyQ regions in both ataxin-2 and ataxin-3 can form both intra-polyQ and inter-polyQ HBs regardless of the number of repeats in the segment. As the number of glutamine repeats increases, the total number of HBs in the polyQ region also increases in both ataxin-2 and ataxin-3 (Figure 5.1 (a) and (b)). However, when the total count of HBs is normalized by the number of glutamines in the polyQ region, this normalized HB count in ataxin-2 slightly decreases with the number of repeats (Figure 5.1 (c)), whereas in atatxin-3, the normalized count of HB does not show any clear trend as the number of repeats increases

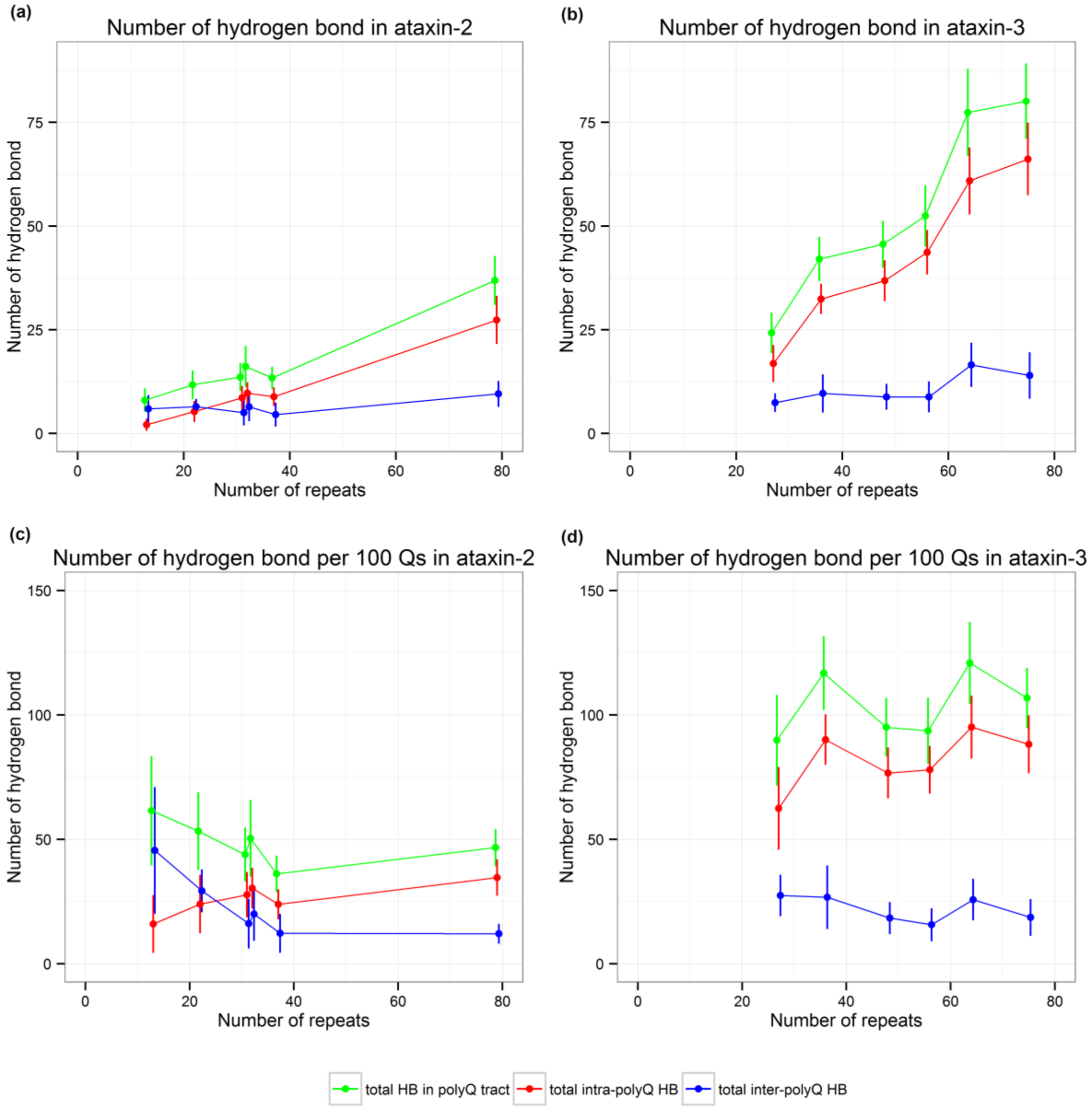


Figure 5.1 Number of hydrogen bond in polyQ region. (a) absolute count of hydrogen bond in ataxin-2; (b) absolute count of hydrogen bond in ataxin-3; (c) normalized count of hydrogen bond in ataxin-2; (d) normalized count of hydrogen bond in ataxin-3. Red: total number of hydrogen bond in polyQ region; green: number of intra-polyQ hydrogen bond; and blue: number of inter-polyQ hydrogen bond. All the data is shown in mean value with standard deviation. The lines are for display only.

(Figure 5.1 (d)). The increase of the count of intra-polyQ hydrogen bond is responsible for almost all the increase observed in the total number of HBs, with the inter-polyQ HBs count remaining constant. When normalized by the number of repeats in the polyQ region, the normalized count of intra-polyQ HBs increases with the number of repeats in both ataxin-2 and ataxin-3 (Figure 5.1 (c) and (d)), whereas the normalized count of inter-polyQ HBs decreases as the number of repeats increases in both ataxin-2 and ataxin-3 (Figure 5.1 (c) and (d)).

The count of sidechain intra-polyQ HBs are similar in ataxin-2 and ataxin-3 for polyQ tracts of similar sizes (Figure 5.2). Considering ataxin-2 and ataxin-3 together, when normalized - the count of sidechain HBs is divided by the number of glutamines of the polyQ region - the number of side-chain HBs per 100 glutamine increases with the number of repeats (Figure 5.2 (b), R^2=0.425, p<0.001). A further analysis is conducted on the origins of the increase of the intra-polyQ HB, and results reveal that all considered categories, including backbone-backbone, backbone-sidechain, and sidechain-sidechain, contribute equally to the overall propensity of the polyQ segment to form intra-polyQ HBs, as the length of the segment increases. These results are a clear indication that as the number of repeats increases, the compactness of the polyQ segment increases.

The absolute SASA of polyQ regions increases with the number of glutamines in the polyQ region; when considering all the data from ataxin-2 and ataxin-3 there is a correlation between the absolute SASA and the number of glutamines in the polyQ region (R^2=0.83, p<0.001, (Figure 5.3 (a)). However, when the SASA of polyQ regions is normalized by the number of repeats, the normalized SASA per glutamine in ataxin-2 is almost independent of the number of repeats (Figure 5.3 (b) red), while for ataxin-3, the

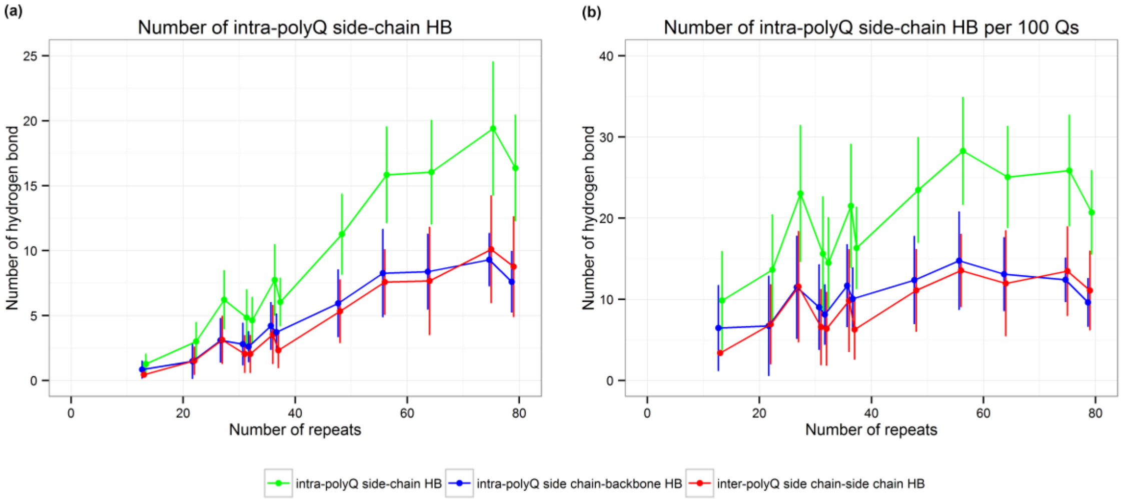


Figure 5.2 The number of sidechain intra-polyQ hydrogen bond. (a) absolute count of sidechain hydrogen bond; (b) normalized count of sidechain hydrogen bond. Blue: total sidechain hydrogen bond; red: sidechain-backbone hydrogen bond; green: sidechain-sidechain hydrogen bond. All data are shown with mean value and standard deviation. The lines are used for display only. Data of ataxin-2 and ataxin-3 are combined together and repeat number of 13, 22, 27, 31, 32, 36, 37, 48, 56, 64, 75, and 79 are plotted.

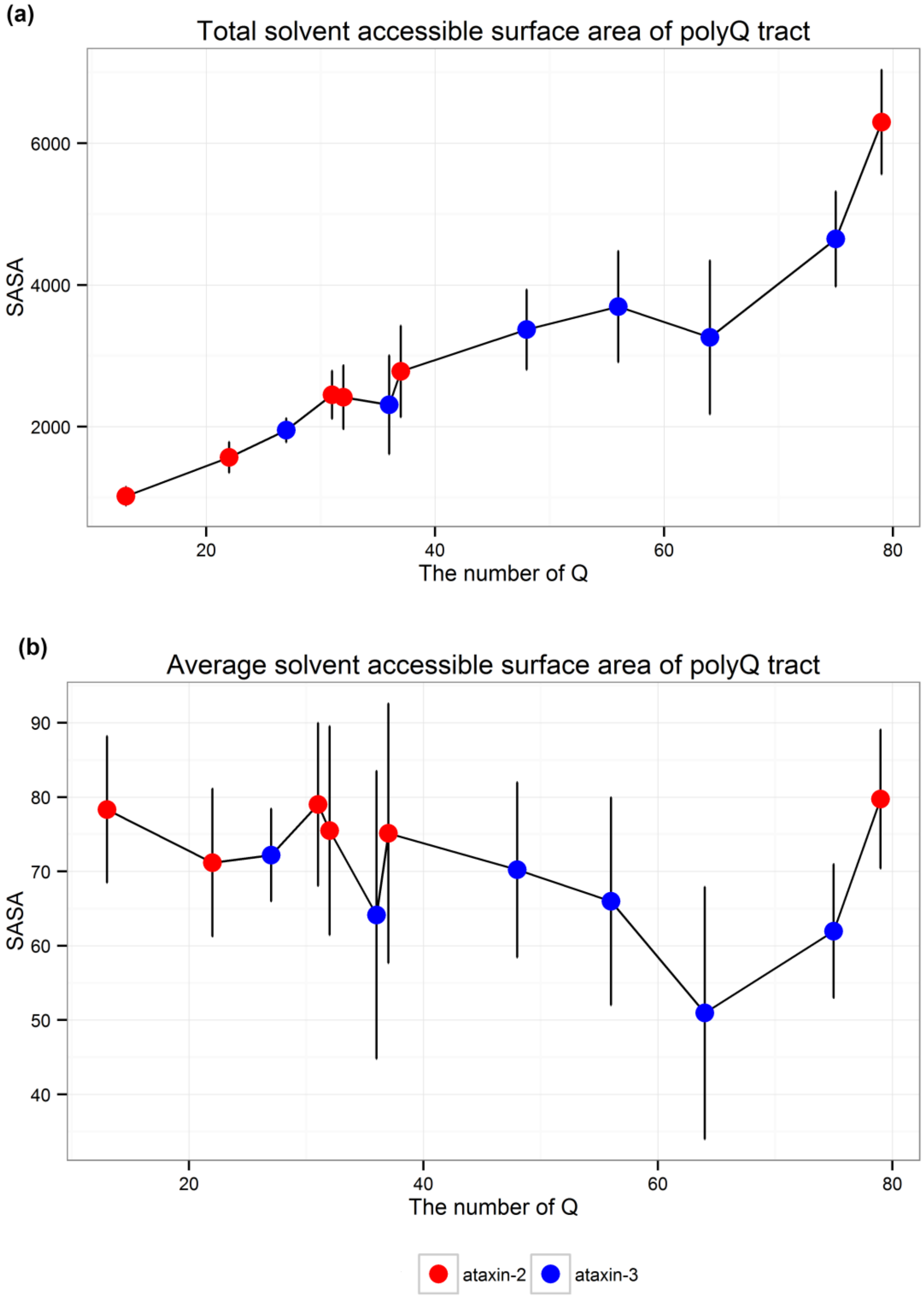


Figure 5.3 Solvent accessible surface area of polyQ regions in ataxin-2 and ataxin-3. (a) total solvent accessible surface area of polyQ regions of ataxin-2 and ataxin-3; (b) normalized solvent accessible surface area of ataxin-2 and ataxin-3. Red: ataxin-2; Blue: ataxin-3.

SASA per glutamine decreases slightly with the number of repeats (Figure 5.3 (b) blue).

When using a precutoff of 20 Å2 per-residue [21] to estimate the SASA for two water molecules, the number of residues in the polyQ region with a SASA value greater than 20 Å2 shows the same trends as those for the total SASA in the polyQ tract (data not shown). Therefore, the calculated SASA properties consistently show that the relative area accessible to the solvent cannot increase as fast as the size of the polyQ segment increases. These results indicate the increasing compactness of the polyQ segment with increasing segment length and is consistent with the observation of the increased tendency to form intra-polyQ HBs, found in this study and in our previous study.

Discussion

Hydrogen bonding and solvent-protein interactions are key factors determining protein folding, stability, and aggregation. In our previous MD simulation study, the HBs and solvent-polyQ interaction of a polyQ monomer are studied in an explicit solvent environment. Our previous study finds that as the polyQ length increases, the polyQ tracts increasingly prefer self-interactions than solvent-polyQ interactions [Chapter 4]. The aforementioned study was restricted to single polyQ segment, therefore question arises, how transferable are the MD results to situations in which the polyQ segment is embedded in a protein? Unfortunately, MD simulations of entire polyQ disease proteins in explicit solvent are unfeasible, therefore this study analyzes the HB and solvent accessibility patterns of the 3D structures of polyQ segments within full-length ataxin-2 and ataxin-3, from our previous work [13].

In all 3D structures analyzed here, the total number of HBs observed in the polyQ segments increases with the polyQ segment lengths. Regardless of the host protein, the normalized count of intra-polyQ HBs increases with the polyQ length, whereas the normalized count of inter-polyQ HBs decreases with the polyQ length. This HB trends are consistent with our MD simulation findings, showing that polyQ regions increasingly prefer self-interactions as the length of the polyQ segment increases. The solvent accessibility properties of the polyQ regions also share common trends, regardless of the host proteins. Although the total SASA of polyQ regions increases with the polyQ lengths, its normalized values decrease with the polyQ lengths. The SASA properties are consistent with both the HB results and SASA results of our previous MD simulations. These results strongly suggest that the increase of the length of the polyQ segments leads to an increase in the compactness of these segments, which increased compactness is independent of protein environment. Moreover, as the structures of the polyQ regions are different for ataxin-2 and ataxin-3 – random coil in ataxin-2 and α-helix in ataxin-3 [13] – the increased preference to adopt intra-polyQ HBs upon polyQ length enlargement appears to be an intrinsic properties of polyQ, which increase does not depend on secondary structure.

Evidence shows that most of the aggregation-prone structures of polyQ peptides are antiparallel β-strands and β-helical structures, which are stabilized by an extensive network of HBs between glutamines in the polyQ regions [23-26]. Therefore an increase of intra-polyQ HBs will increase the probability of the formation of aggregation-prone structures in the polyQ segments.

Using H/D exchange technology, Natalello et al. find that polyQ sidechain HBs are strongly involved in the formation of mature amyloid aggregates in ataxin-3-55Q, and the formation of sidechain HBs is the hallmark of mature fibrils generated by ataxin-3 with expansion polyQ [27]. As shown in this study, the normalized count of sidechain HBs per glutamine in the polyQ segments is positively correlated with the number of repeats in the polyQ segments. This correlation provides additional support for our argument that the increase of the polyQ length is conducive to polyQ aggregation. Wang et al. indicate [12] that glutamine sidechains can compete with water in forming HBs with the glutamine backbone, therefore an increase of the propensity of forming sidechain HBs with the backbone, as observed here, could decrease the ability to solvate the backbone and contribute to aggregation.

Conclusions

Using predicted 3D structures of ataxin-2 and ataxin-3 with polyQ segments of increasing length, we are able to study the changes in hydrogen bonding and solvent accessibility of polyQ regions within the sequence context of two well characterized polyQ proteins. Our results strongly suggest that the HB and solvent accessibility properties of the polyQ regions have an intrinsic dependency on the length of the polyQ segment, regardless of the protein host. The results presented here are consistent with previous studies, including our own molecular dynamics (MD) study on polyQ monomers in explicit solvent, and indicate that polyQ regions increasingly prefer self-interactions, which consistently can lead to more compact polyQ structures. The results also strongly support the notion that the enlargement of the polyQ regions can be the intrinsic force leading to self-aggregation of polyQ proteins, making polyQ aggregation the leading mechanisms in the pathogenesis pathway of polyQ diseases. The results of this study suggest the modulation of solvent-polyQ interaction as a possible therapeutic strategy for the treatment of polyQ diseases.

References

1. Group THsDcR. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. Cell. 1993;72(6):971-83.

2. Yazawa I, Nukina N, Hashida H, Goto J, Yamada M, Kanazawa I. Abnormal gene product identified in hereditary dentatorubral-pallidoluysian atrophy (DRPLA) brain. Nat Genet. 1995;10(1):99-103.

3. La Spada AR, Wilson EM, Lubahn DB, Harding AE, Fischbeck KH. Androgen receptor gene mutations in X-linked spinal and bulbar muscular atrophy. Nature. 1991;352(6330):77-9.

4. Kawaguchi Y, Okamoto T, Taniwaki M, Aizawa M, Inoue M, Katayama S et al. CAG expansions in a novel gene for Machado-Joseph disease at chromosome 14q32.1. Nat Genet. 1994;8(3):221-8.

5. David G, Abbas N, Stevanin G, Durr A, Yvert G, Cancel G et al. Cloning of the SCA7 gene reveals a highly unstable CAG repeat expansion. Nat Genet. 1997;17(1):65-70.

6. Orr HT, Chung MY, Banfi S, Kwiatkowski TJ, Jr., Servadio A, Beaudet AL et al. Expansion of an unstable trinucleotide CAG repeat in spinocerebellar ataxia type 1. Nat Genet. 1993;4(3):221-6.

7. Pulst SM, Nechiporuk A, Nechiporuk T, Gispert S, Chen XN, Lopes-Cendes I et al. Moderate expansion of a normally biallelic trinucleotide repeat in spinocerebellar ataxia type 2. Nat Genet. 1996;14(3):269-76.

8. Fan HC, Ho LI, Chi CS, Chen SJ, Peng GS, Chan TM et al. Polyglutamine (PolyQ) diseases: genetics to treatments. Cell Transplant. 2014;23(4-5):441-58.

9. Polling S, Ormsby AR, Wood RJ, Lee K, Shoubridge C, Hughes JN et al. Polyalanine expansions drive a shift into alpha-helical clusters without amyloid-fibril formation. Nat Struct Mol Biol. 2015;22(12):1008-15.

10. Sun CS, Lee CC, Li YN, Yao-Chen Yang S, Lin CH, Chang YC et al. Conformational switch of polyglutamine-expanded huntingtin into benign aggregates leads to neuroprotective effect. Sci Rep. 2015;5:14992.

11. Vitalis A, Wang X, Pappu RV. Atomistic simulations of the effects of polyglutamine chain length and solvent quality on conformational equilibria and spontaneous homodimerization. J Mol Biol. 2008;384(1):279-97.

12. Wang X, Vitalis A, Wyczalkowski MA, Pappu RV. Characterizing the conformational ensemble of monomeric polyglutamine. Proteins. 2006;63(2):297-311.

13. Wen J, Scoles DR, Facelli JC. Effects of the enlargement of polyglutamine

segments on the structure and folding of ataxin-2 and ataxin-3 proteins. J Biomol Struct Dyn. 2016:1-16.

14. Masino L, Nicastro G, De Simone A, Calder L, Molloy J, Pastore A. The Josephin domain determines the morphological and mechanical properties of ataxin-3 fibrils. Biophys J. 2011;100(8):2033-42.

15. Albrecht M, Golatta M, Wullner U, Lengauer T. Structural and functional analysis of ataxin-2 and ataxin-3. Eur J Biochem. 2004;271(15):3155-70.

16. Roy A, Kucukural A, Zhang Y. I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protocols. 2010;5(4):725-38.

17. Magana JJ, Velazquez-Perez L, Cisneros B. Spinocerebellar ataxia type 2: clinical presentation, molecular mechanisms, and therapeutic perspectives. Mol Neurobiol. 2013;47(1):90-104.

18. Costa Mdo C, Paulson HL. Toward understanding Machado-Joseph disease. Prog Neurobiol. 2012;97(2):239-57.

19. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC et al. UCSF Chimera--a visualization system for exploratory research and analysis. J Comput Chem. 2004;25(13):1605-12.

20. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983;22(12):2577-637.

21. Tsolis AC, Papandreou NC, Iconomidou VA, Hamodrakas SJ. A consensus method for the prediction of ‘aggregation-prone’ peptides in globular proteins. PLoS One. 2013;8(1):e54175.

22. R Development Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing. 2011. http://www.R-project.org/. Accessed 1 Nov 2016.

23. Perutz MF. Glutamine repeats and neurodegenerative diseases: molecular aspects. Trends Biochem Sci. 1999;24(2):58-63.

24. Perutz MF, Finch JT, Berriman J, Lesk A. Amyloid fibers are water-filled nanotubes. Proc Natl Acad Sci U S A. 2002;99(8):5591-5.

25. Ogawa H, Nakano M, Watanabe H, Starikov EB, Rothstein SM, Tanaka S. Molecular dynamics simulation study on the structural stabilities of polyglutamine peptides. Comput Biol Chem. 2008;32(2):102-10.

26. Zhou ZL, Zhao JH, Liu HL, Wu JW, Liu KT, Chuang CK et al. The possible structural models for polyglutamine aggregation: a molecular dynamics simulations study. J Biomol Struct Dyn. 2011;28(5):743-58.

27. Natalello A, Frana AM, Relini A, Apicella A, Invernizzi G, Casari C et al. A major role for side-chain polyglutamine hydrogen bonding in irreversible ataxin-3 aggregation. PLoS One. 2011;6(4):e18789.

CHAPTER 6

DISCUSSION AND CONCLUSION

Summary of Important Findings

By using computational and informatics methods, this dissertation is able to study the structural and folding characteristics of both polyQ proteins and polyQ monomers. The results of this dissertation bring insights into the molecular mechanisms of polyQ aggregation and the pathogenesis of polyQ diseases.

Protein 3D structure prediction is a powerful tool for biomedical research. However, it is necessary to evaluate the prediction performance of these methods on polyQ proteins before large scale predictions are done. In Chapter 2, two of the best-performing protein 3D structure prediction programs, I-TASSER [1] and Rosetta [2], are tested on their prediction performance of the N-terminal huntingtin protein with 17 glutamines (HTT-17Q). Both I-TASSER and Rosetta perform well at predicting the structure of HTT-17Q determined by comparing the predicted models with the X-ray crystal structures of HTT-17Q [3]. Yet I-TASSER performs better than Rosetta, as the I-TASSER models show better agreement on the overall structures of the X-ray structures of HTT-17Q than Rosetta dose and I-TASSER models also better represent the diversity of secondary structures in the polyQ regions.

As a next step, I-TASSER is used to predict the 3D structures of ataxin-2 and ataxin-3 proteins, two of the best representative polyQ proteins. Ataxin-2 and ataxin-3 with polyQ tracts in both normal and pathogenic ranges are studied. In Chapter 3, the structure and conformation changes are described as the function of the length of polyQ tract, while the water accessibility and hydrogen bonding changes are discussed in Chapter 5.

I-TASSER models of ataxin-2 and ataxin-3 indicate that the enlargement of the polyQ tract can affect not only the local structures of polyQ tracts but also the global structures of polyQ proteins. As the lengths of polyQ tracts increase, the proportion of helical structures increases in both ataxin-2 and ataxin-3. The elongation of the polyQ tract can also affect the structure of functional domains in ataxin-2 and ataxin-3 [4]. For example, the UIM1-2 domains in ataxin-3 adopt helix-coil-helix structures, which are required for normal functions [5]. However, the flexible structures are lost in some of the ataxin-3 models with pathogenic polyQ tracts. These results indicate that the polyQ tract can lead to the dysfunction of polyQ proteins by changing the structures of their functional domains. Also, the changes in both local and global structures suggest that the study of full-length polyQ proteins is necessary to get a comprehensive understanding of the pathogenic mechanisms.

In addition to polyQ proteins, the folding dynamics and the solvent-polyQ interactions of polyQ monomers are studied using molecular dynamics (MD) simulation (Chapter 4). PolyQ lengths in the normal, intermediate, and pathogenic ranges are studied. Using GPU-acceleration, it is possible to conduct 105-ns all-atom MD simulations in an explicit solvent environment within a reasonable time frame. The MD simulation results show that, as the lengths of the polyQ segments increase, the polyQ tracts increasingly prefer adopting intra-polyQ hydrogen bonds than adopting solvent-polyQ hydrogen bonds. Long-distance interactions between residues far apart in polyQ are more predominant in long-polyQ segments, and this predominance leads to significant conformational differences. This study demonstrates gains in aggregation propensity with increased polyQ length, and suggests that the modulation of solvent-polyQ interaction may be a possible therapeutic strategy for the treatment of polyQ diseases.

Using the full-length 3D structure models of ataxin-2 and ataxin-3 predicted by I-TASSER, Chapter 5 explores the hydrogen bonding and water accessibility of polyQ tracts within the context of the polyQ proteins, similar results are get as those shown in polyQ monomers using MD simulations (Chapter 4). The results, from the full-length ataxin-2 and ataxin-3 structures, show an increased preference in adopting intra-polyQ hydrogen bonds as the length of polyQ region increases, whereas the preference to forming inter-polyQ hydrogen bonds decreases with the length of the polyQ tract. The absolute SASA values of polyQ regions increase with the number of glutamine repeats in ataxin-2 and ataxin-3, but the normalized SASA values decrease with the polyQ length. These results again indicate that, as the length of polyQ increases, the polyQ self-interaction increases and the propensity of aggregation increases.

As a protein involved in RNA processing pathway [6], ataxin-2 is a component of stress granules [7] which are dense aggregations composed of proteins and RNAs [8]. The stress granules appear when the cell is under stress and it is a mechanism to protect RNAs from reacting with harmful compounds. Stress granules exhibit non-membrane liquid properties [8]. Recent studies show that the low complexity domains in the stress-granular proteins play a role in inducing a liquid phase transition through inter-molecule interactions [9, 10]. The predicted 3D structures of ataxin-2 proteins show low complexity, with a structural dominance of random coils in this study [4], therefore these flexible conformations could be an important structural factor for the normal structures and molecule interaction in stress granules. The structure and solvent accessibility changes in ataxin-2, induced by the enlargement of polyQ tract, might also affect the structure and normal function of stress granules in several possible ways. The structural changes of the functional domains in ataxin-2, induced by the enlargement of polyQ tract, might also change the proteome in the stress granules, by altering the protein-protein interactions. This hypothesis is supported by the results of Jain et al. which show that ATPase-modulated stress granules contain a diverse proteome [11]. The changes of proteome in the stress granules might alter the dynamic equilibrium of the ribonucleoprotein in the cell. Jain et al. also find that stress granules contain a stable substructure core [11]. The changes of the interaction partners might also change the stability and conformation of the core. Also the increase of self-interaction in polyQ regions might affect the phase transition of stress granules between monomer, liquid droplet, and hydrogel states. This is also consistent with the results of Murakami et al. that found that mutations in low-complexity domain of FUS gene, a gene related with neurodegenerative disease ALS, affect the phase transition of stress granules and accelerate transition into fibrillary hydrogens, which will cause neurodegeneration [12]. As the increase of self-interaction in polyQ regions could lead to more compact structures of polyQ and induce self-aggregation, pathogenic polyQ segments might also induce further phase transitions into irreversible fibrillary structures. All these possible changes in the function and structure in stress granules might affect the RNA and protein microenvironment of the cells and might impair granule functions. Further experimental and coarse grain studies are needed to explore the proteome and structures of the stress granules in which ataxin-2 with different polyQ lengths are involved.

The results in structure, hydrogen bonding, and water accessibility together can help us better understand the polyQ length-dependent aggregation pathogenic mechanisms, and suggest possible treatment strategies for polyQ diseases.

Contribution to Biomedical Informatics

This study provides an effective informatics solution and workflow to study the pathogenic mechanism of polyQ diseases at the protein level, and proves that computational and informatics methods are powerful and effective tools to study polyQ diseases. Several informatics methods are created or modified from the existing methods, providing better analysis and visualization of the results. For example, the WebLogo [13], which was originally created for gene/protein sequence comparison, is used as a tool to show secondary structure patterns in this study, and provides a powerful tool for secondary structure visualization and comparison. Also the structure similarity scores are created for a quantitative comparison of the structure similarity in the polyQ regions [3].

The computing time and efficiency are critical factors that limit the application of computational work in the biomedical field. In this study, some state-of-the-art high performance computing technology, such as parallel computing and GPU acceleration, are applied to increase the computational efficiency, which makes it possible to generate all the structural data within a reasonable time frame.

The workflow provided by this dissertation can be easily generalized to study other protein conformation related diseases, including Parkinson’s disease, Alzheimer's disease, and even cancer. At this point, the workflow is successfully applied to study a rare nonsense mutation in cutaneous malignant melanoma and the computational explanation of the pathogenesis of this mutation is consistent with results from experimental studies (results not published). This workflow is also used to annotate oncogene mutations in COSMIC database, and exciting results are found.

References

1. Roy A, Kucukural A, Zhang Y. I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protocols. 2010;5(4):725-38.

2. Leaver-Fay A, Tyka M, Lewis SM, Lange OF, Thompson J, Jacak R et al. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. Methods Enzymol. 2011;487:545-74.

3. Wen J, Scoles D, Facelli J. Structure prediction of polyglutamine disease proteins: comparison of methods. BMC Bioinformatics. 2014;15(Suppl 7):S11.

4. Wen J, Scoles DR, Facelli JC. Effects of the enlargement of polyglutamine segments on the structure and folding of ataxin-2 and ataxin-3 proteins. J Biomol Struct Dyn. 2016:1-16.

5. Song A-X, Zhou C-J, Peng Y, Gao X-C, Zhou Z-R, Fu Q-S et al. Structural transformation of the tandem ubiquitin-interacting motifs in ataxin-3 and their cooperative interactions with ubiquitin chains. PLoS One. 2010;5(10):e13202.

6. Nonhoff U, Ralser M, Welzel F, Piccini I, Balzereit D, Yaspo M-L et al. Ataxin-2 interacts with the DEAD/H-Box RNA helicase DDX6 and interferes with p-bodies and stress granules. Mol Biol Cell. 2007;18(4):1385-96.

7. Ralser M, Albrecht M, Nonhoff U, Lengauer T, Lehrach H, Krobitsch S. An integrative approach to gain insights into the cellular function of human ataxin-2. J Mol Biol. 2005;346(1):203-14.

8. Protter DSW, Parker R. Principles and properties of stress granules. Trends Cell Biol. 2016;26(9):668-79.

9. Shin J, Salameh JS, Richter JD. Impaired neurodevelopment by the low complexity domain of CPEB4 reveals a convergent pathway with neurodegeneration. Sci Rep. 2016;6:29395.

10. Conicella AE, Zerze GH, Mittal J, Fawzi NL. ALS mutations disrupt phase separation mediated by alpha-helical structure in the TDP-43 low-complexity C-terminal domain. Structure. 2016;24(9):1537-49.

11. Jain S, Wheeler Joshua R, Walters Robert W, Agrawal A, Barsic A, Parker R. ATPase-modulated stress granules contain a diverse proteome and substructure. Cell. 2016;164(3):487-98.

12. Murakami T, Qamar S, Lin Julie Q, Schierle Gabriele SK, Rees E, Miyashita A et al. ALS/FTD mutation-induced phase transition of FUS liquid droplets and reversible

hydrogels into irreversible hydrogels impairs RNP granule function. Neuron. 2015;88(4):678-90.

13. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. Genome Res. 2004;14(6):1188-90.