Structural aspects of the Huntingtin protein investigated by biocomputing methods

Von der Fakultät für Mathematik, Informatik und Naturwissenschaften der RWTH Aachen University zur Erlangung des akademischen Grades einer Doktorin der Naturwissenschaften genehmigte Dissertation vorgelegt von

Giulia Rossetti, PhD

aus Rom, Italien

Berichter:	Univ.-Prof. Dr. Paolo Carloni
	Univ.-Prof. Dr.rer.nat. Marc Spehr
Tag der mündlichen Prüfung:	17.05.2011

1. Reviewer: Prof. Paolo Carloni

2. Reviewer: Prof. Marc Spehr

Abstract

Huntingtons disease (HD) is a neurodegenerative disorder which lead to death within a couple of decades. It causes uncontrolled movements, loss of intellectual faculties and emotional disturbance. HD is a familial disease, passed from parent to child through a mutation in a specific gene, the HTT gene. This gene provides the genetic information for a protein called "Huntingtin" (Htt). Part of this gene is constituted by a repeated triplet of the nucleotides cytosine-adenine-guanine. This triplet encodes for a particular amino acid, glutamine (whose one letter code symbol is Q). The number of triplets varies between individuals and generations. If the number of triplets consists of 36 or more the gene encodes for an altered form of the protein, called mutant Huntingtin protein (mut-Htt), with an elongated stretch of Qs (PolyQ). PolyQ, within mut-Htt, may interact abnormally with other proteins. It may also form, in a complex process called misfolding and aggregation, inclusion bodies (aggregates) within cells. The process can be affected rather dramatically by PolyQ's neighboring (or flanking) regions, in particular by the seventeen amino acids preceding the PolyQ, called N17. These regions may also modulate aggregation by interacting with cellular partners.

The resulting aggregates may interfere with the transmission of information within neurons leading to brain damage.

Here I have used computational methods to unravel key aspects of the PolyQ stability (i) (which may in turn explain their formation) and of the interactions between N17 and one of its putative cellular partners (ii).

(i) By using classical molecular dynamics, I have shown that key factors contributing to PolyQ stability depend mainly on their ability to create well organized net of tight intramolecular hydrogen-bond interaction. By using quantum-chemistry methods, I have discovered particular features of those interaction networks that are strictly connected with PolyQ polarity and the shape of their electron density. The latter has never been characterized before for such kind of systems. It may be very important to modulate aggregation properties of the PolyQ in mut-Htt.

(ii) I have used simulation methods to predict the possible shapes (or conformations) that N17 assumes in aqueous solution. The prediction is fully consistent with the available experimental biophysical data. Hence, I used the results obtained from the simulations about N17 conformations to understand how it can interact with possible cellular partners. For this purpose, in the last part of the thesis, I created a model of N17 interaction with F-actin, which was identified experimentally as a possible mut-Htt interactor by Prof. M. Diamond (U. Washington, San Luis, US). F-actin has been observed to affect intracellular aggregation of mut-Htt. Our model has been successfully validated by a variety of in cell experiments performed by Prof. M. Diamond. It is able to give a reasonable explanation for the effect of F-actin on mut-Htt-aggregation. This approach is one of the first investigations employing biocomputing methods, to investigate Htt-interactor binding at the molecular level.

In this thesis, we use several computer-based methods, from atomistic classical and quantum-mechanical simulations to bioinformatics techniques, to investigate a very important disease-linked protein. These approaches successfully allowed me to elucidate some of the crucial facets in HD mechanisms at molecular level. The approaches are generally applicable and hence they are promising tools to clarify issues in other neurodegenerative diseases as well.

Ho visto cose bellissime, grazie alla diversa prospettiva suggerita dalla mia perenne insoddisfazione, e quel che mi consola ancora, é che non smetto di osservare.

Edgard Degas

Acknowledgements

I would like to thank Prof. Paolo Carloni for all these years of fruitful work in his group. I thank him for the shared Science and for all the days spent in constructive discussions. He has always supported me in my professional growth, and he has tirelessly encouraged me in all the steps of my research.

I would like to thank Dr. Alessandra Magistrato and Dr. Annalisa Pastore for all their efforts and their huge help in my work.

I would like to thank Prof. Mueller-Krumbhaar for all his contribution and his great support.

I would like to thanks Prof. Mark Diamond, Prof. Marc Spehr and Prof. Alessandro Laio for their collaboration.

I also thanks all my colleagues, particularly Dr. Emiliano Ippoliti, Dr. Jens Dreyer, Valeria Losasso and Pilar Cossio.

I thanks Elke George and all the administration staff for their daily work, which allow us to concentrate only on Science.

Finally I would like to thanks my family, my friends and all the persons I love. Their support was, is and will always be fundamental.

Contents

List of Figures

LIST OF FIGURES

LIST OF FIGURES

List of Tables

LIST OF TABLES

1

Introduction

Huntingtons disease (HD) is a progressive, fatal, neurodegenerative disorder caused by an expanded CAG repeat in the huntingtin gene (HTT). This encodes an abnormally long polyglutamine tract (PolyQ) in the huntingtin protein (Htt) (132; 195).

HD is inherited in an autosomal dominant manner with age-dependent penetrance: longer Q repeats correlates with an earlier age of onset (112; 259). [1] Clinical features of HD include progressive motor dysfunction, cognitive decline, and psychiatric disturbance (258; 318). They both are likely to be caused by both neuronal dysfunction and neuronal cell death. The prevalence of HD is $4 - 10$ per 100 000 in the western world. Up to date there is no cure for this disease (264).

A seminal contribution to the investigation of HD was given by more than 15 years ago by Mangiarini et al. (196). These authors developed an HD mouse model using transgenic insertion technology (196). They expressed the first exon of mutant human Htt with polyQ expansion (*Mut Htt Exon 1* in Fig. 1.1) as a transgene in the mouse. The resultant mouse lines developed severe disease in as short as 3 weeks, and obvious movement disorders that resembled those of HD, as well as some brain mass as in HD (196). Indeed, *Mut Htt Exon 1* is sufficient to lead to the apoptosis of the cell (25; 55; 150; 206; 262; 297), to cause the disease (196) and to feature the same aggregation properties of the full-length of the mutated protein (mut-Htt) both *in vivo* and *in vitro* (55).

Mut Htt Exon 1 contains the polyQ expansion (1.1). The latter is followed by a polyproline (PolyP) rich region and it is preceded by 17 aminoacids (N17). N17 is

[1] Expansions of 50 and more repeats generally cause the juvenile form of the disease (112).

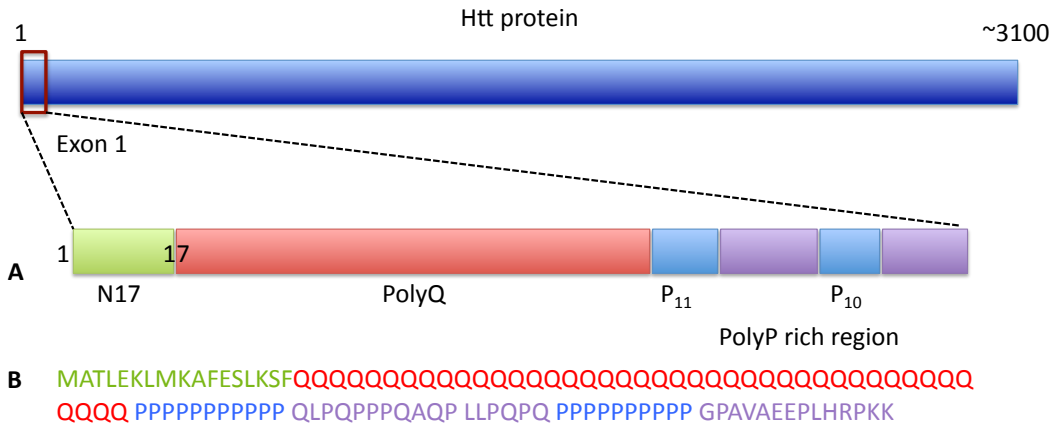


Figure 1.1: ***Htt* and *Exon 1*:** - Scheme of *Exon 1* regions (A) with the corresponding primary sequence (B).

conserved across all vertebrates and modulates the toxicity of mut-Htt (15). Neuronal intra-nuclear and intra-cytoplasmic inclusions of *Mut Htt Exon 1* and other fragments of the mut-Htt, are pathological hallmarks of the disease (81).

In this thesis, we have investigated by computational methods key aspects of the polyQ expansion (i) and N17 (ii).

Our study on the polyQ expansion was prompted by the late, classic work of the Nobel laureate Max Perutz, who was the first to suggest that the structural determinants of PolyQ are insoluble pleated β-sheet-based structures (241). The latter would form tighter interactions upon increasing the PolyQ length, thus explaining the correlation between the latter and the severity of the disease. Perutz (241), based on atomistic models, showed that polyQ can form β-sheets held together by main and side chains atoms in a "polar zipper interaction". Perutz's suggestion was that Q side chains, being similar to the backbone units, are able to establish an extensive hydrogen bond network stabilized in a cooperative mode. Such cooperative effect (CE) causes an extra stabilization of the network with respect to the isolated components. CE on HB has been suggested to be very important for the stability of polyQ chains (158; 241; 245).

Hence, we have investigated the role of the H-bond content, along with the shape of the aggregated structure, for the CE. Then, we have asked ourselves to what extent the CE depends on how the electron density is distributed around the polypeptide structure. Specifically, we have used quantum mechanical approaches to predict the

rearrangement of the electronic density on passing from the isolate Q units to structure of increasing complexity.

(ii) Recently, it has been shown that N17 dramatically affects *Mut Htt Exon 1* (and Mut Htt) aggregation (35; 63; 155; 176; 257; 298; 328). Indeed the deletion of this part brings to a large decrease of *Mut Htt Exon 1* aggregation (298) *in vitro*. In addition, antibodies directed against the N17 (63; 180) decrease the aggregation of Htt protein, whereas antibodies directed against the polyQ increase aggregation (154; 187). The role of N17 for aggregation is currently subject of a very vivid debate. Several hypotheses have been suggested: N17 could nucleate Htt aggregation (136; 298; 299) but may also influence Htt aggregation by affecting mut-Htt cellular localization (15; 306). N17 could also interact with several cellular partners that could in principle affect Htt aggregation (115; 148; 278; 284; 294). In particular, Prof. M. Diamond (Washington University School of Medicine) (10; 250) and others (21; 44; 207; 208; 289) have pointed to a role for actin in Htt aggregation and suggested that actin may bind to N17. In this thesis, we use molecular modeling combined with cellular essays by Prof. Diamond to investigate the role that N17-mediated interactions with actin for regulating Htt aggregation. On the basis of available actin complexes with other binding partner, we have created a structural model of N17 peptide bound on actin surface. On the hypothesis that this bind could decrease Htt aggregation by masking N17 region, we have predicted N17 mutants able to affect the binding. We then test our hypothesis by detecting *Htt Exon 1* aggregation in cell with FRET and fluorescence techniques, in collaboration with prof. Diamond.

2

Huntington's disease and the Huntingtin protein.

2.1 Huntington's Disease.

Huntingtons disease (HD) is a devastating neurodegenerative human disease, for which there is currently no cure. HD is a autosomal dominant [1] disease, characterized by progressive motor, cognitive, and psychiatric symptoms (132).

The gene responsible for HD (HTT) encodes the ubiquitously expressed, large human Huntingtin protein (Htt, molecular weight 348 kDa) (195). The causative mutation is an abnormal expansion of CAG trinucleotide repeats within the coding sequence of the gene. The expansion leads to an elongated stretch of Q residues beyond the first 17 amino acid (N17)(195).

In health individuals, the number of Q repeats is 35 or fewer, with 17-20 repeats found most commonly (215). [2]. Most adult-onset HD cases have 40-50 Qs, featuring a mutated form of the protein (mut-Htt). Expansions of 50 and more repeats are often associated with the juvenile form of the disease (112). There is a strong inverse correlation between the age of onset of HD and the number of Qs (112). The longer polyQ tracts, the earlier age of the onset (112; 259). However, there is a wide variation

[1] Autosomal dominant conditions are achieved in case a mutated gene from one parent is sufficient to cause a disease, in spite of the presence of a normal gene inherited from the other parent.

[2] Repeats between 27 and 35 are rare and are not associated with disease. However they are meiotically unstable. They can expand into the HD range when they are 36 and above. The disease seems not to be fully penentrant in individuals with 36-41 repeats (112)

in the age of onset with a given Q number (137).

Htt is essential for brain development (255), although its exact biological function is unknown. It is located mostly in the cytoplasm and, to a smaller extent, in the nucleus (151). The protein can dynamically travel back and forth between the cytoplasm and the nucleus (151). It may be associated also with the plasma membrane, with the endocytic and autophagic vesicles, with the endosomal compartments, with the endoplasmic reticulum, the Golgi apparatus, mitochondria and microtubules (15; 51; 151; 152; 257; 287). This chapter focuses on molecular and structure aspects of the Htt protein, as well as on some of facets of mut-Htt responsible for HD.

2.2 The Human Huntingtin Protein

Htt is a large, multidomain protein of ($\sim$ 3144 aa) for which structural information at atomic resolution is not avaliable (335). Htt has been proposed to be an elongated super-helical solenoid with a diameter of $\sim$ 200 $\mathring{A}$ (188). We next briefly review key molecular aspects of the protein for which biophysical or biochemical information is available.

Exon 1. The best-characterized part of the protein is the *Htt Exon 1*. It consists of the following regions: N17, the variable polyQ region (less than 36 Q in healthy individuals (195) and a polyP-rich region(296) (Fig. 1). *Htt Exon 1* is sufficient to cause HD-like pathology in animal models (196) and the typical formation of Htt aggregates (20; 75; 196). Hence, investigations of *Htt Exon 1* may help understand key aspects of the disease.

N17 is fully conserved across all vertebrate species. It was originally annotated as unstructured (243). However, mutational analysis *in vivo*, as well as CD(15; 298) and NMR(298) spectroscopy studies *in vitro* have shown that N17 may be an amphipathic α-helix, with membrane-associating properties with regard to the endoplasmic reticulum (15). N17 is particularly susceptible to post-translational modifications, including phosphorylation, ubiquitination, and SUMO attachment (257). In particular, Thr3 phosphorylation influences toxicity(3). Ser 13 and Ser 16 phosphorylation has mostly protective effects in vivo (109).

PolyQ is a key regulator of Htt binding to its cellular partners (115; 241) (8) (155; 335). Its binding capability might be due to its flexible and multifunctional structures able to assume specific conformations and activities depending on its binding partners, sub-cellular location, and time of maturation in a given cell type and tissue (155; 335).[1]

PolyP may stabilize of the polyQ tract by keeping it soluble (35). It may work also as a protein-interaction domain (306). Consistently with these hypotheses, structural data provided hints that the polyQ repeat at the N-terminal is influenced by the COOH-terminal polyP region (35; 155).

HEAT repeats. HEAT repeats are ∼40-amino-acid domains, which fold in two anti-parallel α-helices forming a hairpin (229). Htt features 16 of these repeats (8; 188; 229), organized in 4 clusters (296). HEAT may be involved in protein-protein interactions (9; 217; 240).

Caspase and Calpain Cleavage Sites. These are sites recognized by proteolytic enzymes (including caspases 1, 3, 6, 7 and 8, calpain and an unidentified aspartic protease). As a result, a wide range of fragments are generated (98; 118; 157; 194; 325). The role of Htt proteolysis for its physiological function is not fully elucidated. However, full-length mutHtt is less toxic than its N-terminal fragments (99; 106).

2.3 Mutated Huntingtin

2.3.1 General features

Mut-Htt abnormally interacts with other proteins (129; 178; 179; 271). It causes brain damage (38) producing oxidative stress, excitotossic processes and metabolism deregulation (108; 271). The expression of long Q tracts alone disrupts a wide variety of biological functions in cells and model organisms (142; 196; 334).

[1] The first Htt orthologs multi-alignment provides evidence that the polyQ is an ancient acquisition of Htt (296). Its appearance dates back to sea urchin that features a NHQQ sequence. This sequence consists of a group of four hydrophilic amino acids that can be considered bio-chemically comparable to the four glutamines (QQQQ) found in fish, amphibians, and birds (296). The polyQ has then expanded gradually in mammals. The longest and most polymorphic polyQ is in humans (296). Human Htt function may arise from the binding of different sets of interactors. Many proteins in the cells contain a polyQ, in particular transcription factors and transcriptional regulators (53).

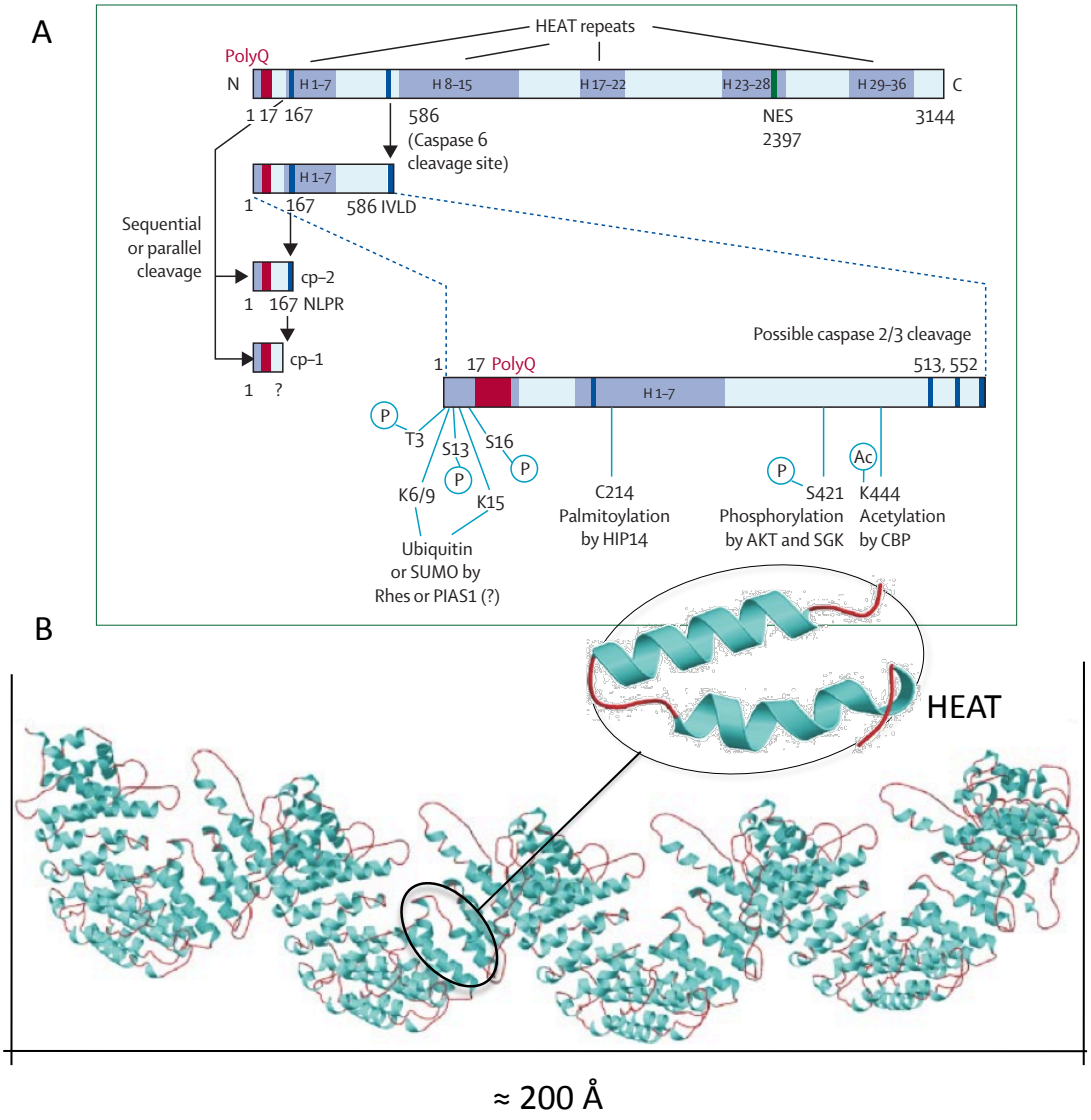


Figure 2.1: Htt - A) Human Htt is predominantly composed of HEAT repeats. The polyglutamine stretch (polyQ) is located at the N terminus. Proteolytic cleavage by caspase 6 and other (as yet, not characterized) proteases may lead to toxic N-terminal fragments, such as cp-1 and cp-2 in the figure. The exact size of these fragments and the relevant cleavage enzymes are currently unknown. Many post-translational modifications (eg, acetylation [Ac], phosphorylation [P], and addition of small ubiquitin-like modifiers [SUMO]) can alter HTTs cell biology and toxic effects. IVLD and NLPR are amino acid cleavage sequences. NES=nuclear export signal. Adapted from ref. (264). **B)** Proposed model of Htt structures. The HEAT domain is magnified.

Neuronal intra-nuclear and intra-cytoplasmic inclusions rich in polyQ are pathological hallmarks of HD (81; 304). The formation of inclusions proceeds through steps that generate different aggregated species, including nuclei, oligomers, protofibrils and large fibres, which form the microscopic aggregates found in patient neurons (262). The aggregation occurs through formation of a reservoir of soluble intermediates whose populations and stabilities may increase with polyQ length (43).

Inclusions have been proposed to be toxic, because they physically block axonal transport between the cell body and the synaptic terminal(335). In addition, they recruit other polyQ-containing proteins, mainly transcription factors. The latter might indeed lose their physiological function and cause cell death (111; 182; 186; 231).

However, these inclusions have been also proposed to result from an attempt of the cells to proteolytically degrade or inactivate mut-Htt (168; 272). This proposal is supported by the fact that cells forming Htt inclusions had an improved survival relative to those that did not form inclusions (13). Accordingly, there is little correlation between inclusion burden and the areas of the brain most affected in HD (113; 168)

The exact degree of toxicity of the other species involved in fibrillation is not known. Little information is available for the protofibrils (335). Recent studies have instead highlighted the roles of oligomeric species. The oligomeric species could be formed in several ways, including via N-terminal interactions or direct polyQ interactions (183; 225; 252). They may be highly reactive toward cellular environment (216; 260; 262; 306). However, these species might not be involved in the pathway to polyQ large inclusion formation (264).

2.3.2 Structural facets of polyQ and its aggregation propensity.

Perutz first showed that polyQ may form β-sheet rich structures. These structures may establish interactions with increasing strength with the number of Qs. These potentially affect the severity of disease (241). He suggested that the latter peculiarity is due to the Q side chain, which is able to establish inter and intra molecular hydrogen bonds interactions.[1] Subsequently other β-rich PolyQ models were proposed. They were stabilized by main and side chains H-bonds (12; 91; 94; 153; 197; 198; 199; 209; 224; 241; 244; 279; 280; 286; 290; 291; 331). Examples include Atkins's model, in which

[1]Notice that aggregates are not observed in proteins expressing polyN. PolyN contains amino acids that differ from Q by only one methylene group(226).

the H-bond network of the Q side chains allows for high-density packing (280) as well as Perutz's circular β-helix (242) and triangular β-helix models (251; 286), which are parallel β-sheets held together by main and side chain hydrogen bonds.

Obtaining experimentally structural information on polyQ in proteins turns out to be difficult, as short wild-type polyQ (lees than 30) lengths tend to be insoluble at the high concentrations required for crystallographic or NMR studies (306). A search of the Protein Data Bank (http://www.rcsb.org/) reveals that polyQ tracts seen in a variety of normal cellular proteins are annotated as "unstructured" or have to be removed to facilitate crystallization (306). Only one structure exists of the N-terminal part of Htt with 17 Qs, obtained by a fusion with maltose-binding protein. It features polyQ stretch that can adopt either an α-helical, random-coil, or an extended-loop conformation (155).

2.3.2.1 Role of the number of Q's.

It is not clear whether the toxicity can be triggered by a specific structural transition that occurs only when the number of Qs is larger than 36. Hence, several efforts have been spent for hunting the elusive toxic polyQ conformer (306). Investigations on anti-polyQ monoclonal antibodies, able to specifically recognize expanded toxic polyQ tracts, have leaded to the suggestion that a generic conformational epitope might form only above a certain polyQ length. Alternatively polyQ tracts could be simply inherently toxic sequences, whose deleterious effect gradually increases with their length (158).

2.3.2.2 Role of the flanking regions.

Originally, Htt aggregation has been thought to involve only the polyQ region (19; 261; 322). However, N17 (i) and polyP (ii) have been found to influence aggregation(306; 335) and to modulate toxicity and aggregation in *Htt Exon 1* (15; 36; 306). [1] Specifically:

(i) The deletion of PolyP in *Htt Exon 1* greatly increases the toxicity of *Htt Exon 1* fragments in yeast, which are otherwise innocuous (78). Therefore, PolyP appears

[1] Also sequences exogenous to *Htt Exon 1* modulate aggregation(86; 136; 241).

to be protective against the effects of expanded polyQ (36; 78). [1]

(ii) N17 modulates the toxicity of mut-Htt in a structure-dependent manner in mouse models (15; 306). A single point mutation in the middle of N17, shown to disrupt the possibility to reach an alpha-helical structure, completely abrogating any visible aggregates of mut-Htt (15). However the role of N17 in both polyQ aggregation and toxicity has been extensively discussed (63; 155; 176; 257; 298; 328). It has been shown that N17 controls sub-cellular localization, aggregation, and cytotoxicity of *Htt Exon 1* fragments in mammalian or yeast cells (15; 257). It has been proposed to nucleate Htt aggregation (136; 298; 299); to regulate the type of aggregate that forms (36; 328); and to mediate binding to cellular partners (115; 148) such as the chaperonin Tric (294), SUMO (284), and F-actin (278). Additionally intracellular antibody (intrabody) fragment binding to N17 decreases Htt exon1 aggregation, possibly by masking N17 (63). Finally deletion of N17 altogether reduces overall Htt aggregation (10), which may indicate a role for N17-mediated protein interactions in regulating protein misfolding (69; 328). Accordingly, the deletion in vitro of this region, strongly reduce polyQ aggregation (298).

A current proposal for *Htt Exon 1* (149; 298) points to the formation of oligomers having N17 in its core and polyQ exposed on the surface. When polyQ increases, the structure would decompact and the oligomers or protofibrils would rearrange into amyloid-like structures. The latter would rapidly propagate via monomer addition (149; 298).

[1] Because PolyP emerged in concomitance with longer polyQ regions (284), it may protect polyQ against its conformational collapse (296). Accordingly, the deletion in vitro of this region reduce polyQ aggregation (35; 72; 298).

3

Structural Properties of Polyglutamine Aggregates Investigated via Molecular Dynamics Simulations

Polyglutamine (polyQ) β-stranded aggregates constitute the hallmark of Huntington disease. The disease is fully penetrant when Q residues are more than 36-40 (disease threshold). Here, based on a molecular dynamics study on polyQ helical structures of different shapes and oligomeric states, we suggest that the stability of the aggregates increases with the number of monomers, while it is rather insensitive to the number of Qs in each monomer. However, the *stability of the single monomer* does depend on the number of side-chain intramolecular H-bonds, and therefore on the number of Qs. If such number is lower than that of the disease threshold, the -stranded monomers are unstable and hence may aggregate with lower probability, consistently with experimental findings. Our results provide a possible interpretation of the apparent polyQ length dependent-toxicity and, they do not support the so-called structural threshold hypothesis, which supposes a transition from random coil to a β-sheet structure only above the disease threshold.

3.1 Introduction

Huntington disease (HD) is an autosomal-dominant polyglutamine (polyQ) disorder.(38) It is caused by expanded CAG trinucleotide repeats in the gene that encodes the protein Huntingtin (Htt). The resulting mutant (mut-Htt), with an extended polyQ tract, interacts abnormally with other proteins and causes brain damage by leading to a generalized neuronal dysfunction,(38; 81; 271) including oxidative stress, excitotossic

processes and metabolism deregulation (Fig. 3.1) [1].(108)

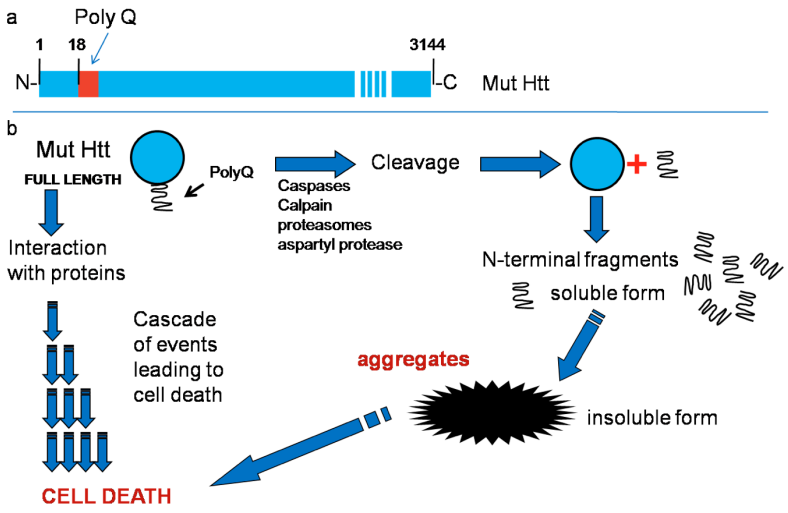


Figure 3.1: Huntingtin and Huntington Disease - (a) Schematic representation of mutated hungtinin (Mut-Htt). (b) Scheme of the events leading to cell death. The figure is based on studies by Borrell et al. (38)

The proteolytic processing of mut-Htt(194) exposes the toxicity of the mutant protein by releasing short N-terminal polyQ-containing fragments of 100-150 residues (exon-1). These fragments form *in vivo* and *in vitro* insoluble, β-sheet-containing aggregates,(33; 56; 64; 75; 275; 276; 295) that constitute the hallmark of the disease (Fig. 3.1).(64; 275) PolyQ peptides, as well as the exon-1 with a pathogenic Q tract, are sufficient to lead to the apoptosis of the cell [2],(25; 55; 150; 206; 262; 297) to cause the disease(196) and to feature the same aggregation properties of the full-length mut-Htt,(55) both *in vivo* and *in vitro*. The polyQs are toxic *per se* and the aggregation can be therefore studied by considering only polyQ peptides.(276)

The polyQ length correlates with the severity of the HD and with the age of its onset.(285; 334) The fully penetrant form of the disease involves polyQ tracts longer

[1]Early neuropathology involves alteration in the expression of neurotransmitter receptors(113; 185; 189; 271; 303; 314) and deregulated mitochondrial homeostasis(23; 230) that consequently disrupts calcium handling.(34) This in turn activates proteases such as calpain and caspase.(37)

[2]The Htt fragments have been observed in nuclear inclusion not only in Huntington but also in other polyQ diseases.(81; 194; 324; 326)

than 36-40 Qs (disease threshold),(268) as observed in other polyQ disorders.(334) This threshold varies, however, among polyQ diseases.(206) Consistent with these observations are the following considerations: (i) polyQ tracts longer than the toxicity threshold can form *in vitro* and *in vivo* SDS insoluble aggregates.(75; 103; 124; 275; 276) (ii) Specific recognition of expanded toxic polyQ tracts by an anti-polyQ monoclonal antibody has been observed, suggesting the existence of a generic conformational epitope formed only above a certain polyQ length.(148; 288; 304)

These facts have led to the so-called "structural threshold hypothesis", stating that only above the critical length of 36-40 Qs the polyQs do undergo a structural transition from random coil to a well defined structure based on β-sheet. This hypothesis has however been challenged by various experimental evidences. First, polyQ fragments shorter than the disease threshold also aggregate adopting similar structures to those peptides longer than threshold(25; 158; 206; 295) and exhibit toxicity in an eukaryote organism (*Caenorhabditis elegans*).(212) Second, the polyQ length turned out to affect more the kinetics of the aggregates formation rather than their stability.(158) In fact, the kinetics of polyQ aggregation is that of nucleated-grow polymerization;(55) aggregation is initiated by a monomer that functions as the critical nucleus with the nucleation event consisting of a random coil to β-sheet transition within an individual monomer (lag phase), then fibril formation proceeds via linear addition of single polyQ chains (elongation phase).(55; 276),(56) It has been shown that polyQ length influences the stability of the initial aggregation seed and that may affect the kinetics of its formation.(56) In contrast, the kinetics of the elongation phase is independent of the polyQ length.(57)

Experimental structural information on polyQ aggregates could shed light on the role of polyQ length for the aggregation process. Unfortunately, due to the insolubility of the aggregates, NMR and X-ray structures are so far lacking. Molecular modeling has therefore been the method of choice to have such insights.(12; 91; 94; 153; 197; 198; 199; 209; 224; 280; 286; 331) Several theoretical models of the aggregated polyQ units, consistent with the available experimental data (electron microscopy and X-ray data)(241; 244; 279; 280; 290; 291) have been proposed. These models are based on Perutzs suggestion that Q side chains, being similar to the backbone units, are able to establish an H-bond net.(241; 243) Thus, the proposed models are -sheet-containing structures stabilized by main and side chains H-bonds. Examples include

Atkinss model, in which the H-bond network of the Q side chains allows for high-density packing(280) as well as Perutzs circular β-helix(242) and triangular β-helix models,(251; 286) which are parallel β-sheets held together by main and side chain hydrogen bonds (Fig. 3.3 a,c and Fig. 3.2).

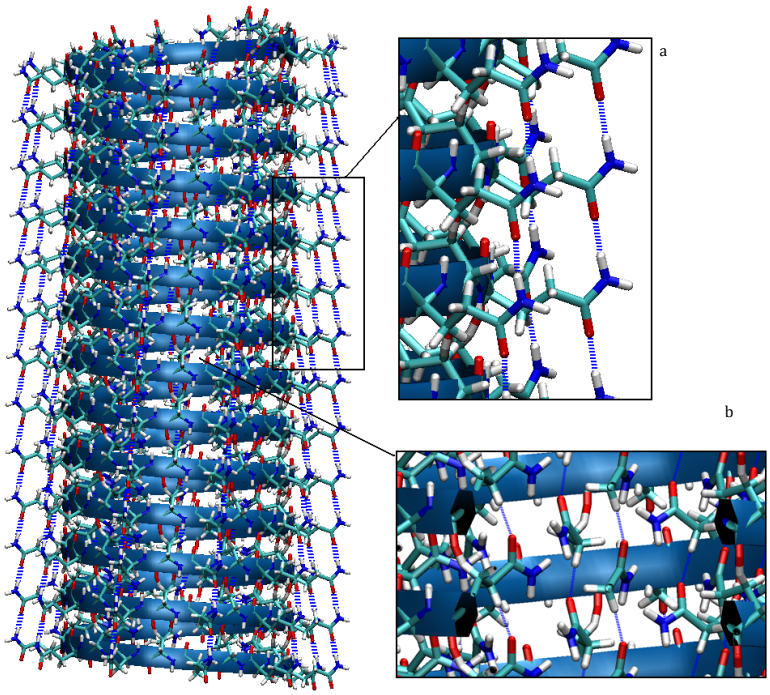


Figure 3.2: Circular β-helix - Molecular view of circular β-helix structure with particular of: (a) external HB net; (b) internal HB net. A similar H-bond scheme is observed also for the triangular β-helix structure.

Molecular simulations of the proposed models have provided valuable insights into their stability,(12; 91; 94; 197; 198; 199; 209; 224; 280; 286; 331) and the dependence of the structural stability on Q number has been partially addressed for the circular β-helix.(153; 209; 224; 286) In particular, an interesting report based on the circular β-helix model suggested that the stability of the structure increases with a repeated number of Q, and that above a critical Q number (30) the structure of the β-helix is kept stable.(224) Unfortunately such conclusions, being based on only one structural model, might depend on the initial structure. Furthermore, as discussed by the authors, the relatively short time-investigated (1 ns) might not allow equilibrating manually-built

models (as opposed to X-ray or NMR structures). In addition, the key issues on if and how the presence of a number of Qs larger than the disease threshold affects the stability of the polyQ structures in different shapes still need to be considered. Here we address these issues by performing 20 ns long molecular dynamics (MD) simulations on models in aqueous solution featuring a number of Qs well beyond the disease threshold, considering the circular β-helix(251) (Fig. 3.3 a) and triangular β-helix(286) (Fig. 3.3 c) models. These models are the only consistent with the structural threshold hypothesis (the outbreak of the disease above the 36-40 residues was explained on the basis of a structural change of the polyQ chain above this number). Therefore these models have been chosen to validate or discard such hypothesis.

Moreover by varying systematically the number as well as the size of the polyQ units in these models, and considering both the monomeric and oligomeric states, our calculations shed light on the dependence of the stability of the -helix structures upon the number of monomers, independently of the structural model chosen.

3.2 Results and Discussion

Predicting the thermodynamics stability of the models considered here (Fig. 3.3) in aqueous solution is currently not feasible with molecular simulations methods.

Structural features, instead, can be evaluated by MD simulations. In fact, previous theoretical works have introduced structural stability concepts using a variety of criteria, including the conservation of a specific conformation,(286) the acquisition of β-strand-like values of backbone dihedral angles(153) and the overall number of hydrogen bonds.(224) To simplify our discussion we qualitatively introduce the structural stability (**SS**) as a quantity which increases with: (i) the compactness of the structure, as measured by the plots of the RMSD of backbone atoms, as well as the Rg versus time; (ii) the hydrogen bond content (HBC), defined as the total number of H-bonds formed within the structural models, divided by the total number of H-bond donor functionalities.

Each quantity on which the **SS** depends is calculated based on 20 ns MD simulations for each model considered. We used **SS** to compare systematically the polyQ stability as a function of shape, number of monomers and number of Qs in a given monomer.

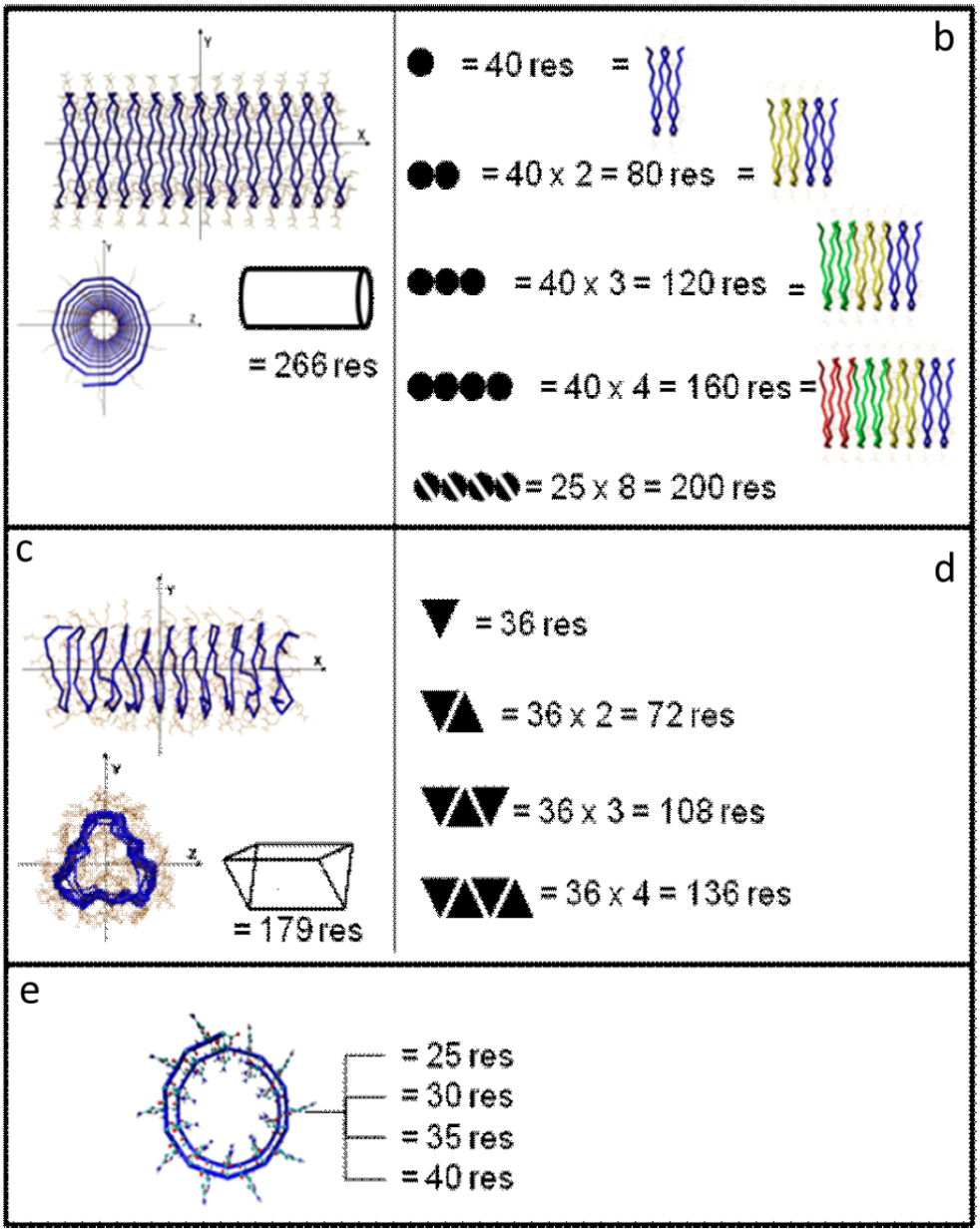


Figure 3.3: PolyQ models - PolyQ models undergoing MD simulations in water (not shown here for clarity). (a,c) side and front views of the large monomeric models. These are the circular (P) and triangular (T) β-helix monomers composed by a Q number well above the disease threshold.(303) (b,d) The oligomeric models are two series of oligomers (P_{AD}, P_{AC}, P_{AB}, PA and T_{AD}, T_{AC}, T_{AB}, T_A, as well as the oligomer P_{AH25}. (e) The small monomeric models are 4 monomers in circular β-helix conformation composed by 25, 30, 35, 40 Qs.

The dependence of **SS** on the shape is investigated in the two large monomeric models of circular (P) and triangular (T) β-helix reported in Fig. 3.3 a,c. These feature a number of Qs (266 Qs for P and 179 Qs for T) well above the disease threshold.(303) This number of Qs has never been observed in patients suffering from HD. Therefore, these models are also used as reference for the other systems considered here.

The dependence of the **SS** on the different number of monomers is studied in *polymeric structures* within the same shape (either circular or triangular shape), but with a different number of monomers, from 4 (P_{AD} and T_{AD}) to 1 (PA and T_A) (Fig. 3.3 b,d) and composed by short monomers of 40 and 36 Qs, respectively.

Finally, the dependence of the **SS** on the number of Qs per monomer is investigated in *small monomeric models* in circular β-helix shape, composed by a number of Qs below and above the disease threshold (Fig. 3.3 e).

In most cases, the quantities on which **SS** depends turn out to fluctuate around an average value after few ns (See A). Few models, specified in the following text, turn out not to equilibrate in the timescale investigated. Obviously, for those systems, averages cannot be taken. However, qualitative comparisons can still be made (see for instance ref. (31)).

3.2.1 Large monomeric models.

In our 20 ns MD runs, the RMSD of these models (P and T) fluctuate around average values of 0.18 and 0.25 nm, after 0.25 ns and 1.0 ns, respectively (Tab. 3.1 and Fig. A.1). The larger RMSD and the longer equilibration time of T, relative to that of P, might be caused, at least in part, by the fact that T is a purely theoretical model, whilst P is a model based on X-ray data interpretation.(244)

Rg decreases by few percent values during the dynamics, indicating that the MD models are more compact than the initial structures (Fig. A.2). The HBC of P and T is $\sim 70\%$ (Tab.1, Fig. 3.4). The contributions of backbone and side-chains are similar (Tab.1, Fig. A.3 A). However, the -sheet content [1] (βSC) of P is large, whilst that of T is just above 50% (Tab.1, Fig. 3.4; see also Fig. A.3 B). Thus, the stability of the two helices, which has been related to their HBC, seems not to be necessarily associated with a large content of secondary structure elements.

[1] The βSC is calculated here using the definition of W. Kabsch and C. Sander.(146)

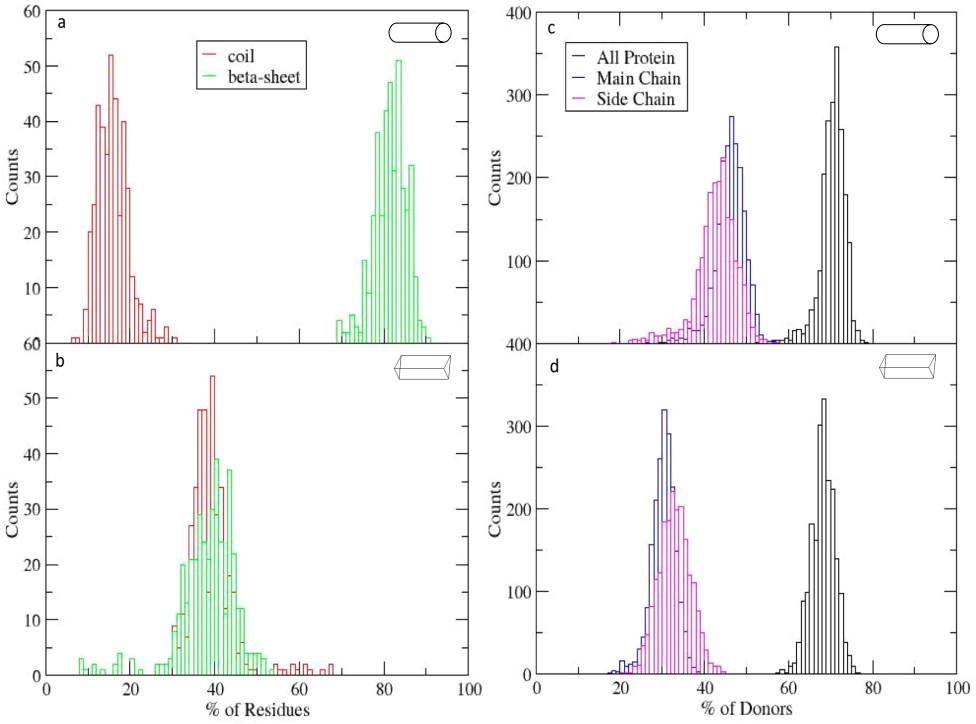


Figure 3.4: Secondary Structure - Percentage of residues in β-sheet (green line) and in random coil (red line) conformations in the large monomeric models considered here: the circular and the triangular β-helices in (a) and (b), respectively. Percentage of donor atoms involved in H-bonds in the whole protein (black line), in the main (blue line) and in the side chains (violet line) for the circular and the triangular β-helices in (c) and (d), respectively.

3.2.2 Oligomeric models.

For both the P and T series, the RMSD increases by decreasing the monomers from four to one (Tab. 3.1, Fig. A.5 A).

The Rg values of the oligomers which feature four and three monomers (Tab.3.1, Fig. **??** B) (P_{AD}, P_{AC}, T_{AD}, T_{AC}), are similar to those of P and T (Fig. **??**). Those with two or one monomers (P_{AB}, PA and T_{AB}, T_A) do not preserve compact folds (See SI).

The HBC of P_{AD}, T_{AD} is not much smaller than those of P and T. Those featuring three monomers are still significantly large. However, the values decrease rapidly by decreasing the number of monomers (3.1, Fig. 3.5). In all circumstances, the largest contribution to HBC is due mainly to the backbone rather than to the side-chains. The secondary structure content follows the same trend of the HBC (Fig. 3.6).

Thus, the **SS** of the assemblies of four and three monomers are similar to those of the single chain systems P and T. This means that in the short monomers forming the oligomeric systems the Qs interact with each other almost as they do in the large monomeric models. This result does not depend on the shape, as P_{AD}, P_{AC}, T_{AD}, T_{AC} behave similarly.

However, the **SS** of P_{AB}, T_{AB} dimers are much smaller than that of T, P and this feature is even more pronounced in the case of the monomers PA and T_A. Also in these cases the **SS** does not depend on the shape. Therefore, the **SS** decreases with a decreasing number of monomers, independently of the shapes.

Because **SS** does not depend on shape and T features a different number of glutamines with respect to P, we suggest that the stability of oligomeric structures may not depend on the number of Qs in each monomer. To test this hypothesis, we construct an oligomer equivalent to P_{AD} (i.e. PAH25), except that each monomer features a smaller number of Qs. This is a circular β-helix oligomer composed by 8 monomers, each 25 Q long (model PAH25 in Fig.2b). The values of the quantities on which **SS** depends turn out to be similar to those of P_{AD} (Fig. A.6), corroborating the hypothesis that **SS** is not influenced by the number of Qs composing each monomer.

We conclude that **SS** depends on the number of monomers *independently of their shape and of their length (number of Qs in each monomer).* This is consistent with

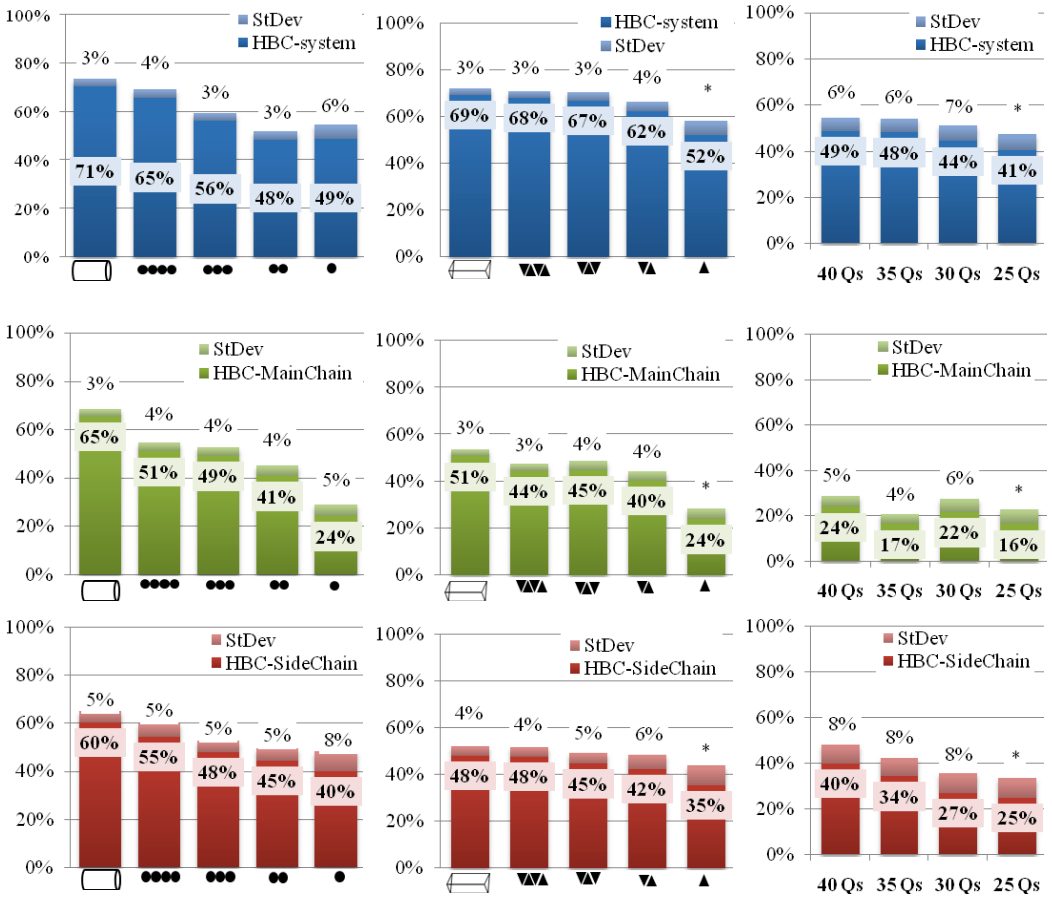


Figure 3.5: HBC in P and T series - Percentage of H-bonds in P series (first column), T series (second column) and monomeric series (third column). Blue histograms represent the total HBC, green and red ones represent main chain and side chain HBC, respectively.

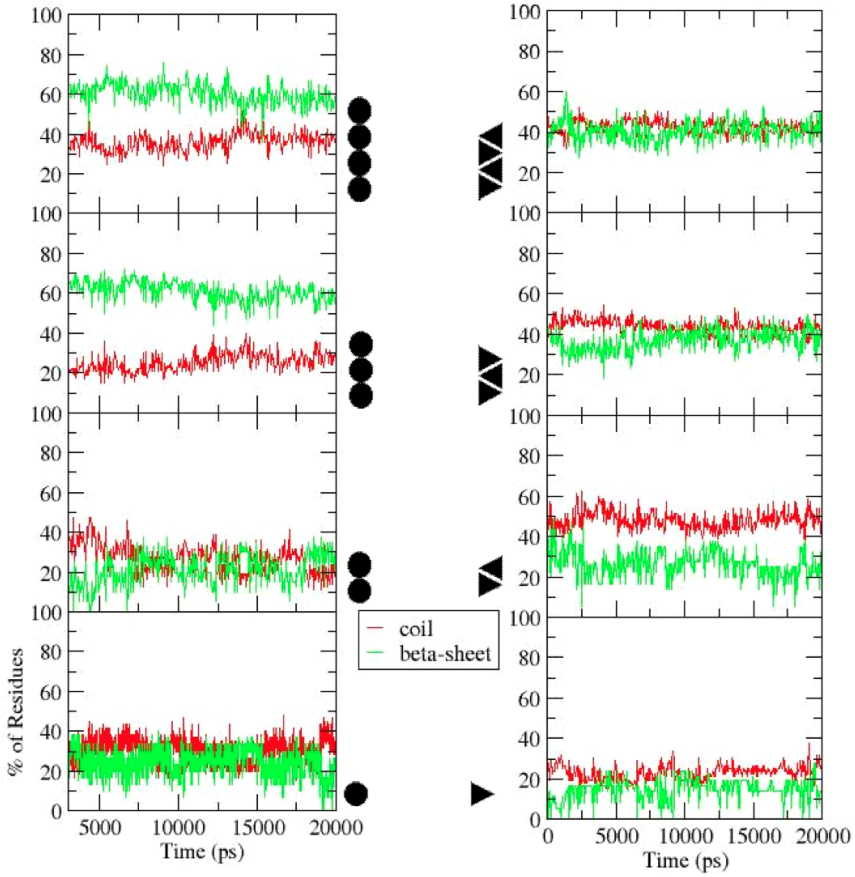


Figure 3.6: Secondary Structure of oligomers - βSC (green) and random coil conformation (red) of oligomers in circular (left panel) and triangular (right panel) β-helix conformations.

NMR data of Klein et al. (158) showing that aggregates formed by short and long polyQ tracts share similar structural properties.

3.2.3 Small monomeric models.

We finally investigate the dependence of the **SS** on the length, focusing only on one shape, because, as seen in the previous section, **SS** turns out not to depend significantly on the latter. We focus here on the circular βhelix shape and we consider monomers of lengths below and above the disease threshold, by choosing four systems containing 25, 30, 35 and 40 Qs (Fig. 3.3 e).

During the MD, P_{25} undergoes the largest structural rearrangements among the models investigated here. Its initial shape is lost after about 10 ns, as shown by a large increase of the RMSD and Rg values (Fig. A.7 A, B). The final MD structure is shown in Fig.7.

In the other small monomeric models, the RMSD decreases by increasing the number of Qs (Tab. 3.1). In fact a similar trend, but with smaller RMSD and Rg values, is found for P_{30} (Fig. A.7 A, B). Remarkably, the RMSD and Rg of P_{40} and P_{35} are similar to those of P (Fig. A.7 A, B), preserving the original β-helix fold.

The overall HBC increases with the number of Qs (Fig.4). Such an increase is caused by the side chains H-bonds, whilst the backbone contribution is similar in the four models (Fig.4). This result is opposite to that we found for the oligomers, in which HBC mainly depends on the H-bonds between backbone atoms (Fig. 3.5). Finally, the βSC of P_{40} is much larger than that of the other three models (Fig. 3.7 A).

Based on these results, we conclude that: (i) the **SS** increases progressively with the number of Qs. (ii) **SS** is associated to an increase of side-chain H-bonds. This suggests that the **SS** of single monomers depends on the possibility of forming intra-side-chains H-bonds.

3.3 Relevance of our results for the toxicity threshold proposals.

In a widely accredited hypothesis, the polyQ length affects the nucleated-grow polymerization kinetics of polyQ aggregation, rather than the stability of the aggregates.(158)

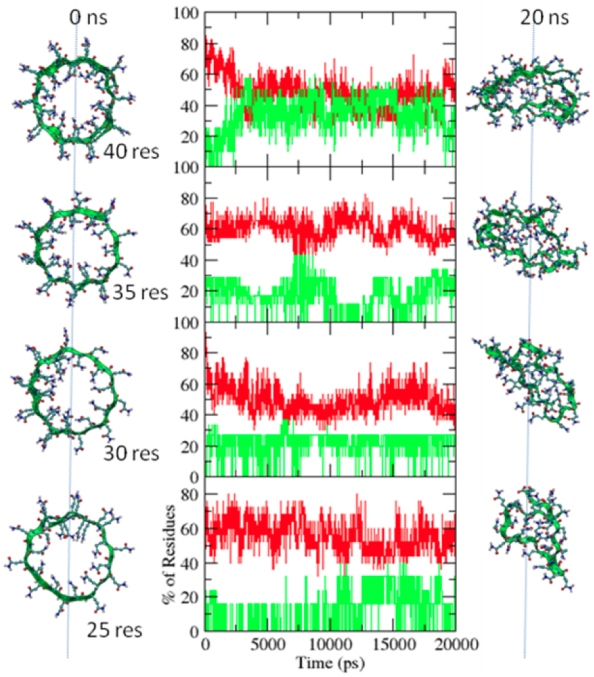


Figure 3.7: Secondary Structure of monomers - βSC (green) and random coil (red) conformations of P_{40}, P_{35}, P_{30}, P_{25} monomers (See Text.) On the left and right side of the graph the initial and final (after 20 ns MD) geometries of each monomer are shown. Water is not shown for the sake of clarity.

3. STRUCTURAL PROPERTIES OF POLYGLUTAMINE AGGREGATES INVESTIGATED VIA MOLECULAR DYNAMICS SIMULATIONS

System	RMSD	$\Delta\sigma$	Rg	$\Delta\sigma$	HBC system	$\Delta\sigma$	HBC MainChain	$\Delta\sigma$	HBC SideChain	$\Delta\sigma$
	nm	nm	nm	nm	% of donors	% of donors	% of donors	% of donors	%of donors	% of donors
P	0.18	0.02	2.11	0.02	70.56	2.93	65.34	3.10	60.49	4.87
T	0.25	0.02	1.72	0.01	69.30	2.81	50.57	3.09	48.26	3.99
P_{AD}	0.27	0.03	1.52	0.02	65.47	3.55	50.91	3.67	54.57	
P_{AC}	0.15	0.03	1.39	0.01	56.08	3.08	48.99	3.61	48.07	4.78
P_{AB}	0.22	0.03	1.25	0.01	48.25	3.49	40.69	4.34	44.57	5.16
P_A	0.36	0.04	1.05	0.02	48.81	5.66	24.07	4.94	40.12	8.19
T_{AD}	0.19	0.03	1.51	0.01	68.22	2.83	44.35	3.24	47.64	4.13
T_{AC}	0.20	0.02	1.28	0.01	67.03	3.49	44.81	3.81	44.70	4.70
T_{AB}	0.17	0.03	1.11	0.01	62.47	4.07	40.33	4.08	42.38	5.91
$*T_A$	0.30	—	0.95	—	52.01	—	23.87	—	35.45	—
P_{AH25}	0.23	0.02	1.47	0.01	67.75	3.35	52.59	3.66	52.71	5.12
P_{40}	0.46	0.04	1.05	0.01	48.81	5.66	24.07	4.94	40.12	8.19
P_{35}	0.61	0.06	0.98	0.02	48.26	5.67	16.51	4.48	34.45	7.89
P_{30}	0.58	0.06	0.99	0.03	44.41	6.87	21.63	5.78	27.18	8.42
$*P_{25}$	0.65	—	0.95	—	40.60	—	16.49	—	24.94	—

$*$ Systems which do not equilibrate in the timescale investigated . Values are however taken from the last snapshot of the dynamics, as the systems are not equilibrated in the 20 ns of dynamics (See SI)

Table 3.1: SS of all the systems studied reported in terms of: average values of RMSDs, Rg and HBC. The properties on which SS depends of PAH25 are displayed in the SI (Fig. A.6).

System	β-sheet content	$\Delta\sigma$
	% of res	% of res
P	81.54	3.82
T	41.30	5.42
P_{AD}	60.51	6.01
P_{AC}	60.15	5.11
P_{AB}	20.43	7.83
P_A	23.26	7.74
T_{AD}	40.50	5.00
T_{AC}	36.89	5.30
T_{AB}	20.14	6.77
$*T_A$	14.50	—
P_{AH25}	59.01	6.02
P_{40}	34.8	7.60
P_{35}	17.8	9.83
P_{30}	17.9	9.94
$*P_{25}$	11.90	—

Table 3.2: βSC. The properties on which βSC of PAH25 are displayed in the SI (Fig. A.6).

At the molecular level, the process would start from a random coil to β-sheet structural transition of an individual monomer, which constitutes the aggregation seed. (55; 56; 57)

Our findings are consistent with this hypothesis, in fact, HBC and βSC of single isolated monomers increase with the number of Qs. Assuming that of HBC and βSC in the transition state leading to the formation of the initial aggregation seed have a similar relevance to that found here for the final products (the aggregates), we formulate the hypothesis that the overall aggregation kinetics may increase upon incrementing the number of Q residues. Although reasonable, this proposal, at present, can be discussed only at a speculative level.

In addition, the HBC and the βSC of the oligomers depend on the number of monomers and not on the number of Qs in each monomer. This is consistent with the hypothesis that, once the initial polyQ seed is formed, the kinetics of the elongation phase does not depend on the number of Q.(56; 57)

Limitations of the calculations.

As any modeling investigation, this study has several limitations. First, the timescale, although longer than that reported in previous studies on folded polyQ peptides,(224) is too short to follow the time evolution of models T_A and P_{25}, (whilst the other models appear to be well equilibrated (Tab. 3.1 3.2 and Fig. A.1, A.5 A, A.7 A)). Thus, any conclusion based on comparisons among the models is necessarily qualitative.

Second, the aggregates have been experimentally characterized at several different ionic strengths, ranging from 20 to 200 mM.(297) Although some of the properties may be affected by varying ion concentrations, we expect that our conclusions on the folded peptides (which have been reported at zero ionic strength) will not change at the qualitative level. In addition, we stress here that no attempt has been made to study the condensation process of the peptides, which are likely to depend dramatically on the simulation conditions.

Finally, our conclusions have been drawn for *a subset of proposed models* (not real structures). Again, this fact prevents to make any quantitative statement, which is anyway beyond the scope of this paper.

3.4 Concluding remarks

Based on our 20 ns MD simulations with GROMOS96 force field for each systems studied (Fig. 3.3), we suggest the following qualitative conclusions:

- The two different -helix shapes, originally invoked to explain the disease threshold, affect only the β-sheet content: the T helix displays a larger number of residues in random coil conformation than the P helix. However, the H-bond content as well as the **SS** of the two shapes is comparable.

- The structural stability of P and T oligomeric systems is not affected by the shape. This has never been shown so far. In addition, **SS** does not depend on the number of Qs in the monomers. Although a recent theoretical study on polyQs suggests that the stability of an aggregate should be regulated by the structural stability of its component monomers, as stated by the authors, the short time-scale investigated (1 ns of MD simulations with AMBER force field(50; 321)) appears not to be sufficient to draw general conclusions.(224) Indeed, we demonstrate that for longer simulations, the oligomer with 4 monomers of 40 Qs and the oligomer with 8 monomer of 25 Qs have similar **SS**. Consistently with our results a recent NMR analysis reveals no structural difference between aggregates formed by short and long polyQ peptides.(158) In summary, our results do not support the structural threshold hypothesis, which suggests a specific conformation for polyQs longer than the threshold.

- Only the number of monomers - thus the concentration in an (*in vivo* or *in vitro*) experiment - contributes to the overall stability of the oligomers because of the additive contribution of single monomer in the H-bond net formed between backbone atoms.

- The H-bonds formed between Q side-chains influence mainly the stability of the single isolated monomer and to a lesser extent the stability of oligomers.

- Interpreting our findings on the basis of the whole landscape of available experimental data,(158) we suggest that the observed length-dependent toxicity threshold may be explained by a faster aggregation kinetics that occurs for longer polyQ tracts.

3.5 Computational Details

3.5.1 Model Systems

Large monomeric models.

The coordinates of the circular β-helix nanotube(244) (named as P From Perutz, who introduced this model in 2002,(244) Fig. **??** a) were kindly provided by Dr. A. Lesk. The structure is characterized by 266 Q residues with ϕ and ψ angles of -162°, 159° and each turn contains 20 residues. The triangular β-helix model (named as T, Fig. **??** c) was constructed starting from the regularly shaped coils from UDP-N-acetyl glucosamine acyltransferase(251) (Protein Data Bank entry: 1LXA),(30) replacing each residue with a glutamine. This model contains 179 residues and each turn is composed by 18 Qs. Both models are composed by a single polyQ chain. Both models have a number of Qs well above the number of polyQs observed at physiological conditions.

Oligomeric models.

Starting from the single chain systems we build the following series of oligomers: I. 4 oligomers in circular β-helix conformation composed by 4, 3, 2, 1 monomers, respectively. These models are named P_{AD}, P_{AC}, P_{AB}, P_A, and each monomer is composed by 40 Q residues (Fig. 3.3 b). II. 4 oligomers in triangular β-helix conformation with respectively 4, 3, 2, 1 monomers. These models are symbolized by T_{AD}, T_{AC}, T_{AB}, T_A, respectively, and each monomer is composed by 36 Q residues (Fig. 3.3 d). III. 1 oligomer in circular β-helix conformation composed by 8 monomers each containing 25 Q residues (Fig. 3.3 b). The model is named P_{AH25}

Small monomeric models.

We considered 4 monomers in circular β-helix conformation composed by 25, 30, 35, 40 Qs and symbolized by P_{25}, P_{30}, P_{35}, P_{40}, respectively (Fig. 3.3 e).

The total charge state of all the systems studied is neutral. The oligomeric and the monomeric model are constructed starting from the coordinates of the continuous systems, selecting a number of Q and considering each residue in the zwitterionic form.

3.5.2 MD Simulation Protocol

Classical MD simulations of all systems were performed with GROMACS(28; 311) software in the canonical (NVT) ensemble with a time step of 2 fs for numerical integration. The LINCS algorithm was used to constrain all bond lengths(120) and the temperature was kept close to 300 K by a weak coupling to an external bath(27; 29) with a coupling constant equal to the time step. GROMOS96(312) force field was employed, which uses an explicit representation of acidic hydrogen atoms.

Long-range electrostatic interactions were treated with the particle mesh Ewald (PME) method,(71; 92) using a grid with a spacing of 0.12 nm combined with a fourth-order B-spline interpolation to compute the potential and forces in between grid points. The cutoff radius for the Lenard-Jones interactions as well as for the real part of PME calculations was set to 0.9 nm. All systems were solvated with SPC waters(26) in a periodic rectangular box large enough to contain the protein and at least 0.9 nm of solvent molecules on each side of the solute.

The systems were initially relaxed by imposing harmonic position restraints of 1000 Kj/(mol*nm) on solute atoms, allowing the equilibration of the solvent without distorting the solute structure. After an energy minimization of the solvent and the solute without harmonic restraints, the temperature was gradually increased from 0 to 300 K using simulated annealing (SA).(214; 218) The SA was performed by increasing the temperature from 0, to 300 K in 12 steps in which the temperature was increased by 25 K in 100 ps of MD.

3.5.3 Calculated properties.

The root mean square deviation (RMSD), the gyration radius (Rg), hydrogen bonds and secondary structure content were calculated using the formulae reported in the Supplementary Information (SI). Acknowledgments. The authors thank Dr A. Lesk for providing the structural model of the circular β-helix.

4

Hydrogen bonding cooperativity in polyQ β-sheets from first principle calculations

Polyglutamine β-sheet aggregates are associated with derangement of Huntingtons disease. The effect of cooperativity of the H-bond network formed both by back-bone and side chains groups is expected to be important for structure and energetics of the aggregates. So far no direct description and/or quantification of the effect is yet available. By per-forming DFT and hybrid DFT/MM simulations of polyglutamine β-sheet structures in vacuo and in aqueous solution, we observe that the cooperativity of glutamine side chains affects both the directions perpendicular and parallel to the backbone. This behavior is not usually ob-served in β-sheets and may provide significant extra-stabilization together with explaining some of the unique properties of polyglutamine aggregation.

4.1 Introduction

Huntington and other neurodegenerative diseases depend on the abnormal expansion of poly-glutamine (polyQ) tracts in proteins which form aggregates rich in β-sheets associated with neurodegeneration. (56; 57; 75; 205; 241; 245; 276) The glutamine side chain is similar to the backbone unit. Thus, polyQ tracts can form particular β-strands stabilized by a hydro-gen bond (HB) net involving both backbone and the side chains.(158; 241; 245) The pres-ence of a Cooperative Effect (CE) on this peculiar HB net may play a role in the misfolding and aggregation of polyQ.(241) CE in hydrogen bonding is very important for both the structure and the energetic of polypeptide systems:(193)

4. HYDROGEN BONDING COOPERATIVITY IN POLYQ β-SHEETS FROM FIRST PRINCIPLE CALCULATIONS

The extra-structural stability of polyQ aggregates due to the CE is related to the number of HBs formed between backbone and side chains.(265) Nevertheless, the conclusions so far were achieved by classical molecular dynamics calculations that cannot answer the critical issue of how to deal with electronic polarizability. This can be described by first principle methods, which have in fact already applied in the study of CE on polypeptides, including polyQ chains. (127; 128; 139; 274; 307; 313; 316; 327; 333) However, the crucial role of Q side chain HBs on CE has not been investigated so far by first principles approaches.

Here we perform first principles DFT-PBE(24; 213; 238) calculations on polyQ peptides of increasing complexity, assembled in parallel [1] β-sheets (Scheme 4.1), a structure well characterized from biochemical and theoretical studies.(22; 160; 237; 307) Our models [2] (Nxn hereafter) differed from each other for the number of strands ($N = 1, 2, 3, 4$) and/or for the number of Q in each strand ($n = 1, 2, 3, 4$). They are terminated by the ad-dition of $-NCH_3$ and $-OCCH_3$ groups. The resulting 16 models range from 29 to 320 atoms. (Scheme 4.1. See caption for more details on notations). Next, because of the obvious role of solvent and temperature effects on polypeptide conformation(274) , we performed 2 ps of hybrid DFT/MM molecular dynamics calculations on a large system, a β-helix [3] nanotube (8 turn of 20 Q, see Fig. B.1 in SI B) in aqueous solution. (28; 175; 310**?**)

Taken together, our calculations suggest that the CE is manifested both by the shortening of HB lengths increasing the number of HBs involved (**structural** aspect) and by the en-ergy stabilization of H-bonded peptides with respect to the isolated ones (**energetic** aspect): We are going to detail in the following some of the crucial feature of our results.

Finally, to prove that such cooperative effects are due only to the peculiarity of polyQ chains, we also considered, as a control study: a) series of models where we

[1] CE turns out to be stronger in parallel β-sheets (like the systems considered here) than in anti-parallel ones.(160)

[2] The models were built using *HyperChem 8.0* program.(**?**)

[3] The structure is characterized by Q residues with ϕ and ψ angles of -162 and 159 degree.(244) Its coordinates were kindly provided by Dr. A. Lesk. Although β-helices have a low probability to form in vivo respect other Q structures, (280; 331) they have been investigated here because: 1) they have been already investigated by classical MD by us;(265) 2) we provide a qualitative description CE, independently from the peculiarity of these conformation. Quantitative predictions, which would require an investigation on a variety of structure proposed, are beyond the scope of the present investigation.

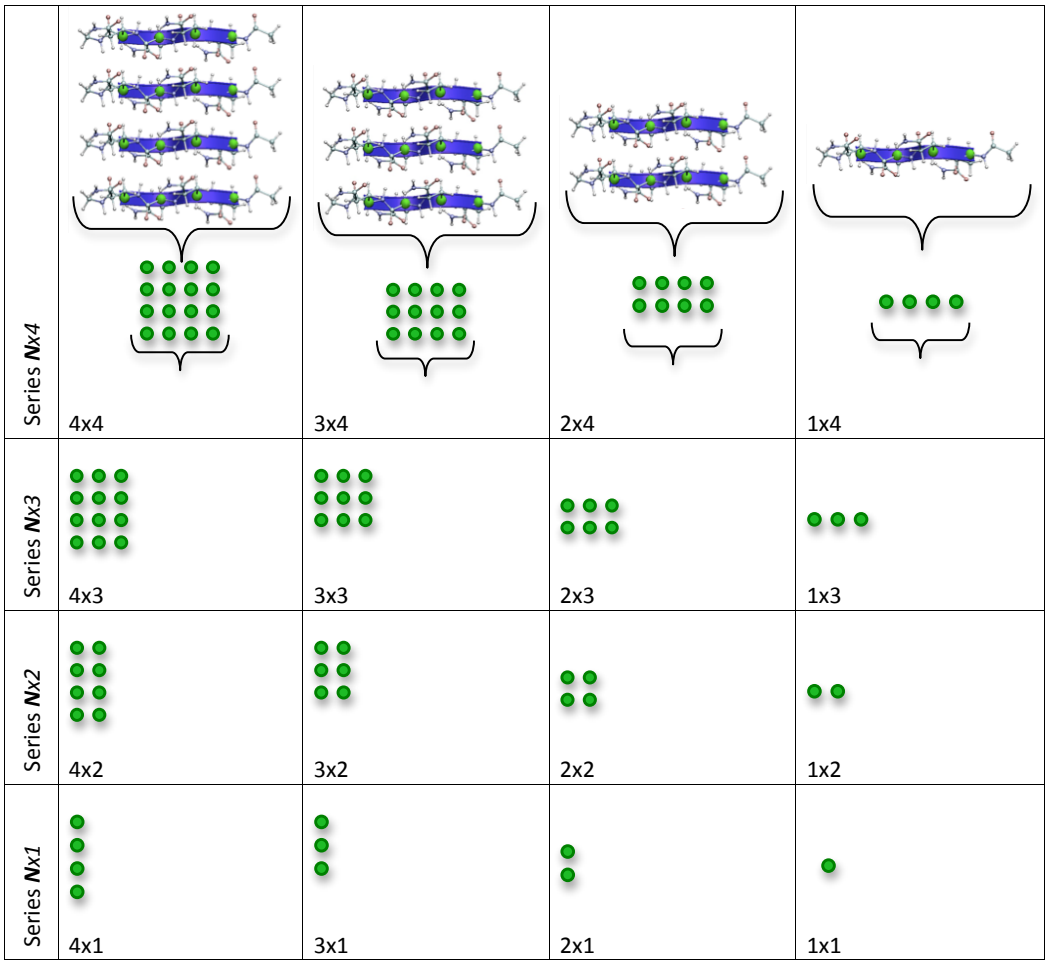


Figure 4.1: Table of Systems - Each system studied here is defined in terms of the n and N integers, ranging from 1 to 4. The first number counts the Qs in each strand. It defines a group of four systems each with the same number of Qs for strand, but with a different number of strands (a series). The second counts the strands in each system. Thus, Nx4 indicate systems formed by peptides of 4 Qs (1x4, 2x4, 3x4, 4x4). Nx3 those formed by 3 Qs (1x3, 2x3, 3x3, 4x3). Nx2 those formed by 2 Qs (1x2, 2x2, 3x2, 4x2). Nx1 those made up of only 1 Q (1x1, 2x1, 3x1, 4x1).We built also other two different Nx3 series: A) Nx3SC polyQ series where we varied the side chain conformations. B) Nx3ALA. This is a polyalanine system.

varied the initial Q side chain conformations; b) series of models built with polyalanine.

4.2 Structural aspects.

CE on a β-sheet system may be present in patterns perpendicular to the peptide elongation ($\perp$CE) or parallel to it ($\|$CE, 4.2 A).(333)

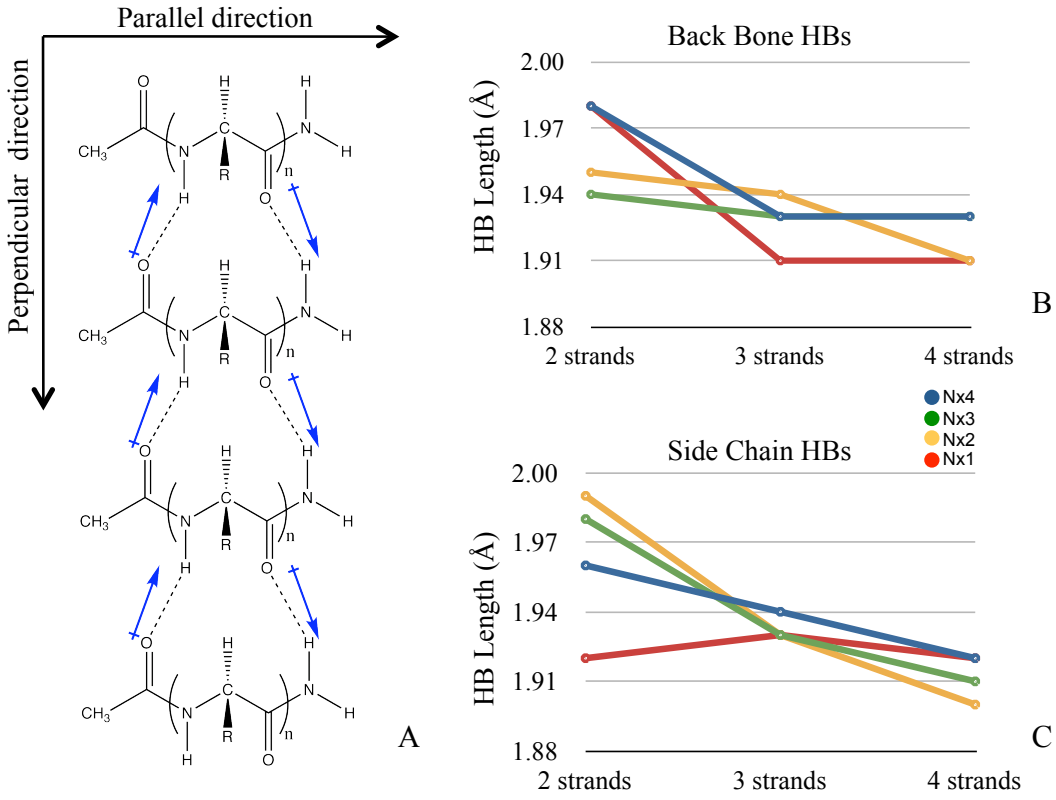


Figure 4.2: Cooperative Effect - A) Parallel ($\|$) and perpendicular ($\perp$) directions of peptides elongation. B)-C) Structural aspects of CE: B) Backbone CE ($\perp$CE-effect a): Mean values of HB lengths of the backbone atoms versus the number of strands for each series of n Q. C) Side chains CE ($\perp$CE-effect a): Mean values of HB lengths for the side chain atoms versus the number of strands for each series of n Q.

1. The $\perp$CE is manifested (a) by a decrease of HB length with an increasing number of piling strands, and (b) by HBs at the center of the pile shorter than in the rim.(160; 333)

In all the series considered, HB distances decreased with an increasing number of

piling strands in both the backbone and the side-chains (Effect a in Fig. 4.2 B-C, Tab. B.2).

In addition, HB lengths turned out to be shorter at the center of H-bonded chains than at the rim, in the case where at least three HBs are piled up in the perpendicular direction (N=4) (Effect b in Fig 4.3, 4.4, 4 A and Fig. B.3). This feature was observed both for the side-chains (Fig. 4.3) and the backbone (Fig. 4.4, 4.5 A, B.2, B.3).

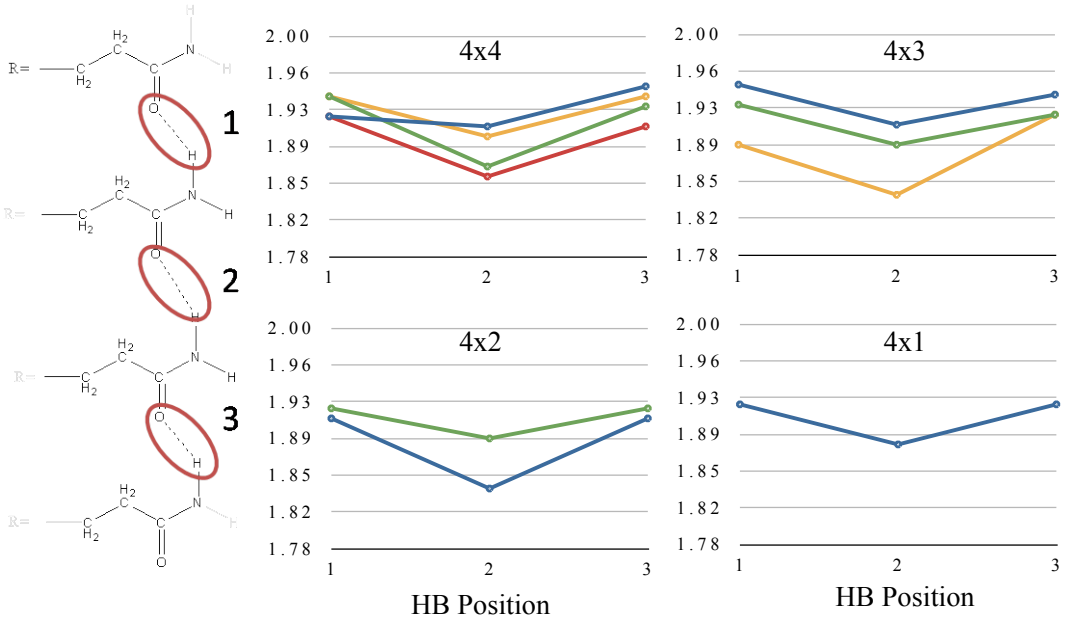


Figure 4.3: ⊥**CE** - ⊥CE-effect b: in the Q side chains: Systems 4x4, 4x3, 4x2, 4x1. HB length versus HB position.

For the backbone, the trend was however observed only when taking averages: the inner HBs turned out not always to be the shortest of the column [1] (Fig. 4.4). This fact can be explained, at least in part, by the polarization of the dipoles associated with the HB functionalities (C=ON-H) of both backbone and side chains. It has been already observed that the backbone dipoles along the same column of β-strands have the same orientations (in contrast to those of the adjacent column) and can therefore sum up increasing the polarization of the systems.(160) How-ever, in the case of polyQ β-strands, the glutamine side chains counterbalance this polarization, affecting the inner HBs (Fig. 4.4). Therefore in the columns where the HB dipole orientations were

[1] With the term column we in-dicate the HB chain in perpendicular direction

enhanced by similar side chain HB dipole orientations, a CE is present: the shorter HB was the one in the center of the column; on the other hand, when neighboring side chain columns had HB dipoles oriented in opposite directions (with respect the column considered): the inner HB was not the shortest of the column. (Fig. 4.4)

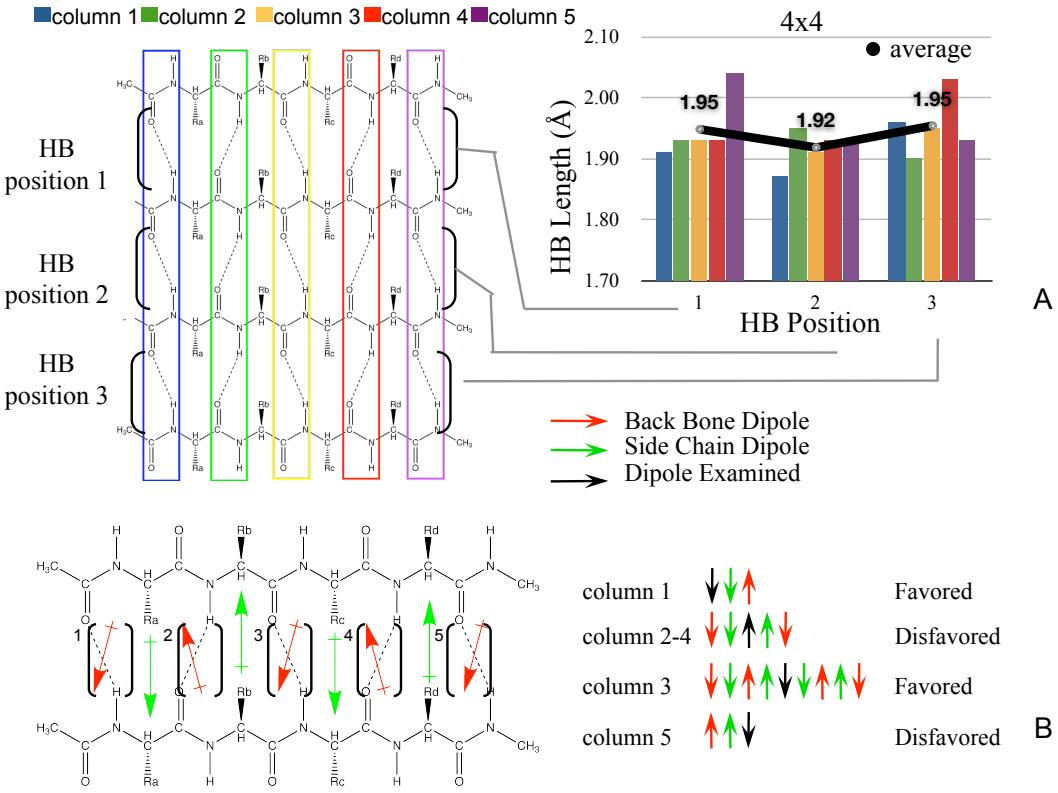


Figure 4.4: $\perp$CE-effect b in system 4x4. - A) In the histograms: HB length of backbone for different positions inside each strand as a function of the position across the different strands). Color of the histogram corresponds to the HBs circled on the top-left picture. The black line represents the mean values over the rows. B) Orientation of dipoles associated with the HBs for 4x4 (4x3, 4x2, 4x1 treated in SI, Fig. B.3).

To prove this conclusion, we perform the same calculations on the Nx3 polyQ series varying the side chain conformations ($Nx3_{SC}$ hereafter). Here glutamine side chains are twisted in such a way they are not able to establish HBs, thus only backbone HBs are present. As expect, we found both type of $\perp$CE (effect a and b). However, remarkably $\perp$CE-type b, is not affect by side chain HB dipole orientations due to the absence of side-chain HBs. Thus backbone HB lengths turned out to be shorter at the

center of H-bonded chains than at the rim, in each sin-gle columns d considered, not only taking the average. (Fig. B.5 - Tab B.4) According to this, we have similar results also for the $Nx3_{ALA}$, where there is no possibility for the alanine to form side chain HBs. (Fig. B.5 - Tab. B.4) Indeed, we found only a ⊥CE of type b: HB lengths are shorter at the center of H-bonded chains than at the rim, in the case where at least three HBs are piled up in the perpendicular direction (N=4). (Fig. B.5)

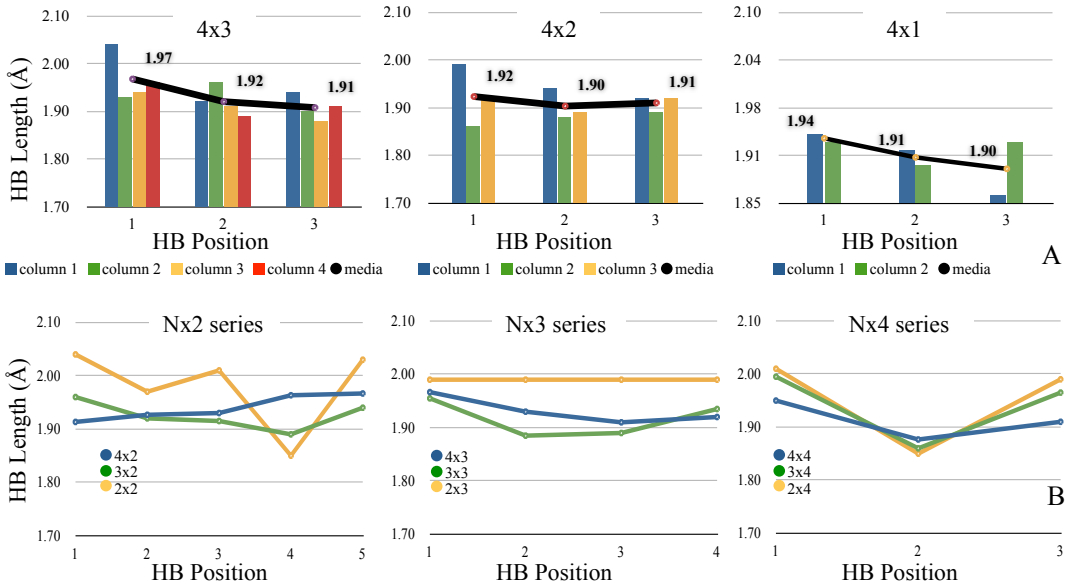


Figure 4.5: Backbone CE - A) Backbone CE in the direction perpendicular to strand elongation (⊥CE-effect b): Systems 4x3, 4x2, 4x1. In the histograms: HB length for each column (the position along the strand) versus the HB position (the position perpendicular to the strand direction); the black line represents the mean values over the rows. B) Backbone ‖CE: Series Nx2, Nx3, Nx4. HB length versus HB position.

2. We observed a ‖CE as reflected from shortening of central HB lengths between two adjacent strands.(333) ‖CE is usually not present in β-sheets because of the alternative orientation of backbone HB dipoles along the strands (Fig. 4.2 A).(123; 163; 333) However, the dipoles associated with the Q side chains add up in a coherent way for the central HBs between two strands (position 2 in Nx2 series, position 2 and 3 in Nx3 series, posi-tion 2, 3 and 4 in Nx4 series). As a result, the latter turned out to be shorter than those of the rim (Fig. 4.5 B). We performed the same calculation on the Nx3SC systems, where there is not the con-tribution of side chains HB. As expected,

no ∥CE is found.(Fig. B.6 a) These results point to the relevance of glutamine side chains for the structure of polyQ systems.

To confirm that such cooperative effects are specific only for polyQ and not a general feature of polypeptide chain, we performed a control study also on a series ($Nx3_{ALA}$) of poly-alanine systems (Tab. B.4). As expected, no ∥CE or ⊥CE type a, are found. (Fig. B.6 b)

Similar conclusions can be drawn by our hybrid QM/MM calculations of the circular β-helix of polyQ chain, in which the QM region correspond to 4x4 to 4x3 and 3x4, and the rest of the polyQ tracts as well as the water molecules were included in the MM region ($\sim$ 45000 atoms). These systems were labeled as $4x4_{MIX}$, $4x3_{MIX}$, $3x4_{MIX}$. Although the trend of HB lengths in the first two systems qualitatively resembled that of the corresponding in vacuo models, we have to remark that the HB lengths were larger (Tab. B.8 and Fig. B.9). Moreover the side chains formed mostly HB with the solvent. These differences are probably due to the presence of the solvent and to temperature effects, which are completely neglected in the in vacuo calculations [1]. No CE was observed in the last system ($3x4_{MIX}$), possibly because of the small number of strands.

4.3 Energetic aspects.

The stabilization energy associated with the formation of HBs between the different strands [2] of the systems *in vacuo* is calculated as follows. First, we define the stabilization energy *per strand* (ΔE_N) as the energy associated with the addition of the Nth Q strand to the $Q_{N-1}(E_{Nxn})$, minus the formation energy of N isolated strand.

$$\Delta E_N = E_{Nxn} - N \cdot E_{1xn} \tag{4.1}$$

E_{Nxn} is the energy of a system belonging to the n series and containing N strands; E_{1xn} is the energy associated to an isolated polyQ strand with n glutamines. This is the

[1] We further notice that because of the very short time-scale of our QM/MM simulation, our structural parameters may also not have reached equilibration.

[2] Unfortunately the stabilization associated to the addition of a Q unit starts with the fourth amino acid unit, (127)so cannot be investigated here: in fact the number of amino acids in our systems is never greater than four. This issue must be addressed in a further study.

for-mation energy of a strand constituted by n glutamines free from long-range effects (i.e. isolated not-interactive strand).

The stabilization energy per hydrogen bond (ΔE_{HB}) was then defined by dividing ΔE_N by the number of hydrogen bonds (n_{HB}) in each system.

$$\Delta E_{HB} = \Delta E_N / n_{HB} \tag{4.2}$$

ΔE_{HB} decreased nonlinearly with the number of strands (Fig. 4.6 and Tab. B.1): the variation of ΔE_{HB} in each series is ∼0.8 kcal/mol passing from two-strands systems to four-strands systems. This quantity is smaller than the typical DFT-PBE error.(249) However, here we consider differences of energies in similar systems; thus we can reasonably assume that fortuitous error cancellations errors may increase the accuracy of our calculations. ΔE_{HB} ranged from -5.0 kcal/mol in the smallest system to - 6.5 kcal/mol in the larger system (4x4), suggesting that a CE effect exists and that for the present systems this is at maximum of 1.5 kcal/mol per HBs.

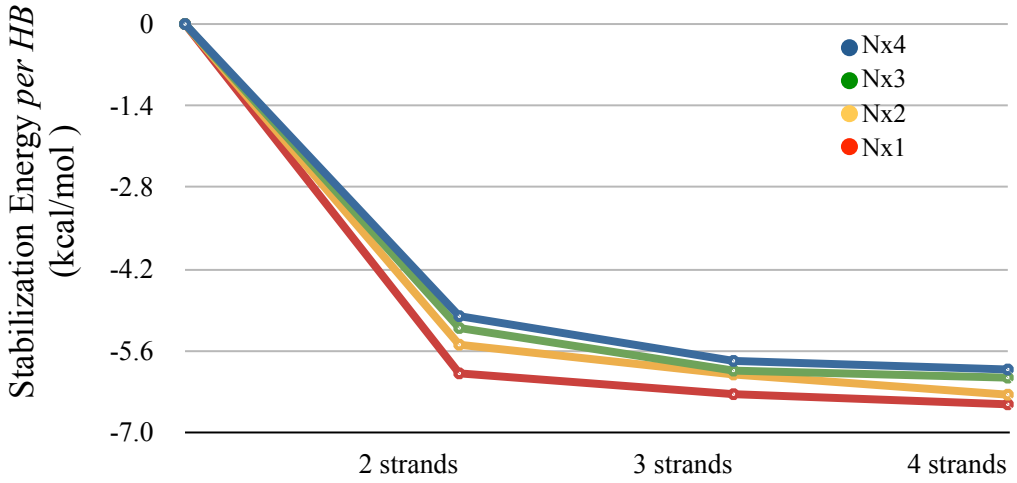


Figure 4.6: Stabilization energy per hydrogen bond (ΔE_H) for the addition of an Nth Q strand to the Q_{N-1} - The gradual change of ΔE_H versus the number of strands showed that the strength of the HBs between layers increases nonlinearly with the number of strands.

As expected the stabilization energy depending on CE is smaller for polyA systems with respect the polyQ; clearly for the absence of side chain HBs stabilization. According to this hypothesis, if we compute the CE for the $Nx3_{SC}$ series, where the glutamine

side chains are not able to form HBs, we find result comparable with polyA ones. (Tab. B.7)

In summary we found that: 1) both parallel and perpendicular CEs affect the geometry of polyQ β-strands because of the key role of the Q side chains. 2) The formation of cooperative hydrogen bonds stabilized multiple polyQ β-sheet strands with respect to a single iso-lated strand. 3) Within the limitations of the calculations on a single β-stranded structure in water solution, we suggest that environmental effects on hydrogen bonding CE affects only the magnitude of CE, while the qualitative trend is the same as that found in the *in vacuo* calculation.

Acknowledgements. Annalisa Pastore acknowledges funding from MRC grant No U117584256).

5

Conformational ensemble of Huntingtin N-term in aqueous solution explored by atomistic simulations

The 17-amino-acid N-terminal sequence (N17) of the huntingtin protein, plays a crucial role in its aggregation mechanism.N17 has been intensively investigated experimentally. The peptide is mainly unstructured at room temperature, and the conformational states it visits have not yet been characterized. Here we predict the free energy landscape of N17 in aqeuous solution by using bias-exchange metadynamics, together with an all atom description in explicit solvent. N17 turns out to populate four main kinetic basins, interconverting on the microsecond time-scale. The most populated basin (about 75%) is a random coil, with an extended flat exposed hydrophobic surface. This , might create a hydrophobic seed around which flanking polyQ tracts may collapse. In addition, it may also promote hydrophobic-force driven associations between Htt N-terminal fragments. The other basins contain helical conformations, which could facilitate the binding on the target surface of N17. The detailed structural characterization of N17 might help further investigations of N17 binding to its cellular partners and its effect on huntingtin aggregation.

5.1 Introduction

Huntington disease (HD) is an autosomal-dominant neurodegenerative disorder, for which there is no cure (142; 195). It is caused by expanded CAG trinucleotide repeats in the gene that encodes the large ($\sim$ 3500 aa) protein Huntingtin (Htt). The resulting mutant, with an extended polyQ tract in the N-terminal region, interacts abnormally with other proteins leading to neuronal dysfunction (38; 75; 108; 271).

Recently, *in vitro* (136; 155; 299; 328), *in vivo* (3; 15; 86; 207; 257; 306) and *in silico studies*(176) showed that the N-terminal 17 amino acids fragment (Sequence: MATLEKLMKAFESLKSF - N17 hereafter), cause a increase of polyQ fibrillation. The *in vivo* studies suggest that this effect might arise by a variety of mechanisms. N17 could alter subcellular localization (306). It could nucleate aggregation (136; 176; 273; 298; 328). It might also modulate aggregation by interacting with cellular partners (115), including the SUMO protein (284), an immunglobulin fragment (63), F-actin (10) and chaperones (294).

Understanding these mechanisms would greatly benefit from structural information.

A wealth of experimental biophysical techniques, including NMR (298), CD (15; 298; 328) and FRET (298) have established that N17 in aqueous solution does not possesses a unique, folded structure, adopting predominantly unfolded, random-coil conformations with transient helical conformation. Detailed information on the possible conformations present in solutions is so far lacking. Atomistic simulations are a powerful tool to predict the structural determinants and energetic of peptides in solution, in cases where experiments do not provide straightforward structural information. These approaches include molecular dynamics (MD) simulations running on tailored machines or on massive collective calculations initiatives (such as folding@home, http://folding.stanford.edu/). Alternatively, in the case that the free energy may be described in terms of few collective variables (CVs), different techniques can be used (reviewed in (61; 79; 184)). Here, we characterize the thermodynamics and the kinetics of N17 at room temperature using Bias Exchange Metadynamics (BE), which has been already applied to similar problems (200; 247; 248).

5.2 Results and Discussion

Our calculations suggest that there are four kinetic basins (or metastable states) at 300 K. Each of them is characterized by a population and by an attractor, which is defined as the cluster with lower free-energy in the basin. The clusters belongs to each basins are not necessarily identical in structure, but are kinetically connected, namely the transition time between two clusters belonging to the same basins is typically much smaller than the one between clusters belonging to different basins (200).

The conformers of N7 are here analyzed in terms of basins. The far most populated state is basin B2 (75%), followed by B1 and B3 (11% and 10%, respectively) and B4 (4%) (Fig. 5.1).

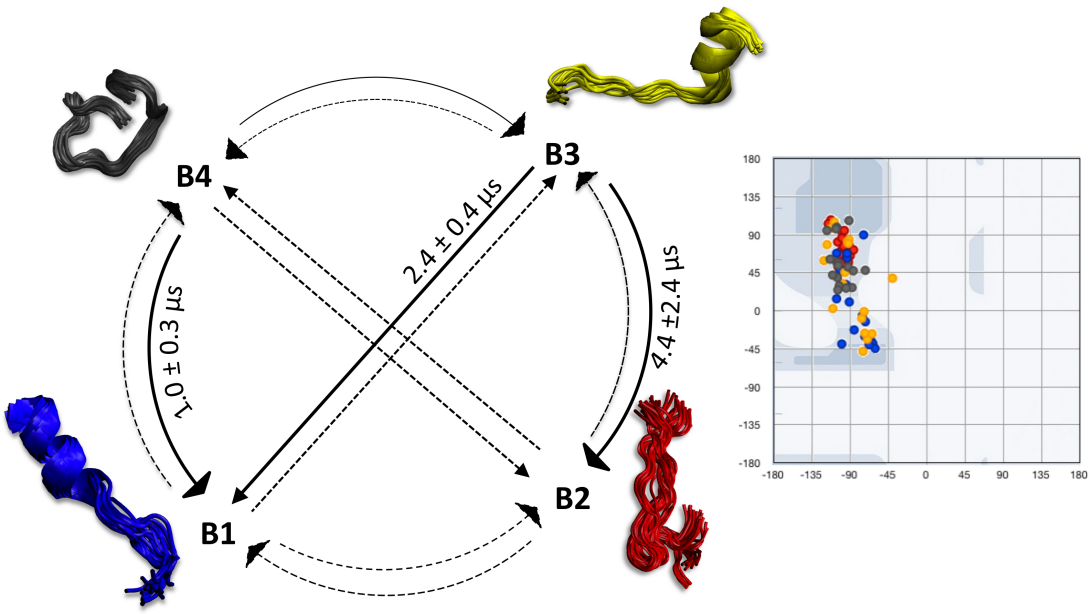


Figure 5.1: N17 Basins - Four basin (B1, red, B2, blue, B3, yellow, B4, grey) of N17 in aqueous solution. Each of them is characterized by a population and by an attractor, which is defined as the cluster with lower free-energy in the basin. The clusters belonging to each basins are not necessarily structurally similar, but are kinetically connected, namely the transition time between two clusters belonging to the same basins is typically much smaller than the one between clusters belonging to different basins (32; 131; 200). Because of the large number of clusters we show here the backbone structures only of the attractors for the sake of clarity. The calculated interconvertion rates between the kinetic basins are reported. Dotted lines used for rates ¿ 2 μs. The statistical error on the transition times between states is shown in figure. The kinetic basins differ in terms of secondary structure content, as also evident from the Ramachandran Plot in the inset.

B1 and B3 include structures with a relatively high helical content. However, the structure corresponding to a fully formed alpha helix is rather high in free energy (28 kJ/mol): in fact, residues 11 to 17 for B1 and 7 to 17 for B3 are preferentially coils. B1 is more compact than B3, as evident from a calculation of its gyration radius and its Solvent Accessible Surface Area (SASA, Tab. 5.1). However, it has a similar phobic area (Tab. 5.1). Hence, the conformation of the hydrophobic side chains in B1

should differs significantly from that of B3, within the same helical fold: indeed, the hydrophobic side chains of the B1 attractor are located on one side of the helix, whilst in the B3 attractor they are spread around all the helix (Fig. 5.2). The most occupied basin at 300 K is, according to our simulation, basin B2, which includes extended-coiled and highly solvated structures. Notably, the B2 attractor presents a well-ordered organization of the hydrophobic side chains only on one side of the structure like B1. B4 is constituted by globular compact coiled structures. It is clearly distinguished from B2 from a significantly lower gyration radius and a smaller SASA.

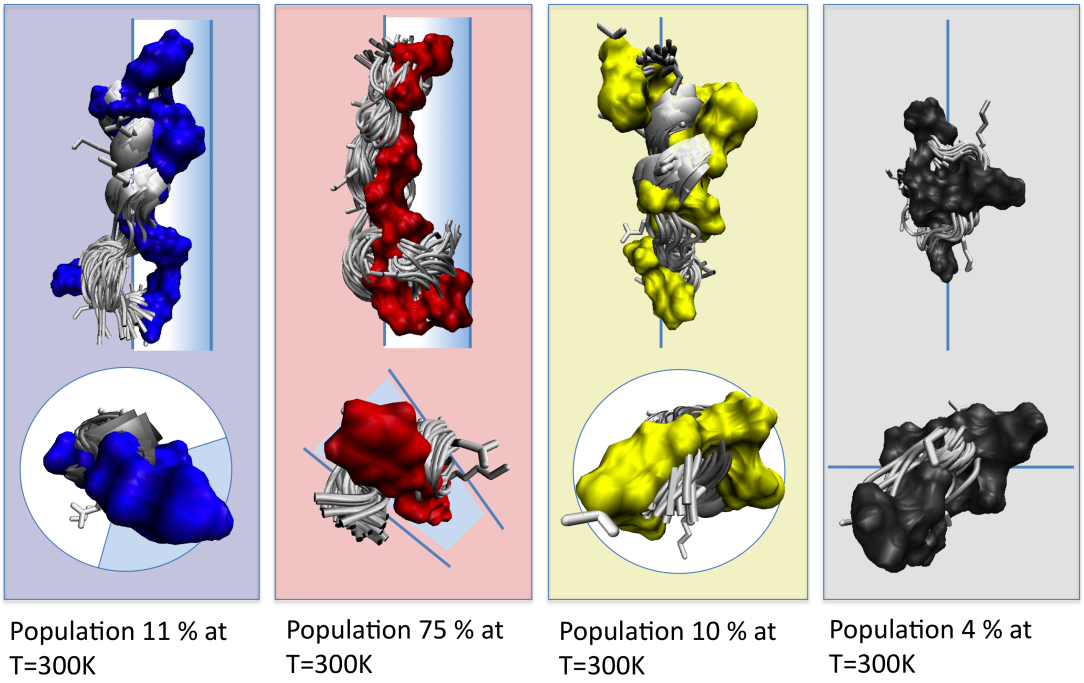


Figure 5.2: Hydrophobic Side Chain Distribution - The hydrophobic side chain distribution of N17's four attractors is shown.

Next, we check the consistency of our findings with available NMR (298) (i), CD (15; 298; 328) (ii) and FRET (298) (iii) derived structural determinant.

To make a meaningful comparison, we have first to assess on which time scale the conformations interconvert into each other. If the interconversion time is small with respect to experimental time resolution, the properties are measured as an average over the four conformations. Otherwise, it is more appropriate computing the properties of each basin separately. Here we have estimated the interconversion kinetics constants

(Fig. 5.1) following ref. (222).

Basin	Populatio (%)	Gyration Radius (nm)	SASA (nm^2)	Phobic Area (nm^2)	Phylic Area (nm^2)	End-to-End distance (nm)
B1	11	0.83	27.3	12.8	14.5	1.8
B2	75	1.08	29.5	13.3	16.2	2.5
B3	10	0.96	28.3	12.9	15.4	1.1
B4	4	0.84	27.9	12.6	15.3	2.2

Table 5.1: Selected properties of the four basins

We find that all of them are in the 1 to 200 μs timescale. This is at least 1 order of magnitude smaller than that time-scale typical of NMR and CD experiments, usually in the ms. It is instead 3-4 orders of magnitude larger than the timescale of FRET experiments. The NMR and CD derived structural determinants are hence calculated as averages, whilst those derived by FRET-derived are calculated for each single basin.

(i) The bidimensional proton TOCSY and NOESY spectra show that N17 adopts a predominantly unfolded, random-coil conformations (298). Our calculations are consistent with this finding, as our most populated cluster is in a random coil conformation. In addition, the NOESY spectrum (298) shows a cross-peak (i+2) connecting the Thr3-Ha and Glu5-HN protons in the weak-range, indicating that the two atoms are at a distance d between 0.4-0.5 nm. This points to the transient existence of a few residues in the α-helix conformation (298). Our calculations are also consistent with this finding: the calculated d is 0.49 nm, close to the α-helix characteristic value (0.44 nm). (ii) CD studies (15; 298; 328) provided approximate estimate of the helical content (HC) of the peptide, which turns out to be HC $\cong$ 37%. Our calculated percentage of helical content is in fair agreement to this value (HC $\sim$ 29%).

(iii) FRET experiments provide an approximate estimate of N17's end-to-end distance of 2.40 $\pm$ 0.05 nm (298). Again, the calculations are in fair agreement with this finding (Tab. 5.1). The value far most populated basin (B2) is close to the experimental value.

5.3 Conclusions

Determining the conformation that N17 assumes in solution may help understand the mechanism by which N17 modulates polyQ aggregation in Huntington disease. Our

free energy calculations suggest that N17 is present in four main kinetic basins, interconverting with each-others in the microsecond time-scale. Our findings are fully consistent with a wealth of experimental biophysical data. The main populated state (75%) is a random coil, with extended flat exposed hydrophobic surface. This may increase oligomerization properties of N17 because could both create a hydrophobic seed around which flanking polyQ tract may collapse (263) and promote hydrophobic-force driven associations between Htt N-terminal fragments (298). The other significantly populated basins assume a helical conformation from residues from 1 to 10. The latter could facilitate the binding on N17's target surface. This is in agreement with the proposed binding conformation for N17 to a variety of cellular partners (10; 63; 284; 294) that is helical one.

5.4 Computational Details

As the initial configuration of N17 we have taken its extended coil conformation build with Modeller 9v8 (293). E, K residues were considered in their charged state. The peptide was inserted into a cubic box (box vector 7.18 nm) of $\sim$ 4100 water molecules. 1 Cl- ion was added so as to achieve electroneutrality to the system. Periodic boundary conditions were applied. The AMBER(parm99) (50; 320), Aqvist (1), and TIP3P force field (1), was used for the protein, the counter ions, and water, respectively. Long-range electrostatic interactions were treated with the particle mesh Ewald (PME) (71; 92) method, using a grid with a spacing of 0.12 nm, combined with a fourth-order cubic spline interpolation (107) to compute the potential and forces in between grid points. The cutoff radius for the real part of electrostatics, as well as that for the Lennard-Jones interactions, was set to 0.9 nm. Simulations were performed in periodic boundary conditions in the NPT ensemble, with temperatures and pressure kept close to the desired value (T = 300 K, P = 1 bar) through the Nosé-Hoover (126; 220) and Andersen-Parrinello-Rahman (219; 234) coupling schemes respectively. The LINCS algorithm (120) was used to constrain all bond lengths involving hydrogen atoms and the time-step used was 2 fs. The system was first energy-minimized imposing harmonic position restraints of 1000 Kj/(mol*nm2) on solute atoms. This allowed the solvent to equilibrate without distorting the solute structure. After 200 ps of energy minimization without restrains, the N17 was gradually heated from 0 to 300 K. N17 was gradually

heated from 0 to 300 K. The temperature was increased by 25 K every 100 ps of MD. After the latter procedure, the entire system underwent to other 200 ps of energy minimization and finally to MD The simulations were performed with the GROMACS software package (28). g_sas was used for computing the solvent accessible surface area with a probe radius of 0.14 nm. g_rama an g_gyrate package where used for calculating Ramachandran angles and radius of gyration respectively. The average value of an observable O (Gyration Radius, SASA, etc) is calculated by using the estimated free energies, as

$< O >= \sum_i (O_i * exp(-F_i/T))/\sum_i (exp(-F_i/T))$.

The free energy as a function of six dimentionless collective variable (CVs) was calculated using Bias Exchange Metadynamics (247) (A brief summary of the basic principles of the method is offered in the SI). The CVs are: i) CV_1 counts number of $C_\gamma$contacts (hydrophobic contacts), CV_2 counts the number of $C_\alpha$contacts, CV_3 counts number of backbone hydrogen bonds, CV_5, CV_6 is the dihedral correlation between successive Ψ dihedrals. The Gaussian widths for these CVs were 3.0, 6.0, 3.0, 0.6, 0.6 and 0.6 respectively. These parameters have been optimized using the technique in (66).

The total bias simulation time was 240 ns (40 ns for replica). Convergence was reached after 12 ns in each replica.

Using the approach of (200), the CV space is divided in a grid of clusters. The free energy of each cluster is estimated by a weighted-hystogram approach (169), using the thermodynamic model described in ref (200). The transition rates between each cluster are valuated introducing the kinetic model described in (200), where a Markovian diffusive behavior is assumed (see SI) (32; 131).

Both the thermodynamic and the kinetic model require the minimal number of independent CVs able to describe with a good statistic the behaviors of the clusters. For our system, the numbers of such CVs turns out to be 4, namely CV_2, CV_4, CV_5, CV_6 (see SI). The set of clusters was thus defined by partitioning this 4 dimensional CV space in small hype-rectangles of sizes 9.2, 1.19, 0.6 and 1.27 respectively for each CV.

6

Actin binding by Htt blocks intracellular aggregation.

The Huntington neurodegenerative disease is associated by extended polyQ tracts in the protein Huntingtin. The first seventeen aminoacids of the N-term region (N17) modulates aggregation and toxicity of the mutated form of the protein (Mut-Htt). Because N17 might also mediate binding to F-actin, and modulation of F-actin cytoskeleton affects Mut-Htt aggregation, N17-actin interactions might also play a role for Htt aggregation. To address this issue, here we predicted the structural determinants of such interaction by biocomputing approaches. Then, we identified mutations that modulate actin binding. The effects of these mutations and of replacing N17 with actin binding domains were tested experimentally on intracellular aggregation of the Htt exon 1-CFP/YFP proteins. Taken together, our results corroborate the hypothesis that N17/F-actin binding might stabilize Mut-Htt.

6.1 Introduction

Huntington disease (HD) is an autosomal dominant polyglutamine (polyQ) disorder. It is caused by expanded CAG trinucleotide repeats in the gene that encodes the protein Huntingtin (Htt)(195). The resulting mutant (mut-Htt), with an extended polyQ tract, causes neurodegeneration(264). Proteolytic processing appears to play a key role in the progression of HD by releasing short N-terminal polyQ-containing fragments of 100-150 residues (194). Short Htt fragments form β-sheet-containing aggregates (56; 64; 75; 275; 276; 295), which are a hallmark of the disease (64; 275). Because the precise pathogenic

fragment is not known, we have chosen to study a peptide consisting of the first exon of Htt, termed Htt exon 1. For cellular studies, we have used cyan and yellow fluorescent protein fusions to Htt exon 1, termed Htt exon 1-CFP and Htt exon 1-YFP.

The first seventeen amino acids of the Htt amino terminus, termed N17, play a key role in its aggregation and toxicity (63; 155; 176; 257; 298; 328). N17 controls subcellular localization, aggregation, and cytotoxicity of Htt exon 1 fragments in mammalian or yeast cells(15; 257). Multiple roles for N17 have been suggested. It has been proposed to nucleate Htt aggregation (136; 298; 299); to regulate the type of aggregate that forms (36; 328); and to mediate binding to protein partners (115; 148) such as the chaperonin Tric (294), SUMO (284), and F-actin (278). Additionally intracellular antibody (intrabody) fragment binding to N17 decreases Htt exon1 aggregation, possibly by masking N17 (63). Some of us (10; 250) and others (21; 44; 207; 208; 289) have posited a role for actin as an influence on Htt aggregation. While deletion of N17 altogether reduces overall Htt aggregation (10), which may indicate a role for N17-mediated protein interactions in regulating protein misfolding, N17 deletion could also affect protein misfolding via effects on primary protein structure (69; 328). It remains unclear whether actin binding by Htt increases or decreases Htt aggregation within cells. In this study we have used molecular modeling combined with cellular approaches to investigate the role that an N17-mediated interaction with actin might play in regulating Htt aggregation.

6.2 Results

The N17 region and the WH2 proteins (235) (a family of actin binding proteins characterized by the WiskottAldrich syndrome protein (WASP)-homology domain-2 (235)) contain elements of similarity and identity. We used existing crystal structures of the WH2/actin interaction to create a model for how N17 might mediate Htt binding to F-actin (Fig. 6.1). Based on this model, we predicted mutations in N17 that should decrease actin binding. To predict mutations, which should increase N17/actin affinity, we chose amino acid substitutions that increase sequence and/or structural similarity with the WH2 proteins (Fig. 6.2). We subsequently determined the effects of each of these amino acid substitutions on intracellular aggregation of Htt exon 1 (Fig. 6.2). Fi-

nally, we determined the effect on Htt intracellular aggregation of complete substitution of the WH2 and Lifeact F-actin binding domains for the N17 region (Fig. 6.3).

6.2.1 Structural model of the F-actin/N17 complex.

Suitable templates to predict structural determinants of N17/ actin complexes are WH2 proteins/ actin complex X-ray structures (Tab. 6.1 in SM). Indeed, N17 and WH2 proteins are similar in primary structure: the sequence similarity (SS) between Homo sapiens WH2 proteins and N17 is as high as ∼63% (Fig. 6.2 A). Second, they may share similar secondary structure and binding modes. The WH2 proteins feature an amphipathic[1] N-terminal α-helix when bound to actin in the fully conserved, mostly hydrophobic binding site (16; 59; 82; 114; 125; 133; 156; 191; 235; 253; 302).[2]

N17 can adopt an α-helix fold (90% of probability, with AGADIR server (171)) with an amphipathic distribution of the residues (see hydropathy plot, Section 2 in SM), consistently with the proposed binding conformation to a variety of cellular partners (15; 63; 155; 294; 298). Third, one of them, the N-WASP protein, has been linked to Htt exon1 aggregation (207; 208).

In our modeled structure, N17 residues M8, F11, F17, E12, S13, and S16 bind to hydrophobic residues of the binding site (Fig. 6.1 and Tab. 6.1). In addition, K9, K15, and S16 form H-bonds with T133 and M355 backbone (Fig. 6.1). This binding mode is fairly similar to that of the WH2 proteins. Some differences are to be expected because of the presence of a larger number of polar residues in N17 than in WH2 (Fig. 6.1 and Fig. 6.2 A).

6.2.2 Predicting mutations affecting F-actin/N17 interaction.

According to the F-actin/N17 model mutations predicted to disrupt N17 H-bond interactions with F-actin should inhibit the binding reducing the affinity for F-actin. These include: (i) K9A, K15A, K15C, S16A, E12T. On the other hand, increasing the SS with WH2 proteins, for which binding in cell has been demonstrated (16; 133; 156;

[1] An amphipathic molecule contains both polar (water-soluble) and nonpolar (not water-soluble) regions.

[2] The hydrophobic binding site is made of residues Y143, A144, G146, T148, G168, I341, I345, L346, L349, T351, M355. It is composed by a cleft (between domains 1 and 3) and a contiguous pocket at the front end of the cleft (82). The hydrophobic pocket is solvent accessible both in F- and G-actin (82) (Section 3 in SM).

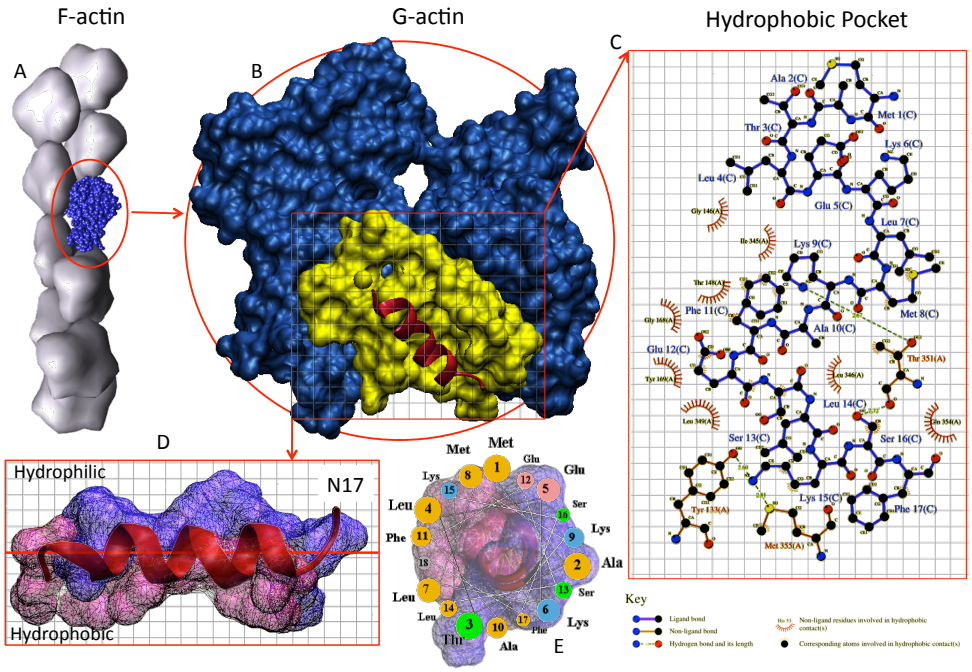


Figure 6.1: Model of F-actin/N17. - **(A)** Cartoon representing a monomer of F-actin (blue) (see SM for modeling details). **(B)** monomer of F-actin (blue) in complex with N17 (red cartoon). The hydrophobic binding site is represented as a yellow surface. **(C)** Scheme of N17/AHP interactions. **(D)** The orientation of hydrophobic and hydrophilic residues of N17. **(E)** N17 helical wheel.

Mutation	Model	SS to WH2	Helix fold propensity	Other Experimental Validation	Observed Aggregation Effects
T3A	Stabilize helix fold	Increase	Increase	"Prevent the phosphorylation of T3 (3)"	Increase
L7S	Stabilize helix fold	Increase	Increase		Decrease
K9A	Disrupts HB between Lys 9 in N17 and Thr 351 in F-actin	Decrease	Decrease		Increase
F11G	Replace hydrophobic interactions	Increase	Increase		Decrease
E12T	Disrupt hydrophobic interaction with L346, L349, T351	Decrease	Decrease		Increase
S13A/S16A	"Disrupt HB Between S16 and T351"	Decrease	Decrease	"Mimic phosphorylation at S13 and S16 changes Htt aggregation(109)"	Increase
S13D/S16D	SB interaction with K6 and K9	Equivalent	Increase	"Prevent phosphorylation at S13 and S16 changes Htt(109)"	Decrease
K15C	Disrupt bifurcated HB between Lys 15 in N17 and T133 and M355 in F-actin	Decrease	Decrease		Increase
K15A	Disrupt bifurcated HB between K15 in N17 and T133 and M355 in F-actin	Decrease	Decrease		Increase
F17V	Replace hydrophobic interactions	Increase	Equivalent		Decrease

Table 6.1: Mutation of N17 and their effect on aggregation. - Comparisons with experimental available data are also included. HB=hydrogen bond; SB=salt bridge).

191; 302),are instead expected to increase the affinity for F-actin. Mutations which increase SS include T3A, F11G, F17V and L7S. Finally, mutations which stabilize the N-terminal helix are expected to bind more tightly to F-actin, as APBs bind with such fold. This is the case of the S13D mutation, which stabilizes the N17 helix fold, due to D salt bridge with K9.

6.2.3 Cell essays

To evaluate the effect of various point mutations in the Htt exon 1 protein, we used a combination of fluorescence imaging to discern effects on inclusion formation, and an intracellular aggregation assay based on fluorescence resonance energy transfer (FRET)(250). These approaches have now been validated in our prior work (10; 80; 250; 278). Mutants predicted to increase binding affinity between N17 and actin reduced aggregation. Conversely mutants predicted to decrease binding affinity, increased aggregation (Fig. 6.2).[1] These results were consistent with prior studies that indicate that mutation S13/16A and S13/16D, increase and decrease aggregation, respectively(109).

Our previous results support the idea that the N17 domain reduces Htt aggregation by binding to actin, although we cannot rule out its binding to other cellular structures. To directly test whether actin binding would reduce Htt aggregation, we created a chimeric protein in which we replaced the N17 region of Htt exon1 with two domains for which the direct interaction with F-actin is established: a 17aa domain (from WH2 proteins) that binds the hydrophobic pocket of actin(2; 41; 58; 59; 83; 119; 235; 277), and Lifeact, a 17 aa F-actin binding peptide used as an F-actin marker (256). These substitutions strongly reduced Htt exon 1 aggregation as measured by FRET and fluorescence imaging in HEK293 cells (Fig. 6.3).

Some consideration on the accuracy of our model. As in any model, these results are subject to caveats. Here, we consider only one of the possible orientations of the N17 helix (front-to-back). However the opposite orientation is also possible (back-to-front). According to our modeling procedure (see Methods session), the first creates more hydrophobic contacts and more hydrophobic binding events than the second (Table S2-S3 in SM). Hence, N17 is expected to bind more tightly than in the back-to-front orientation. In addition, this modeling suffers from the typical limitations

[1] T3A mutation experimentally shows conflicting results and hence it is not discussed here.

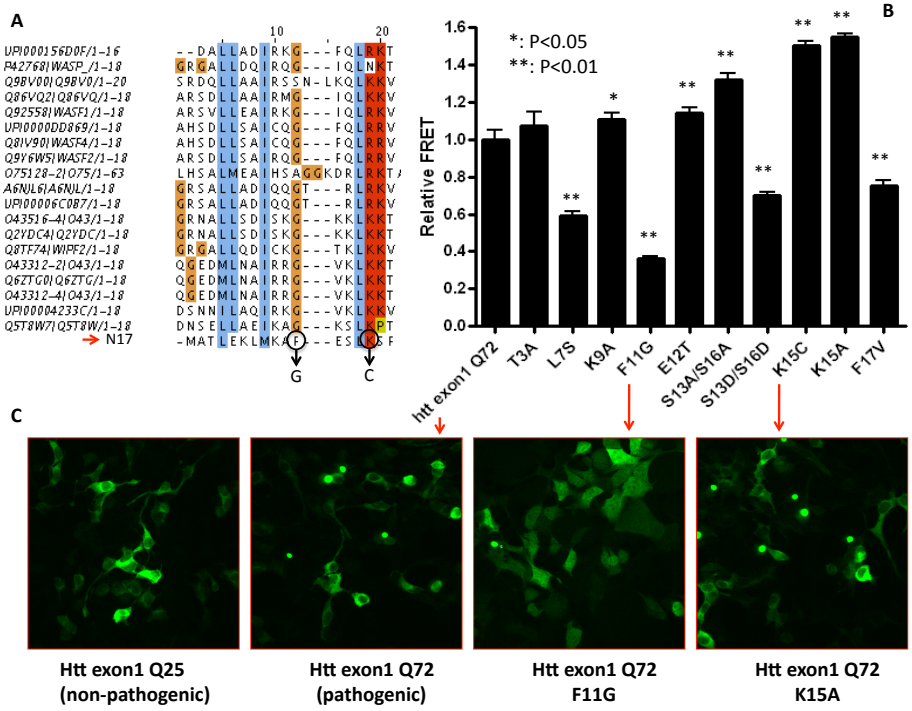


Figure 6.2: Experiment 1 - A) *Sequence Alignment between N17 and WH2 domains in Human Homo sapiens proteins (SS=63%).* Two of the suggested mutations in FRET and Fluorescence experiments are underlined. **B)** *Effects of mutations in N17 on Htt Aggregation.* HEK293 cells were transfected with either wild-type Htt exon 1 (Q72)-CFP/YFP or the indicated mutants in the N17 region. Relative aggregation vs. wild-type, expanded Htt was determined by calculating the relative FRET signal derived from Htt aggregation. Note that some mutations strongly increase aggregation (e.g. K15A), while others suppress aggregation (e.g. F11G). **C)** *Representative fluorescence images of cells expressing various Htt constructs.* HEK293 cells were transfected with either wild-type Htt exon 1-YFP or the indicated mutants in the N17 region. Images were taken by fluorescence microscopy after 48h in culture. Htt exon 1 (Q25)-YFP does not form inclusions, whereas the expanded form (Q72) forms cytoplasm inclusions. The K15A mutant increases inclusion formation, whereas the F11G mutant blocks aggregation. These images are representative of the cells studied in the FRET analysis.

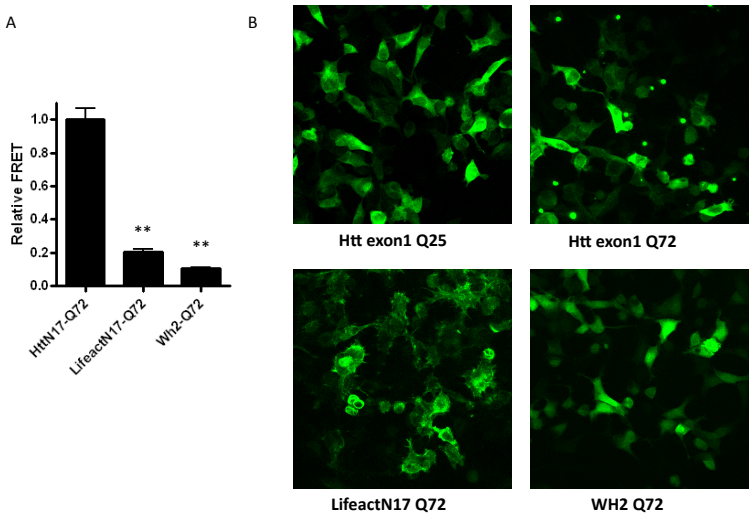


Figure 6.3: Experiment 2 - A) *Substitution of the N17 region with a peptide actin binding sequences reduces Htt exon 1 aggregation.* The N17 region of Htt exon 1was removed and replaced with the 17aa lifeact peptide sequence (which binds F-actin), or the 17aa WH2 domain (which binds both G and F actin). HEK293 cells were transfected either with wt Htt exon 1(Q72)-CFP/YFP or with the indicated chimeric proteins. Substitution with either the Lifeact sequence or the WH2 domain strongly reduced Htt aggregation, as measured by FRET. **B)** *Representative images of cells expressing chimeric Htt constructs.* HEK293 cells were transfected with the indicated Htt exon 1-YFP constructs, and imaged by fluorescence microscopy. Htt exon 1 (Q25) remains diffusely distributed, whereas Htt exon 1 (Q72) forms intracellular inclusions. Lifeact-Htt localizes to the actin cytoskeleton, but does not form inclusions. WH2-Htt is diffusely distributed, and does not form inclusions either. These images are representative of the cells used to obtain the FRET measurements in Fig. 6.3 A.

of rigid protein modeling (201). Finally, our model does not take into account the recent claim that WH2 proteins are incompatible to F-actin filament in crystallization studies (254), because of two reasons. First, such claim is mainly based on the superposition of a loop of F-actin (the so-called D-loop) on the cleft, which would hamper ligand binding. However, a simple analysis of the loop conformations based on the available structural information (See Section 4 in SM) shows that this loop is highly flexible, and its conformation is dictated by the type of ligand bound the cleft. So this claim appears to be unjustified. Second, our model is made for in cell conditions. These are expected to be different from those of the X-ray diffraction experiments. Indeed, WH2/F-actin binding in these conditions have been demonstrated (16; 133; 156; 191; 302).

6.3 Discussion

The protein context surrounding the polyQ repeat is important for modulating its aggregation potential and toxicity (306). In particular, N17 has a critical influence on Htt aggregation and toxicity (3; 15; 257; 284; 298). Prior work has suggested that the N17 region of Htt might mediate binding to F-actin at a $\sim$ 1M affinity (10), although this has been difficult to quantify precisely. However, the role of an actin interaction in Htt aggregation has been somewhat uncertain. Based on existing structural information regarding the binding of the WH2 proteins with the hydrophobic pocket of G-actin, we created a molecular model to identify 10 mutations in N17 that are predicted to modulate actin binding. We then directly tested the effects of these mutations on intracellular aggregation of the Htt exon 1-CFP/YFP proteins. We found that all mutations predicted to weaken the interaction of Htt exon 1 with actin increased intracellular aggregation, whereas those predicted to increase Htt/actin affinity uniformly reduced intracellular aggregation. Finally, substitution of bona fide actin binding domains (WH2 and Lifeact) each dramatically reduced Htt exon 1 intracellular aggregation. These results are consistent with a model that F-actin binding could serve to stabilize the expanded form of the Htt protein, and can help explain why modulation of the F-actin cytoskeleton has been observed to have profound effects on intracellular aggregation of Htt.

6.4 Methods

6.4.1 Modeling.

The structure of N17/actin was predicted by homology modeling in several steps (see SI for details):

1. Identification of to ABPs homologous to N17, using BLAST(4).

2. Alignment between N17 and Homo Sapiens ABPs sequences using CLUSTALW (300). The proteins which exhibit the largest sequence similarity are the class of WH2 proteins (235) (63%) Instead, it is less of $\sim$ 36% with other actin binding protein in the same hydrophobic region (62) (Fig S5 SM). Similar trend is found when aligning the sequences of all eukaryotic species (See Material Online). Structural information of the complexes with actin of members of both the class and the family is available.

3. Structural prediction of the actin monomer (A) and F-actin (B) in complex with N17. (A) is based on structural information of the WH2 proteins presented in Tab. D.1 in SM (PDBID 2A3Z, 2A40, 2A41 from (59), 2D1K from (181), and 2VCP (unpublished). The superposition of the target sequence (N17) on a template structure (WH2) implies the calculation for of an 'average structure' or 'framework' (283; 292). The templates are averaged into the framework using weights corresponding to their similarity with target sequence (283). Among the WH2 complexes, the largest weight has been assigned here to N-WASP/Actin complex (PDBID 2VCP) because it is the only one from Homo Sapiens and because N-WASP is the only WH2 protein linked to Htt exon1 aggregation (207; 208). B) is based on the structural determinants of 2VCP actin and 3BYH-based model of F-actin (100) (see SM). We used the Modeller 9v.6 package (95; 201; 269; 293).

6.4.2 Experimental methods:

1. Plasmid construction: Plasmids Htt exon1 Q25 CFP/YFP and Htt exon1 Q72 CFP/YFP were described previously (250). A series of mutants in the N17 region of Htt exon1 Q72 CFP/YFP were generated using the QuikChange^{TM} Site-Directed Mutagenesis Kit from Strategene. The Htt exon1 N17 region (MATLE

KLMKAFESLKSF) was also replaced by lifeactN17 (MGVADLIKKFESISKEE) or a WH2 domain (MRDALLDQIRQGIQLKSV), and named as LifeactN17 Q72 CFP/YFP and WH2 Q72 CFP/YFP, respectively.

2. Cell culture and transfection: HEK293 cells were cultured in Dulbecco's modified Eagle's medium-5% fetal bovine serum (FBS) and transfected with lipofectamineTM reagent. For FRET assay, 24 hours post trasfection, cells were passaged onto a 96-well black plate, cultured for another 24 hours, and then fixed with 4% paraformaldehyde. For microscopic imaging, 24 hours post transfection, cells were fixed with 4% paraformaldehyde. Fluoresence images of cells were taken using a confocal image system from Zeiss.

3. FRET assay: FRET assay was performed as described previously (Desai, 2006). The relative FRET of all mutants or fused constructs were generated by normalizing to the FRET of Htt exon1 (Q72)-CFP/YFP, which was expressed as 1.0. The FRET assay was performed using a fluorescence plate reader (TECAN, Infinite M1000).

6. ACTIN BINDING BY HTT BLOCKS INTRACELLULAR AGGREGATION.

7

Conclusion

The first exon of the huntingtin protein (*Htt Exon 1* in Fig. 1.1) contains the polyQ region. It is implicated in Huntington disease (75; 196). It can form aggregates similar to those observed in the neurons of afflicted patients (75; 196). It is therefore very important to get insights on the structure of *Htt Exon 1* and of its complexes with cellular partners (306).

In this thesis, I have investigated by computational methods two relevant structural aspects of *Htt Exon 1*.

First, I have studied the effect of cooperativity on the polyQ expansion. The length of the polyQ region in *Htt Exon 1* is inversely correlated to the age of onset of symptoms (112; 259). Although neither the Huntingtons disease onset mechanism nor the structural properties of polyQ aggregates are know, the similarity between the glutamine side chain and the amino acids backbone attributes a crucial role to the formation of intramolecular hydrogen bonds (HB) in polyQ aggregates (241). Hence, a characteristic feature of polyQ repeats is that they can form β-sheets stabilized by HBs not only between backbone atoms, but also between atoms of the side chains. By performing molecular dynamics calculations, I have further corroborated the hypothesis (241) that the HB content plays a major role in the structural stability of polyQ systems. Indeed, on passing from large monomeric systems to oligomeric ones, PolyQs tracts always tried to establish the higher number of HBs between both side-chain and backbone atoms, independently of the β-sheet content and the number of Qs in each structure. Furthermore, quantum-mechanical calculations point to the fact that, as expected, the HB network of polyQ β-sheets is associated with a CE other than that of the amino

acids (193). According to our calculations, side chain HBs affect the cooperativity in both the directions perpendicular and parallel to the backbone (Fig. 7.1). The presence of such extra-stabilization may help explain some of the unique properties of polyQ. These are (i) their aggregation propensity and the ability of creating tighter interactions, which increase by increasing the number of Qs, first suggested by Perutz in 1994 (241). (ii) The propensity to create high stable HB networks, which may provide possible explanation for the known PolyQ ability in sequestrating transcriptor factors rich in polyQ(335).

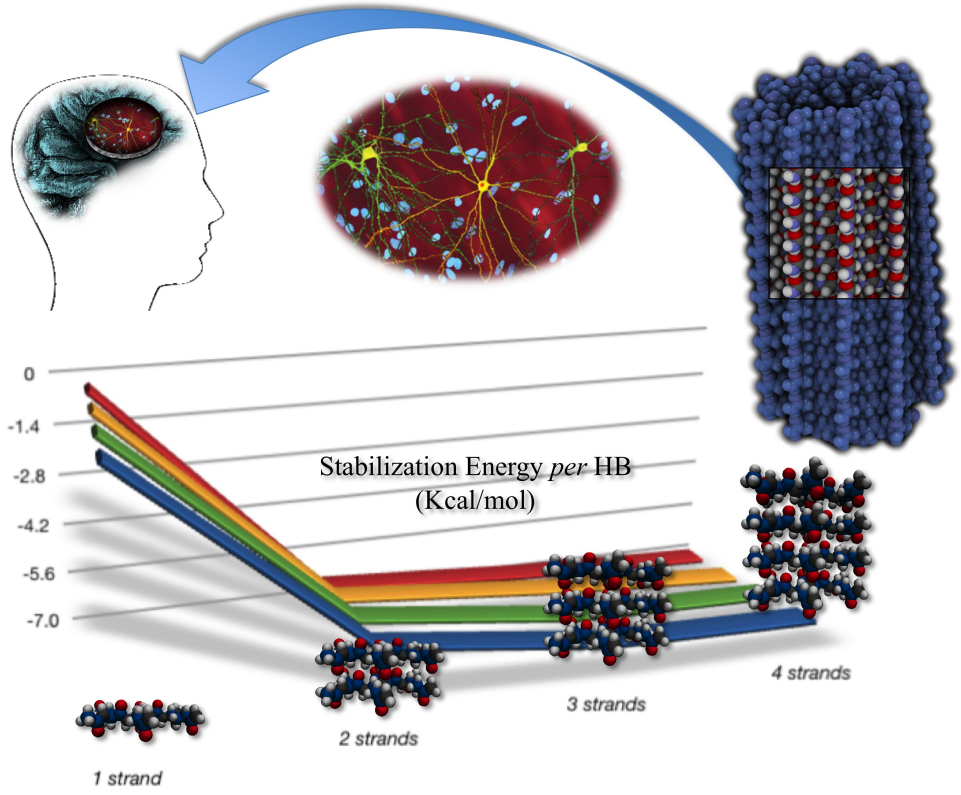


Figure 7.1: Cooperative Effect of PolyQ - Stabilitation Energy *per* HB: the strength of the HBs between layers increases nonlinearly with the number of strands, according to our calculations.

Next, I have investigated the structure of one of the *Htt Exon 1* flanking regions. This is the first seventeen amino acids of the N-term region (N17). N17 modulates aggregation and toxicity of *Htt Exon1* (306). First, I have predicted the conformational properties of the peptide in solution. The peptide is mainly unstructured at room

temperature. However, the conformational states visited by N17 have not yet been characterized. I have predicted the free energy landscape of N17 in aqueous solution by using computational methods, together with an all atom description in explicit solvent. N17 turns out to populate four main conformational ensembles, interconverting on the microsecond time-scale. The most populated ensemble (about 75%) is a random coil, with an extended flat exposed hydrophobic surface. At the speculative level, this might create a hydrophobic seed around which flanking polyQ tracts may collapse. In addition, it may also promote hydrophobic-force driven associations between Htt N-terminal fragments. The other ensembles contain helical conformations. The results are in accord with experiments (15; 298).

The detailed structural characterization of N17 alone helps investigate N17 interactions to one of its cellular partners, F-actin. Because a modulation of F-actin cytoskeleton affects Mut-Htt aggregation, N17-actin interactions might also play a role for Htt aggregation. This study has been carried out in collaboration the neurobiology Lab of Prof. M. Diamond (Washington University School of Medicine).

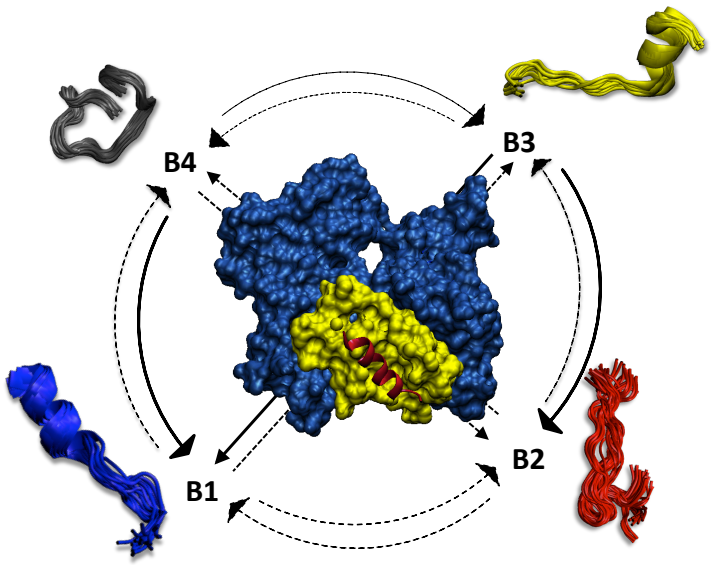


Figure 7.2: N17 - Conformations visited by N17 in solution (B1-B4). The helical conformation bind to the actin surface, according to our model (structure in the middle of the figure)

To address this issue, here we predicted the structural determinants of such inter-

action by biocomputing approaches. Only the helical conformation among the four found in aqueous solution turns out to be compatible with the binding to actin. The correspondent model is shown in Fig. 7.2. Then, I identified mutations that modulate actin binding. The effects of these mutations and of the replacement of the entire N17 with actin binding domains were tested experimentally by Prof. Diamond's Lab in HEK293 cells using procedure described in (80). A combination of fluorescence imaging to discern effects on inclusion formation, and an intracellular aggregation assay based on fluorescence resonance energy transfer (FRET) was used (152). Taken together, our results corroborate the hypothesis that N17/F-actin binding might stabilize *Htt Exon 1*.

8

Materials & methods

8.1 Introduction

"Certainly no subject or field is making more progress on so many fronts at the present moment, than biology, and if we were to name the most powerful assumption of all, which leads one on and on in an attempt to understand life, it is that all things are made of atoms, and that everything that living things do can be understood in terms of the jigglings and wigglings of atoms."

Richard P. Feynman from *Six easy pieces* 1963

Biomolecular modeling is a fertile and growing area, with exciting opportunities for molecular biophysics. Its application required a deep understanding of the tremendous complexity of the system of interest. Vast amounts of data are being provided by large-scale research efforts in genomics, proteomics, glycomics and structural biology. The challenge for computational techniques is to help in efforts to investigate features not directly accessible to experiments (309). Whereas it is indeed possible to take "still snapshots" of crystal structures and probe features of the motion of molecules through NMR, no current experimental technique allows access to all the biologically relevant time scales of motion with atomic resolution (309). Increasingly, computer simulations of biological macromolecules are helping to meet this challenge. Simulations based on fundamental physics offer the potential of filling-in these crucial 'gaps', modeling how proteins and other biomolecules move, fluctuate, interact, react and function. Improvements in computer hardware continue to deliver more computational power, which, when combined with theoretical and algorithmic developments, have led to an increasing range and depth of applications of molecular modeling and molecular dynamics in biology.

8.2 Homology Modeling

Knowledge of the three-dimensional structure of proteins is a prerequisite for molecular dynamic simulations. Only two spectroscopic techniques, nuclear magnetic resonance (NMR) and X-ray, can produce high-resolution three-dimensional coordinates of macromolecules. Most other spectroscopic techniques either add information to such three-dimensional coordinates, or require these coordinates for detailed interpretation of their results. NMR and X-ray are very elaborate techniques, and worldwide only about

30 protein structures are solved per day, wilts about 50 sequences per ten minutes were determined and deposited (315). Consequently, the necessity for homology modelling is tremendously increasing.

In its most elementary form, *homology modelling involves calculating the structure of a protein for which only the sequence is known using its alignment with a homologous protein for which the structure is known.*

Homology model is based on a major observation: During evolution, the structure is more stable and changes much slower than the associated sequence; hence similar sequences adopt practically identical structures, and distantly related sequences still fold into similar structures (62). This relationship was first identified by Chothia and Lesk (1986) (62) and later quantified by Sander and Schneider (1991)(270). Thanks to the exponential growth of the Protein Data Bank (PDB), Rost (1999) (266) could derive a precise limit for this rule (Fig. 8.1). As long as the length of two sequences and the percentage of identical residues fall in the region marked as "safe" the two sequences are practically guaranteed to adopt a similar structure (Fig. 8.1).

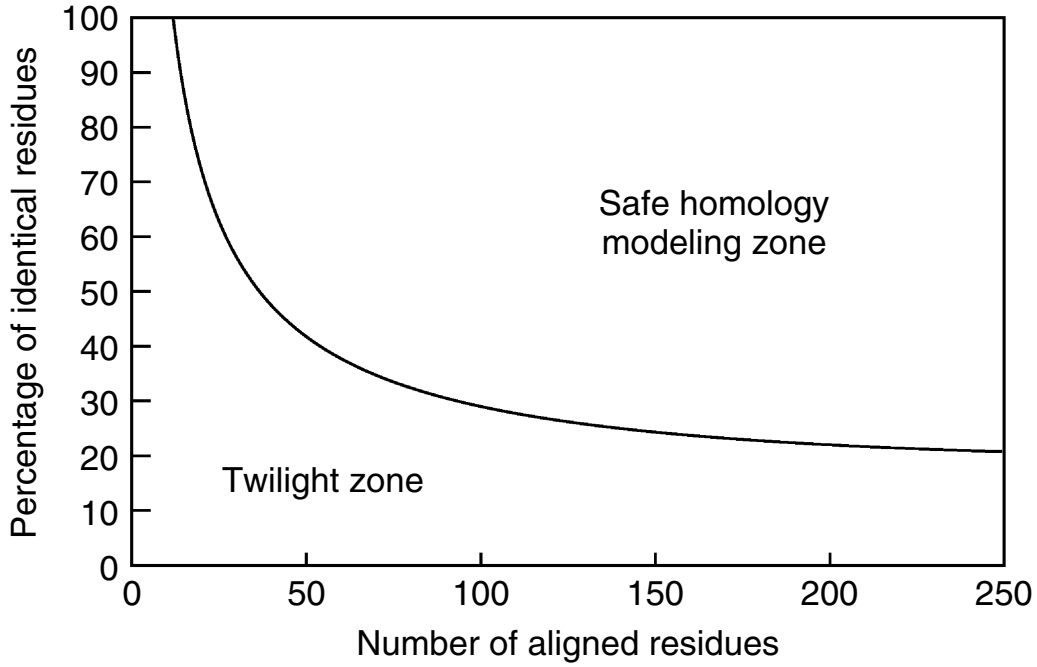


Figure 8.1: The two zones of sequence alignments. - Two sequences are practically guaranteed to fold into the same structure if their length and percentage sequence identity fall into the region marked as "safe" (266)

Homology modelling is usually described as a multi-step process, that can be summarized in seven steps:

1. Template recognition and initial alignment
2. Alignment correction
3. Backbone generation
4. Loop modeling

5. Side-chain modeling
6. Model optimization
7. Model validation

8.2.1 Step 1: Template recognition and initial alignment

In the safe homology modeling zone (Fig. 8.1), the percentage identity between the sequence of interest and a possible template is high enough to be detected with simple sequence alignment programs such as BLAST (4) or FASTA (236). To identify these hits, the program compares the query sequence to all the sequences of known structures in the PDB using mainly two matrices:

- A residue exchange matrix. It is 20x20 matrix where any couple of the 20 amino acids have a score of likelihood to be aligned. Along the diagonal there are the conserved residues with the highest score. Exchanges between residue types with similar physicochemical properties (for example F→Y) get a better score than exchanges between residue types that widely differ in their properties.
- An alignment matrix. The axes of this matrix correspond to the two sequences to align, and the matrix elements are simply the values from the residue exchange matrix for a given pair of residues. During the alignment process, one tries to find the best path through this matrix, starting from a point near the top left, and going down to the bottom right. To make sure that no residue is used twice, one must always take at least one step to the right and one step down.

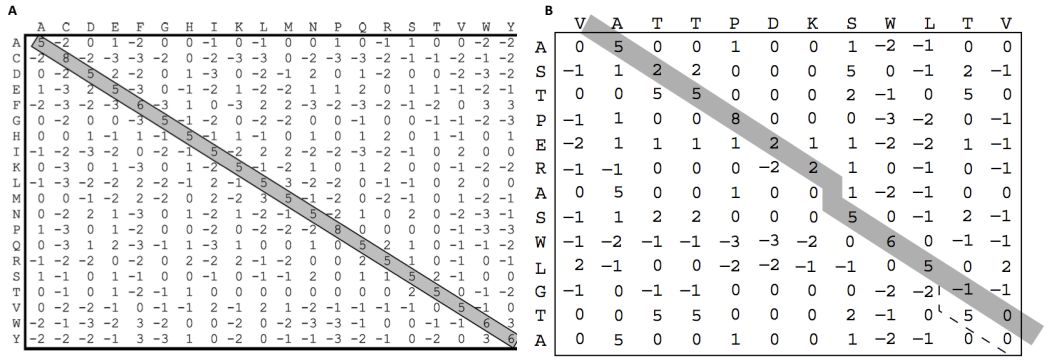


Figure 8.2: Matrix - A)residue exchange or scoring matrix used by alignment algorithms. This matrix is symmetric since the score for aligning residues A and B is normally the same as for B and A. B) alignment matrix for two sample sequences, using the scores from A. The optimum path corresponding is shown in grey.

Four general types of scoring have been applied to alignments:

Identity: considers only identical residues

Genetic Code: considers the number of base changes in DNA or RNA to inter convert the codons for the amino acids

Chemical Similarity: considers the physico-chemical properties (e.g., polarity, size, charge) with greater weight given to alignment of similar properties

Observed Substitutions: considers substitution frequencies observed in alignments of sequences (i.e. frequency with which a given amino acid is observed to be replaced by other amino acids among proteins for which the sequences can be aligned.)

In this thesis we use the BLOSUM matrix based on Observed Substitutions.

BLOSUM Matrices The substitution matrices derived by Dayhoff and co-workers (76) were based on substitution frequencies from global alignments of very similar sequences (The mutation probability matrix that they derived gives the probability of one amino acid mutating to a second amino acid within a particular evolutionary time). Henikoff and Henikoff (117) extended this approach by developing substitution matrices using local multiple alignments of more distantly related sequences. A database was assembled that contained multiple alignments (without gaps) of short regions of related sequences. These sequences were clustered into groups (blocks) based on their similarity at some threshold value of percentage identity. Blocks substitution matrices (BLOSUM) were derived based on substitution frequencies for all pairs of amino acids within a group. The different BLOSUM matrices were obtained by varying the threshold. For example, a BLOSUM80 matrix is derived using a threshold of 80% identity.

8.2.1.1 Definition of sequence identity and sequence similarity from (266)

Pairwise **sequence identity** was defined by the percentage of residues identical between two aligned sequences (e.g. aspartic matching aspartic counts 1: D - D = 1; aspartic on glutamic was a non-match: D - E=0). Pairwise **sequence similarity** was defined by the percentage of residues similar between two sequences (e.g. D - D $\leq$ 1; and aspartic on glutamic was now considered a match: D-E $>$ 0). Similarity scores depend on:

The percentage sequence identity between template and target. If it is greater than 90%, the accuracy of the model can be compared to crystallographically determined structures, except for a few individual side chains(62). From 50% to 90% identity, the root main square deviation error in the modeled coordinates can be as large as 1.5 $\AA$, with considerably larger local errors. If the sequence identity drops to 25%, the alignment turns out to be the main bottleneck for homology modeling, often leading to very large errors.

BLAST and FASTA are successful when the query is highly similar to structures in the database. In contrast, templates that are close to the possible homology modelling threshold are harder to find or may even remain undetected. The multiple sequence methods for fold identification (5; 144) are especially useful for finding significant structural relationships when the sequence identify between the target and the template drops below 25%. This class of methods appears to be one of the most sensitive fully-automated approaches for detecting remote sequence-structure relationships.

8.2.2 Step 2: Alignment correction

Once identified one or more possible modelling templates using the initial screening described above, more sophisticated methods are needed to arrive at a better alignment. Many programs are available to align a number of related sequences, for example, CLUSTALW(300), which take care of also of amino

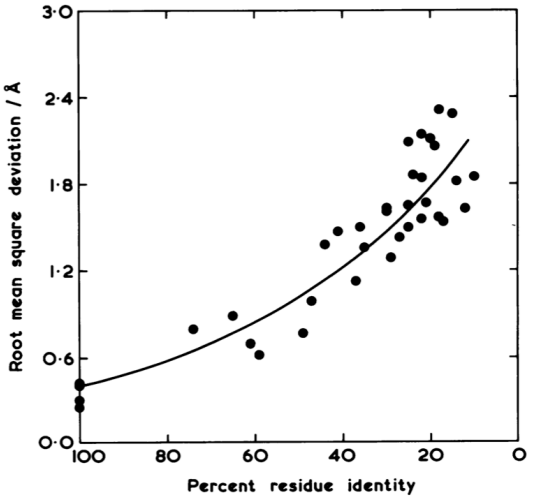


Figure 8.3: SI and RMSD - The relation of residue identity and the RMSD deviation of the backbone atoms of the common cores of 32 pairs of homologous proteins (62). $RMSD = \sqrt{\frac{\sum_{i=1}^{N} d_i^2}{N}}$ where N=total number of atoms and d_i=distance between the coordinates of an atom i at t=0 and tn when the structures are superimposed. Picture taken from Chotia and Lesk, EMBO J 1986 (62).

acids chemical features. For example, if at a certain position only exchanges between hydrophobic residues are observed, it is highly likely that this residue is buried. Considering this knowledge during the alignment, it is possible to use the multiple sequence alignment to derive position-specific scoring matrices, which are also called profiles(300). In recent years, new programs such as MUSCLE and T-Coffee have been developed that use these profiles to generate and refine the multiple sequence alignments(88; 221).

8.2.3 Step 3: Backbone generation

Once an initial target-template alignment has been built, a variety of methods can be used to construct a 3D model for the target protein. We can identify three classes (201):

1. Modeling by rigid-body assembly. This method assembles a model from a small number of rigid bodies obtained from aligned protein structures. The approach is based on the natural dissection of the protein folds into conserved core regions, variable loops, and side chains.
2. Modeling by segment matching. It relies on the approximate positions of conserved atoms in the templates. Comparative models can be constructed by using a subset of atomic positions from template structures as guiding positions, and by identifying and assembling short, all-atom segments that fit these guiding positions.
3. Modeling by satisfaction of spatial restraints. It uses either distance geometry or optimization techniques to satisfy spatial restraints obtained from the alignment.

I will treat only the latter one, since it is the one used in this thesis.

8.2.3.1 Modeling by satisfaction of spatial restraints

The methods in this class generate many constraints or restraints on the structure of the target sequence, using its alignment to related protein structures as a guide. The restraints are generally obtained by assuming that the corresponding distances and angles between aligned residues in the template and the target structures are similar. These homology-derived restraints are usually supplemented by stereochemical restraints on bond lengths, bond angles, dihedral angles, and nonbonded atom-atom contacts obtained from a molecular mechanics force field. The model is then derived by minimizing the violations of all the restraints. This can be achieved either by distance geometry or real-space optimization. A real-space optimization method, such as that implemented in the computer program MODELLER(269), starts by building the model using the distance and dihedral angle restraints on the target sequence derived from its alignment with template 3D structures. Then, the spatial restraints and an empirical force field terms, which enforce proper stereochemistry, are combined into an objective function. Finally, the model is generated by optimizing the objective function in Cartesian space. Modeling by satisfaction of spatial restraints can use many different types of information about the target sequence. Therefore this is a very powerful tool for proteins structure predictions. One of the strengths of modeling by satisfaction of spatial restraints is that constraints or restraints derived from a number of different sources can easily be added to the homology-derived restraints. For example, restraints might be obtained from NMR experiments, cross-linking experiments, fluorescence spectroscopy, image reconstruction in electron microscopy, site-directed mutagenesis, etc. In this way, a comparative model, especially in the difficult cases, could be improved by making it consistent with available experimental data.

8.2.4 Step 4: Loop modeling

Any insertion or deletion in the alignment implies a structural change of the backbone, and can thus not be modelled in the previous step. Since these changes usually take place outside regular secondary structure elements, their prediction is referred to as loop modelling.

There are two major approaches to the problem:

1. *Knowledge based*: which search the PDB for known loops with high sequence similarity to the target and endpoints that match the anchor residues between which the loop has to be inserted (210). All major molecular modeling programs and servers support this approach including Modeller, the one used here.
2. *Energy based*: which sample random loop conformations whith an energy function used to judge the quality of a loop (329).

8.2.5 Side-chain modeling

The most successful approaches to side-chain prediction are knowledge based. They use libraries of common side- chain rotamers extracted from high-resolution X-ray structures (60; 87; 192). An essential feature of these libraries is backbone dependence, hence they store the distribution of the side-chain dihedral angles (χ_1, χ_2 etc.) as a function of the backbone dihedrals ϕ and ψ. This not only increases the accuracy, but also decrease the search space: certain backbone conformations strongly favor certain rotamers (allowing, for example, a hydrogen bond between side chain and backbone). The prediction accuracy is usually quite high for residues in the hydrophobic core where more than 90% of all χ_1-angles

fall within ±20 degree from the experimental values, but much lower for residues on the surface where the percentage is often even below 50% (45). There are two reasons for this:

1. Experimental reasons: flexible side chains on the surface tend to adopt multiple conformations, which are additionally influenced by crystal contacts. So even experiment cannot provide one single correct answer.
2. Theoretical reasons: the energy functions used to score rotamers can easily handle the hydrophobic packing in the core (mainly Van der Waals interactions), but are not precise enough to get the complicated electrostatic interactions on the surface, including hydrogen bonds with water molecules and associated entropic effects.

Nevertheless, the surface residues are among the most important ones to get right; they mediate all the interactions, and applications such as drug design or protein docking thus critically depend on them (315).

8.2.6 Step 6: Model optimization

Once all these steps are completed, one obtains the initial homology model, which hopefully looks broadly similar to the target structure. The minor details, however, such as the precise backbone conformation, hydrogen-bonding networks or certain side-chain rotamers, are often wrong. Predictors try to bridge the gap between model and target using various optimization and refinement techniques, such as molecular dynamics and Monte Carlo simulations(315). However, for a given model, there are unfortunately many more paths leading away from the target than towards it, and combined with the limited accuracy of empirical force fields, this makes it very easy to reduce the model accuracy during the refinement. Consequently, the best optimization is often no optimization(315). Indeed at every minimization step, a few big errors (like bumps, i.e., too short atomic distances) are removed while many small errors are introduced(164; 165). When the big errors are gone, the small ones start accumulating and the model moves away from the target (164; 165)(Fig. 8.4).

Moreover, several protein structures published in the Protein Data Bank (PDB) were discovered to be erroneous (54). Even when the structure determination process is correct, the determined structure may adopt a non-physiological fold, for example, due to non-physiological constraints imposed by the crystal in the case of X-ray crystallography(147). Such errors increase in computationally derived structures, which are built by extrapolating from a homologous protein (95) and many alternative conformations might be generated during the simulation (147). Programs that try to numerically assess the correctness of a given structural model for a protein are called Model Quality Assessment Programs (MQAPs). The need for such programs is widely recognized by the structural biology community, as evidenced by the inclusion of a category for assessing MQAP performance in the biennial Critical Assessment of Techniques for Protein Structure Prediction (CASP) experiment (67).

Four are the approaches that address this problem and that perform well at the 2008 CASP (68; 147):

- YASARA (165), which runs molecular dynamics simulations of models in explicit solvent, using a new partly knowledge-based all atom force field derived from Amber(50), whose parameters have been optimized to minimize the damage done to protein crystal structures.
- The LEE-SERVER, which makes extensive use of conformational space annealing to create alignments, to help Modeller(269) build physically realistic models while satisfying input restraints from templates and CHARMM(42) stereochemistry, and to remodel the side-chains.

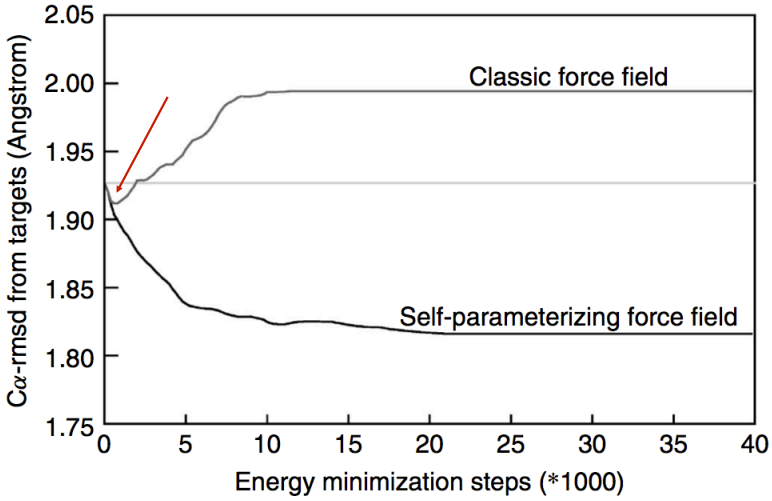


Figure 8.4: Model optimization - The average rmsd between models and targets during an extensive energy minimization of 14 homology models with two different force fields (AMBER(50) and YAMBER(164; 165)). Both force fields improve the models during the first 500 energy minimization steps but then the small errors sum up in the classic force field and guide the minimization in the wrong direction, away from the target while the self-parameterizing force field goes in the right direction. This figure was taken from Chapter 25 of Homology Modeling, by Elmar Krieger, Sander B. Nabuurs, and Gert Vriend (166)

- ROSETTA (74), whose high resolution refinement protocol combines a physically realistic all atom force field with Monte Carlo minimization to allow the large conformational space to be sampled quickly.

- UNDERTAKER (11), which creates a pool of candidate models from various templates and then optimizes them with an adaptive genetic algorithm, using a primarily empirical cost function that does not include bond angle, bond length, or other physics-like terms.

The single-structure MQAPs that performed best in CASP8 was LEE-server(147) and it is the ones used in this thesis.

8.2.7 Step 7: Model validation

The number of errors (for a given method) mainly depends on two values:

1. The percentage sequence identity between template and target. See Previous Sections.

2. The number of errors in the template. Protein structures were error free until the landmark article on Procheck by the Thornton group (177). This article can be seen as the beginning of the realization that crystallographers and NMR spectroscopists actually use experimental techniques to determine their coordinates. Structure validation became a common household technique for most scientists however a mayor bottlenecks remain: the detection of an error does not implicitly mean that the error can be removed.

8.2.8 Last Step: Iteration

If the model is not good enough, (part of) the modelling process has to be repeated. For instance, wrong side-chain conformations can be improved by iterating the process from step 5 onwards. Sometimes, this iteration step means that one has to start the modelling process all over again using another template or alignment. Alternatively, one can start several modelling processes using different templates. The resulting models can be combined in the end to produce a hybrid model that consists of the strongest points of each separate model.

8.3 From Microscopic to Macroscopic: Simulations as a bridge between theory and experiments

8.3.1 Introduction

An experiment is usually made on a macroscopic sample that contains an extremely large number of atoms or molecules sampling an enormous number of conformations. On the other hand computer simulations generate information at the microscopic level, including atomic positions and velocities. The conversion of this microscopic information to macroscopic observables such as pressure, energy, heat capacities, etc., requires statistical mechanics.

8.3.1.1 Statistical Mechanics

Statistical mechanics provides the mathematical expressions that relate macroscopic properties to the distribution and motion of the atoms in the (bio)molecular system.

Thermodynamic state and microscopic state. The *thermodynamic state* of a system is defined by a set of parameters (state variables) that are physical observables, like for example, the temperature, T, the pressure, P, and the number of particles, N.

The *microscopic state* of a system is defined by all the N atoms' positions,

$$(q_1,q_2,...,q_N) \equiv q^N$$

and momenta

$$(p_1,p_2,...,p_N) \equiv p^N$$

, i.e., points in the multidimensional phase space called phase space (Γ). A single point in phase space, denoted by G, describes the state of the system.

Thus, the thermodynamic state is interpreted in terms of all the accessible microscopic states. These are assumed to have all the same probability.

An *ensemble* is a *collection* of points in phase space satisfying the conditions of a particular thermodynamic state. Or we can also say that it is a collection of all possible systems, which have different microscopic states but have an identical macroscopic or thermodynamic state.

Common techniques to sample the configurations assumed by a system at equilibrium are Molecular Dynamics (MD), Monte Carlo (MC) sampling and Stochastic/Brownian dynamics. We focus here on MD because this is the approach used in this thesis. A MD simulation generates a sequence of points in phase space as a function of time; these points belong to the same ensemble, and they correspond to the different conformations of the system and their respective momenta. Thermodynamic quantities are determined in computer simulations as *ensemble averages* over a large number of microscopic configurations assumed by the system under study. This is allowed by the ergodic hypothesis assumption, described in the next session.

8.3.1.2 The Ergodic hypothesis

In statistical mechanics, averages corresponding to thermodynamic observables are defined in terms of ensemble averages. Thermodynamic observables can be modeled by considering at once a collection of identical systems. Each system represents one of all the accessible microstates. The macrostate is

allowed to evolve in time in MD simulation. Its behaviour can be characterized by a time dependent distribution function, (qN(t), pN(t)) for the microstates. The instantaneous average value of the observable, O, over the phase space is interpreted as:

$$\frac{\int_{\Gamma} O(q^N(t)p^N(t))\rho(q^N(t)p^N(t))dq^N dp^N}{\int_{\Gamma} \rho(q^N(t)p^N(t))dq^N dp^N} \tag{8.1}$$

If we assume equal probability for all microstates, then the distribution of points in phase space is frozen into one single shape, i.e., the distribution function is time invariant, and the condition:

$$\frac{\rho(q^N(t)p^N(t))dq^N dp^N}{dt} = 0 \tag{8.2}$$

describes the thermodynamic equilibrium. The so-called ensemble average is defined as:

$$\langle O \rangle_{ensemble} = \frac{\int_{\Gamma} O(q^N(t)p^N(t))\rho(q^N(t)p^N(t))dq^N dp^N}{\int_{\Gamma} \rho(q^N(t)p^N(t))dq^N dp^N} \tag{8.3}$$

In a molecular dynamics simulation, the points in the ensemble are calculated sequentially in time, thus, assuming that the equations of motion of the system are solved, each observable can be empirically associated with a function (O), of the instantaneous microstate, $(q^N(t), p^N(t))$, of the system. Thus any microscopic observable is assumed to be a time averaged value:

$$\langle O \rangle_{ensemble} = \lim_{\tau \to \infty} \int_{\tau_0}^{\tau_0 + t} O(q^N(t)p^N(t))dt \tag{8.4}$$

The **ergodic hypothesis** states that in thermodynamic equilibrium, the **time average** and the **ensemble average** are equal.

That is, if one allows the system to evolve indefinitely in time, the system will pass through all possible microstates, and the experimental measurement will coincide with the calculated time and *ensemble averages*.

Such a procedure is well founded only for the so-called *ergodic systems*, which are assumed to fully sample the accessible phase space (hyper)volume during the "observation" (i.e. "simulation") time. It is generally assumed that complex systems, such as the majority of those biologically relevant, are ergodic. Although this is a plausible assumption, it should be pointed out that the ergodic hypothesis is not always true for biological systems (for some example see e.g. Ref. (96)).

8.3.1.3 Trajectory Accuracy: Shadow Orbits and the Liapunov instability

A further drawback affecting (in principle) all kinds of MD simulations is the so-called Liapunov instability (190):

Two trajectories differing initially by an infinitesimal amount will diverge exponentially in time.

For chaotic systems, like almost all those simulated by MD, the trajectory is extremely sensitive to initial conditions. Any error in the integration of the equation of motion, no matter how small, will always cause the simulated (numerical) trajectory to diverge exponentially from the "true" trajectory starting from the same initial conditions. Thus, any imperfect integrator (and all are imperfect) introduces errors that guarantee the trajectory diverges from the true trajectory. How can we know that we are generating the correct results in MD? In some cases, there is actually a "shadow orbit" that closely follows the integrated trajectory. Or from other point of view, the true trajectory, to which

the numerical one overlaps for a certain period of time, is called "shadow orbit" [1]. The shadow orbit is an exact trajectory for the system, but it starts from a slightly displaced initial point. At the time the Liapunov instability raises up, the numerical trajectory will get far from that specific shadow orbit, but there always will be another one of these to which it is superimposed.

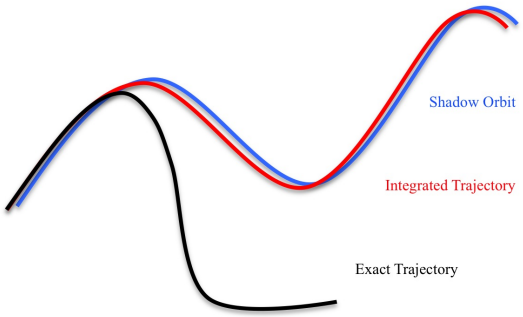


Figure 8.5: Shadow Orbit - Schematic representation of a Shadow Orbit

8.3.1.4 How Long? How Large?

Molecular dynamics evolves a finite-sized molecular configuration forward in time, in a step-by-step fashion. There are limits on the typical time scales and length scales that can be investigated and the consequences must be considered in analyzing the results.

Correlation Time and Correlation Length. Simulation runs are typically short corresponding to few nanoseconds of real time, and in special cases extending to the microsecond regime. This means that we need to test whether or not a simulation has reached equilibrium before we can trust the averages calculated in it. Moreover, there is a clear need to subject the simulation averages to a statistical analysis, to make a realistic estimate of the errors. How long should we run? This depends on the system and the physical properties of interest. If we have an observable a, the time (t) correlation function is $\langle a(t_0)a(t_0+t)\rangle$, where t_0 is a constant that fix the initial time. Assuming that the system is in equilibrium, this function is independent of the choice of time origin and may be written $\langle a(0)a(t)\rangle$, fixing $t_0 = 0$. From 8.4 we can easily define a correlation time

$$\tau_a = \int_0^\tau dt \, \langle a(0)a(t)\rangle \, / \, \langle a^2 \rangle \tag{8.5}$$

for which the measure of a(0) and a(t) became uncorrelated. At the same way, we can define a spatial correlation function $\langle a(0)a(t)\rangle$ relating values computed at different points r apart. Spatial isotropy allows us to write this as a function of the distance between the points, r, rather than the

[1] These are known to exist for hyperbolic systems. Can sometimes be shown to exist, for long times, for more general systems [e.g., see Quinlan and Tremaine, Mon. Not. R. Astron. Soc. 259, 5050 (1992)].

vector r: notably this symmetry is broken in a liquid crystal. Spatial homogeneity, which applies to simple liquids (but not to solids or liquid crystals) allows us to omit any reference to an absolute origin of coordinates. This function decays from a short-range nonzero value to zero over a characteristic distance ξa, the correlation length. It is almost essential for simulation box sizes to be large compared with ξa, and for simulation run lengths τ to be large compared with τa, for all properties of interest a. Only when these two conditions are respected it is guaranteed that reliably-sampled statistical properties are obtained.

Accessible time and length scale. An important issue of simulation studies is the accessible time and length scale coverable by microscopic simulations. Figure 8.6 shows a schematic representation for different types of simulations in a length-time-diagram. It is clear that the more detailed a simulation technique operates, the smaller is the accessibility of long times and large length scales. Therefore quantum simulations, where fast motions of electrons are taken into account, are located in the lower left corner of the diagram and typical length and time scales are of order of and ps. Classical molecular dynamics approximates electronic distributions in a rather coarse-grained fashion by putting either fixed partial charges on interaction sites or by adding an approximate model for polarization effects. In both cases, the time scale of the system is not dominated by the motion of electrons, but by the time of intermolecular collision events, rotational motions or intramolecular vibrations, which are orders of magnitude slower than those of electron motions. Consequently, the time step of integration is larger and trajectory lengths are of order ns and accessible lengths of order 10–100 A. If tracer particles in a solvent medium are considered, Brownian dynamics can be applied, where the effect of the solvent is hidden in average quantities. Since collision times between tracer particles is very long, larger time steps may be applied. Furthermore, since the solvent is not simulated explicitly, the length scales may be increased considerably. Finally, if one is interested not in a microscopic picture of the simulated system but in macroscopic quantities, the concepts of hydrodynamics may be applied, where the system properties are hidden in effective numbers, e.g. density, viscosity, sound velocity.

8.3.1.5 Design a Molecular Dynamic Simulation in biomolecular field

A key decision in beginning a simulation of a biomolecular system is the choice of an appropriate method for that particular system and for the questions of interest. A simulation method should be capable of delivering a reliable result in a reasonable time (309).

Studies involving multi-nanosecond dynamics simulations are now common. However, expert knowledge is still required, and care needs to be taken to ensure that the application of a biomolecular simulation method to a particular problem is meaningful and useful.

Concerning MD simulations the "ingredients" are basically three:

1. A model for the interaction between system constituents (atoms, molecules, surfaces etc.) is needed.

2. An integrator is needed, which propagates particle positions and velocities from time t to t + δt. It is a finite difference scheme that moves trajectories discretely in time. The time step δt has properly to be chosen to guarantee stability of the integrator, i.e. there should be no drift in the systems energy.

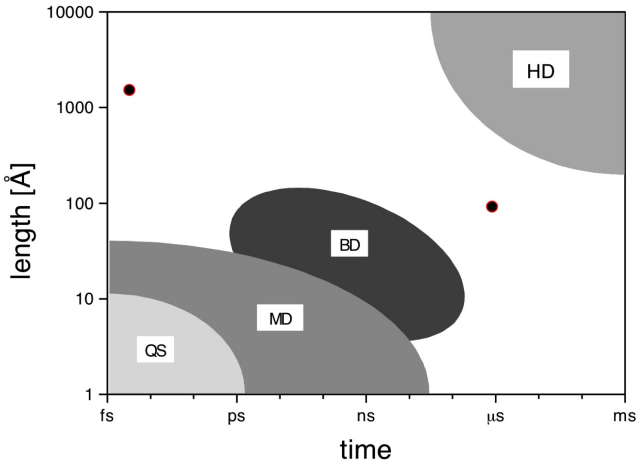


Figure 8.6: Time and Length scales - Schematic comparison of time and length scales, accessible to different types of simulation techniques (quantum simulations (QM), molecular dynamics (MD), Brownian dynamics (BD) and hydrodynamics/fluid dynamics (HD)).

3. A statistical ensemble has to be chosen, where thermodynamic quantities like pressure, temperature or the number of particles are controlled.

These choices essentially define an MD simulation.

8.4 Molecular Dynamics Simulations

Introduction

Molecular systems and motion of their constituents, both nuclei and electrons, are known to be accurately described only by laws of quantum mechanics. Implementation of classical laws of mechanics in this scheme involves a series of approximations for the quantum description:

1. The molecular wavefunction (solution) of the molecular Schrödinger equation is separated into nuclear and electronic parts, so that motion of nuclei is decoupled from electronic motion due to the fact that nuclei are much heavier. Such decoupling allows their equations of motion to be separated and solved.
2. Nuclei are approximated as classical particles.
3. Third, electronic variables could be either integrated out beforehand and an approximate single potential energy surface (usually representing the electronic ground state) is constructed; or they could be treated within a suitable approximation as active degrees of freedom via the electronic Schrödinger equation, and forces on nuclei are computed by electronic structure calculations that are performed for each generated trajectory.

Accordingly to how the electronic part is treated, MD simulation is branched out into two methodologies:

- Classical Molecular Dynamics, where forces are derived from predefined potential models by analytical gradient applications.
- Ab initio Molecular Dynamics, where forces on nuclei are obtained from the electronic structure calculations.

I will describe both.

8.4.1 The semiclassical approximation

A molecular system with the positions of N nuclei, $R = \{R_1, R_2..., R_N\}$, and the n electrons located at $r = \{r_1, r_2..., r_n\}$, is completely described non relativistically by the molecular Schrödinger equation:

$$i\hbar\frac{\partial}{\partial t}\Phi(r, R; t) = H\Phi(r, R; t) \tag{8.6}$$

Where H is the non-relativistic molecular Hamiltonian, in SI units, of the N nuclei and n electrons. In absence of external fields, it has the form:

$$\begin{aligned} H(R, r) = &- \sum_{i=1}^{N} \frac{\hbar}{2M_I}\nabla^2_{R_I} - \sum_{i=1}^{n} \frac{\hbar}{2M_e}\nabla^2_{r_i} + \sum_{I=1,J>1}^{N,N-1} \frac{Z_i Z_j e^2}{4\pi\varepsilon_0|R_I - R_J|} + \\ &\sum_{i=1,j>1}^{n,n-1} \frac{e^2}{4\pi\varepsilon_0|r_i - r_j|} + \sum_{I=1,i=1}^{N,n} \frac{Z_I e^2}{4\pi\varepsilon_0|R_I - r_i|} \end{aligned} \tag{8.7}$$

Where M_I is the mass of nucleus I whose atomic number is Z_I, m_e is the mass of the electron, and e is its charge. The operator ΔR_I and Δr_i act on the coordinates nucleus I and electron i respectively. Thus we can write as:

$$H(R, r) = T_N(R) + T_e(r) + V(r, R) = T_R(R) + H_e(r, R)$$

The electron-nucleus interactions bind electrons to nuclei, and leads to a mathematical inseparable Hamiltonian.

At room temperature the thermal wavelength λ is about 0.1 A, while typical interatomic distances, in liquids and solids, are of the order of 1 A. Thus, a good approximation is to neglect quantum correlations between wave functions of different nuclei, i.e to consider the nuclear wavefunction as an incoherent superimposition of individual nuclear wave packets. In addition, nuclear masses are large enough that such individual wave packets are usually well localized. Formally, we can separate electronic from nuclear degrees of freedom by writing the wavefunction as product of terms depending only on electronic or nuclear positions ("one-determinant" Ansatz):

$$\Phi(r, R; T) \approx \Psi(r; t)\chi(R; t) \exp\left[\frac{1}{\hbar}\int_{t_0}^{t} dt' E_e(t')\right] \tag{8.8}$$

where the electronic and nuclear wavefunctions are separately normalized:

$$\langle\Psi(r; t) \mid \Psi(r; t)\rangle = 1, \langle\chi(R; t) \mid \chi(R; t)\rangle = 1$$

and the phase factor has the form:

$$\int dr dR \Psi^*(r; t)\chi^*(R; t) H_e \Psi(r; t)\chi(R; t)$$

Inserting this last expression in Schrödinger equation, multiply from left by $\langle\Psi(r;t)|$ and $\langle\chi(R;t)|$, integrate over R and r, and apply the energy conservation:

$$\frac{d}{dt}\int \Phi^*(r,R;t)H\Phi(r,R;t) = 0$$

the following system of coupled equations is obtained:

$$i\hbar\frac{\partial\Psi}{\partial t} = -\sum\frac{\hbar}{2m_e}\nabla^2_{r_i}\Psi + \left\{\int drdR\chi^*(R;t)V(r,R)\chi(R;t)\right\}\Psi \tag{8.9}$$

$$i\hbar\frac{\partial\chi}{\partial t} = -\sum\frac{\hbar}{2M_I}\nabla^2_{R_I}\chi + \left\{\int drdR\Psi^*(r;t)H_e\Psi(r;t)\right\}\chi \tag{8.10}$$

which defines the basis of the TDSCF, method introduced as early as 1930 by Dirac.

Each wavefunction above obey Schrödinger equation but with time dependent effective potential obtained by appropriate averages over the other degrees of freedom; both electrons and nuclei move quantum-mechanically in time-dependent effective potentials (or self-consistently obtained average fields) obtained from appropriate averages (quantum mechanical expectation values $\langle ...\rangle$) over the other class of degrees of freedom. The next step in the derivation of classical molecular dynamics is the task to **approximate the nuclei as classical point particles**. A classical description of nuclei dynamics is achieved by expressing χ in terms of an amplitude factor A and a phase S which are both considered to be real and A > 0 in this polar representation:

$$\chi(R;t) = A(R;t)\exp\left[\frac{iS(R;t)}{\hbar}\right] \tag{8.11}$$

Using the polar representation, in the classical limit $\hbar \to 0$, on the 8.9 and 8.10, the following set of equations is obtained:

$$\frac{\partial S}{\partial t} + \sum_{I=1}^{N}\frac{(\nabla_I S)^2}{2M_I} + \int dr\Psi^* H_e\Psi = 0 \tag{8.12}$$

$$\frac{\partial A^2}{\partial t} + \sum_{I=1}^{N}\nabla\cdot\left(A^2\frac{\nabla_I S}{M_I}\right) = 0 \tag{8.13}$$

The equation for A is a continuity equation for the density probability $A^2 = |\chi|^2$ of nuclei, which move with classical velocities $\nabla_I S/M_I = p_I/M_I$.

More important for our purpose is the equation 8.12 . Indeed, if we use the connection $P_I \equiv \nabla_I S$, it becomes isomorphic to the Hamilton-Jacobi equation of classical motion for action S and Hamiltonian *H(R,P)=T(P)+V(R)* defined in terms of (generalized) coordinates $\{R_I\}$ and their conjugate momenta $\{P_I\}$. The Newtonian equation of motion $\dot{P}_I = -\nabla_I V(\{R_I\})$ corresponds to 8.12:

$$\begin{aligned} &\tfrac{dP_I}{dt} = -\nabla_I\textstyle\int dr\Psi^* H_e\Psi \\ &or \\ &M_I\ddot{R}_I(t) = -\nabla_I\textstyle\int dr\Psi^* H_e\Psi \end{aligned} \tag{8.14}$$

Thus, the nuclei move according to classical mechanics in an effective potential V_e^E due to the electrons. This potential is a function of only the nuclear positions at time t as a result of averaging H_e over the electronic degrees of freedom, i.e. computing its quantum expectation value $\langle\Psi|\,H_e\,|\Psi\rangle$, while keeping the nuclear positions fixed at their instantaneous values $\{R_I\}$. In other words, nuclei are driven by a mean-field potential due to electrons and containing also a contribution from their

kinetic energy. Finally, to get off the nuclear wavefunction also from 8.9 one replaces the nuclear density $|\chi(\{R_I(t)\};t)|^2$ by a product of delta functions $\prod_I \delta(R_I - R_I(t))$, i.e. incoherent wave packets extremely localized. At the classical limit, the electronic wave equation is:

$$i\hbar\frac{\partial\Psi}{\partial t} = -\sum_i \frac{\hbar^2}{2m_e}\nabla^2_{r_i}\Psi + V(r, R(t))\Psi \tag{8.15}$$

which evolves selfconsistently as the classical nuclei are propagated via 8.14. Note that now He and thus Ψ depend parametrically on the classical nuclear positions $\{R_I(t)\}$ at time t through $V_e^E(\{R_I(t)\})$. These equations (8.14 and 8.15) represent the so called "Ehrenfest molecular dynamics" scheme. It is clear now that the motion of the nuclei is dictated by the Hamiltonian *He*, which basically contains the quantistic information on the electronic system.

Thus, the major task of quantum mechanics concerns the solution of the Schrödinger equation for the electrons, whose solution allow to know the dynamical behaviour of the system.

8.4.2 Derivation of classical molecular dynamics equations

Although the TDSCF approach underlying Ehrenfest molecular dynamics clearly is a *meanfield theory*, transitions between electronic states are included in this scheme. Thus, at this stage a further simplification can be invoked by restricting the total electronic wave function Ψ to be the ground state wave function Ψ_0 of He at each instant of time. This should be a good approximation if the energy difference between Ψ_0 and the first excited state Ψ_1 is everywhere large compared to the thermal energy k_{BT}, roughly speaking.

In this limit the nuclei move on a single potential energy surface:

$$V_e^{Ehr} = \int dr \Psi_0^* H_e \Psi_0 \equiv E_0(\{R_I(t)\}) \tag{8.16}$$

which is computed by solving the electronic time-independent Schrödinger equation only for the ground state:

$$H_e\Psi_0 = E_0\Psi_0 \tag{8.17}$$

Now, E_0 is a function of nuclear positions R, and both Ehrenfest and the ground state Born-Oppenheimer potentials are identical.

As a consequence of this observation, it is conceivable to decouple the task of generating the nuclear dynamics from the task of computing the potential energy surface. Assuming the possibility to solve the stationary Schrödinger equation for as many nuclear configurations as possible, the classical molecular dynamics approach is derived by the following three steps scheme:

1. Solving 8.17 for many representative nuclear configurations to compute the ground state energy E_0.
2. The generated data points R, $V^{Ehr}(R)$ or some equivalent experimental data points are fitted to a suitable analytical functional form to construct a global potential energy surface.
3. The following Newtonian equation of motion:

$$M_I\ddot{R}_I(t) = -\nabla_I V_e^{Ehr}(\{R_I(t)\}) \tag{8.18}$$

is solved by applying analytically the gradient for many different initial conditions to produce the nuclear classical trajectories on this global potential energy surface.

Furthermore the overall internal interaction potential is approximated to V^{appr} which is expanded to pair-wise, three-body, four-body and up to n-body contributions, these contributions are categorized as intermolecular long-range and intramolecular short-range interaction:

$$V^{Ehr} \approx V^{appr}(R) = \sum_{I<J}^{N} V_{IJ}(R_I R_J) + \sum_{I<J<K}^{N} V_{IJK}(R_I R_J R_K) + \sum_{I<J<K<L}^{N} V_{IJKL}(R_I R_J R_K R_L) + ... \tag{8.19}$$

Potential expansion is practically truncated at some term to reduce the dimensionality resulting from the increase of the number of active nuclear degrees of freedom. Within the same potential expansion, electronic degrees of freedom do no longer appear explicitly but are effectively included in a functional form of V^{appr} potential.

As a result of this derivation, the essential assumptions underlying classical molecular dynamics become transparent: the electrons follow adiabatically the classical nuclear motion and can be integrated out so that the nuclei evolve on a single BornOppenheimer potential energy surface (typically but not necessarily given by the electronic ground state), which is in general approximated in terms of fewbody interactions.

8.5 Empirical Force Fields

As relevant biological processes usually involve large systems (thousands of atoms or more), and occur in relatively long timescales (from nano to microseconds or more), it is necessary to develop effective parametrized potentials, which are faster to integrate, albeit less accurate, in order to study this kind of systems. The term force field indicates a functional form for this approximation, which relates the configuration of the system ($\{R_i\}, i = 1, ..., N$) to its internal energy U, along with the set of parameters used in that function. In this work, the AMBER(234) force field for description of macromolecules in solution has been used, while the GROMACS(28; 121) package has been used to integrate the equation of motion. The functional form can be written as:

$$\begin{aligned} U = &\sum_{bonds} K_r(r - r_{eq})^2 + \sum_{angles} K_\theta(\theta - \theta_{eq})^2 + \sum_{dihedrals} \tfrac{V_n}{2}[1 + \cos(n\varphi - \gamma)] + \\ &+ \sum_{i<j} \left[4\varepsilon_{ij} \left(\tfrac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\tfrac{\sigma_{ij}}{r_{ij}} \right)^{6} + \tfrac{q_i q_j}{\varepsilon r_{ij}} \right] \end{aligned} \tag{8.20}$$

Atom bond stretching and angle bending are represented as harmonic terms, while dihedrals or torsional are described by a sinusoidal term. Non-bonded interactions comprise two terms, the first is a Lennard-Jones 6-12 which describes atom-atom repulsion and dispersion interactions, the second is the Coulomb electrostatic term. In eq. 8.20, r and θ are respectively the bond length and valence angle; ϕ is the dihedral or torsion angle and r_{ij} is the distance between atoms i and j. Parameters include the bond force constant and equilibrium distance, K_r and r_{eq}, respectively; the valence angle force constant and equilibrium angle, K_θ, and θ_{eq}, respectively; the dihedral force constant, multiplicity and phase angle, V_n, n, and γ, respectively. The functional form used for out-of-plane distorsions (e.g. in planar groups) is different in different force fields. For instance, in the AMBER force field this term has the same form as that used for proper dihedrals, while in CHARMM an harmonic term is used. Collectively, these parameters represent the internal or intramolecular ones.

Non bonded parameters between atoms i and j include the partial atomic charges, qi, along with the LJ well-depth, ϵ_{ij}, and σ_{ij}, the (finite) distance at which the inter-particle potential is zero. These

terms are also referred to as the interaction or external parameters. Typically, ϵ_{ij}, and σ_{ij} are obtained for individual atom types and then combined to yield ϵ_{ij}, and σ_{ij} for the interacting atoms via combining rules. The dielectric constant ϵ is typically set to 1 (corresponding to the permittivity of vacuum) in calculations that incorporate explicit solvent representations.

Van der Waals and electrostatic interactions are calculated between atoms belonging to different molecules or for atoms in the same molecules separated by at least three bonds. For the van der Waals potential, this truncation introduce only a small error in the energy.

This is not the case for the electrostatic potential, because the Coulomb interaction between charges q_i and q_j decays slowly with distance. Hence it can not be truncated, but when periodic boundary conditions are used, it is computed with efficient schemes such as Particle Mesh Ewald in conjunction with periodic boundary conditions, which approximate the exact result to an acceptable error similar to the error in the van der Waals potential.

8.5.1 Long Range Interactions

In simulations of biological systems it is highly convenient to avoid the calculation of all non-bonded pair interactions, as the computational cost would be proportional to the square of the number of atoms. These interactions primarily dictates the dynamics of biomolecules, and cannot be merely truncated beyond a given cutoff when long-ranged. The difference between short and long interactions is the spatial extent of the potential. If the potential drops down to zero faster than r^d, where r is the separation between two particles and d the dimension of the problem, it is called **short ranged**, otherwise it is **long ranged**. This becomes clear by considering the integral:

$$I = \int \frac{dr^d}{r^n} = \{\infty : n \leq d; finite : n > d$$

i.e a particles potential energy gets contributions from all particles of the universe if n $\leq$ d, otherwise the interaction is bound to a certain region, which is often modeled by a spherical interaction range. Long range interactions essentially require to take all particle pairs into account for a proper treatment of interactions. Coulomb ($\sim \phi_{-1}$) and dipole-dipole ($\sim \phi_{-3}$) should be considered long-range when dealing with three-dimensional systems. This may become a problem, if periodic boundary conditions are imposed to the system, i.e. formally simulating an infinite number of particles (no explicit boundaries imply infinite extent of the system). Therefore one has to devise special techniques to treat this situation.

8.5.2 Ewald Summation Method

Considering the electrostatic energy of a system of particles in a cubic box and imposing periodic boundary conditions, leads to an equivalent problem. At position r_i of particle i, the electrostatic potential, ϕ, can be written down as a lattice sum:

$$\varphi(r_i) = \frac{1}{8\pi\varepsilon_0} \sum_{|n|=0}^{\infty} \sum_{j=1}^{N} \frac{q_j}{|r_{ij} + nL|}$$

thus, for N particles:

$$E = \frac{1}{8\pi\varepsilon_0} \sum_{|n|=0}^{\infty} \left[\sum_{j=1}^{N} \sum_{j=1}^{N} \frac{q_j q_i}{|r_{ij} + nL|} \right]$$

where L is the length of the periodic box , N is the total number of atoms, and n are the direct lattice vectors. i is not equal to j for $|n| = 0$. This equation is conditionally convergent, i.e. the result of the outcome depends on the order of summation. Moreover, the sum extends over infinite number of lattice vectors. Thus, the procedure has to be modified in order to get an absolute convergent sum and to get it fast converging. The original method of Ewald consisted in introducing a convergence factor e^{-ns}, which makes the sum absolute convergent; then transforming it into different fast converging terms and then putting s in the convergence factor to zero. The final result of the calculation can be easier understood from a physical picture. If every charge in the system is screened by a counter charge of opposite sign, which is smeared out, then the potential of this composite charge distribution becomes short ranged (it is similar in electrolytic solutions, where ionic charges are screened by counter charges - the result is an exponentially decaying function, the Debye potential). In order to compensate for the added charge distribution this has to be subtracted again. The far field of a localized charge distribution is, however, again a Coulomb potential. Therefore this term will be long ranged. The efficiency gain shows up, when the short range interactions are calculated as direct particle-particle contributions in real space, while the long range part of the smeared charge cloud are summed up in reciprocal Fourier space. Choosing as the smeared charge distribution a Gaussian charge cloud of half width $1/\alpha$ the corresponding expression for the energy becomes:

$$\varphi(r_i)\sum_{n}\sum_{j=1}^{N} q_j \frac{erfc(\alpha\,|r_{ij}+nL|}{|r_{ij}+nL|} + \frac{4\pi}{L^3}\sum_{k\neq 0}\sum_{j=1}^{N}\frac{q_j}{|k|^2}e^{-|k|^2/4\alpha}e^{ikr_{ij}} - q_i\frac{2\alpha}{\sqrt{\pi}}$$

k is the reciprocal vector in a cubic box. The parameter α tunes the relative weights of real and reciprocal sums, although the final result is independent of it. An optimal choice for it makes the Ewald sum converge as $N^{3/2}$, which can be further improved to NlnN with the use of Particle-Mesh methods (as the Particle-Mesh Ewald, PME or the Particle-Particle Particle-Mesh, PPPM)(71; 92), making advantage of the Fast Fourier Transform. The last term corresponds to a self-energy contribution which has to be subtracted, as it is considered in the Fourier part. The new equation is an exact equivalent of the first, with the difference that it is an absolute converging expression. Therefore nothing would be gained without further approximation. Since the complimentary error function can be approximated for large arguments by a Gaussian function and the k-space parts decreases like a Gaussian, both terms can be approximated by stopping the sums at a certain lattice vector n and a maximal *kvalue*, *kmax*.

8.5.3 Boundaries

Restriction on the size of a time step is not the only challenge in molecular dynamics methods. Another concerns the finite size effects of the simulated system as its number of particles is far fewer than that in any natural sample, and is most from thousands to maximum few millions. Enclosing the system with a rigid-walled container, most particles would be under the influence of its boundaries through collisions. If we ignore the boundaries most particles would lie at surface whose area tends to be minimized, distorting thus the shape of the system whenever it is a non-spherical. It is of no help to increase particles in a system as the more particles exist, the more particles are at the surface and more undesired effects are encountered. Those peculiarity due to the size limit and the improper treatment of boundaries, makes it unreliable to statistically extract macroscopic bulk properties since the later are calculated in the limit $N \to \infty$, where N is the number of particles. To go over both practical difficulties, periodic boundary conditions are imposed on the relatively small systems in such a way

that particles experience forces as if they reside in the bulk.

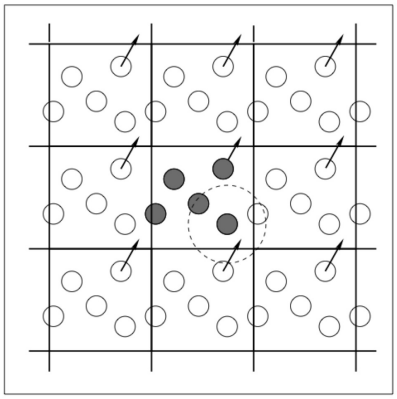


Figure 8.7: Periodic boundary conditions. - As a particle moves out of the simulation box, an image particle moves in to replace it. In calculating particle interactions within the cutoff range, both real and image neighbours are included.

8.5.3.1 Periodic boundary conditions (PBC)

When applying periodic boundaries, the fundamental (primitive) simulation cell is replicated infinitely and periodically in all directions. There is no restriction on the shape of the cell other than having the characteristic to completely fill all of space translationally with no overlaps nor voids. It is appropriate to choose a cell shape that reflects the underlying geometry of the system in question. When the interactions are of a short range each side of the replicated primitive cell must be of a length that is at least twice the radius of the spatial cutoff so to keep accuracy. Particles in this case are subjected to the condition such that when a particle leaves the primitive cell, its image from the cell on the opposite side reenters the cell with the same velocity. Herein, boundaries of the cell are no longer rigid but imaginary and their effects are completely absent. When subjecting the system to this condition, the system is not any more invariant (symmetric) under space rotation; henceforth the angular momentum is no longer conserved whereas the linear momentum and mechanical energy are still conserved.

8.5.3.2 Minimum image convention for short range interactions

For short ranged forces, PBC are used in conjunction with the minimum image convention. In this scheme each particle interacts at most with only one image of every other particle n the system. To exclude interactions of a particle with its own images (self-interaction), the assumed cubic simulation cell, as already mentioned, must have a side length of at least as twice as the radius of the cutoff. Interactions terms between pairs further away from each other than the cutoff radius are obviously zero.

8.5.4 Neighbors List

For short-range potentials, not all the n-permutations represent a set of interacting particles since particles at a larger separation than a spatial cutoff radius do not interact. Nevertheless computing the non-bonded contribution to the interatomic forces in an MD simulation involves, in principle, a large number of pairwise calculations: we consider each atom i and loop over all other atoms j to calculate the minimum image separations rij. Let us assume that the interaction potentials are of short range, $\nu(r_{ij}) = 0$ if $r_{ij} > r_{cut}$, the potential cutoff. In this case, the program skips the force calculation, avoiding expensive calculations, and considers the next candidate j. Nonetheless, the time to examine all pair separations is proportional to the number of distinct pairs, $1/2N(N-1)$ in an N-atom system, and for every pair one must compute at least r_{ij}^2; this still consumes a lot of time. Some economies result from the use of lists of nearby pairs of atoms. Verlet suggested such a technique for improving the speed of a program. The potential cutoff sphere, of radius r_{cut}, around a particular atom, is surrounded by a 'skin' of rlist radius. At the first step in a simulation, a list is constructed of all the neighbours of each atom, for which the pair separation is within rlist. Over the next few MD time steps, only pairs appearing in the list are checked in the force routine. Therefore, in a force routine, not all particles have to tested, whether they are in a range $r < r_{cut}$, but only those particle pairs, stored in the list. Since particles are moving during the simulation, it is necessary to update the list from time to time. List update must be at the correct frequency, a common update is between 10 to 20 time steps. To avoid double counting in the energy summation, only neighbors where (j > 1) are stored. In some cases of three/four body interactions, it is a must to only exclude equal indices, i.e., the list must contain all the pairs (j $\neq$ 1) for evaluation of three-body terms defined by the valence bond angle.

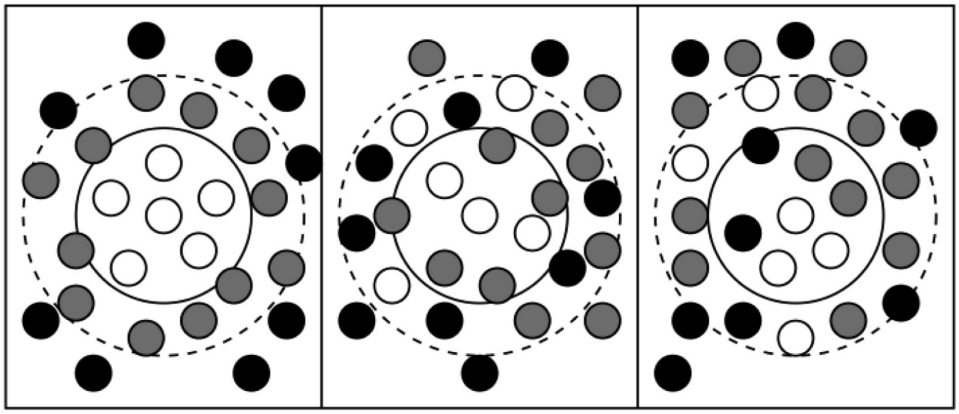


Figure 8.8: The Verlet list - The Verlet list on its construction, later, and too late. The potential cutoff range (solid circle), and the list range (dashed circle), are indicated. The list must be reconstructed before particles originally outside the list range (black) have penetrated the potential cutoff sphere.

8.5.5 Constrains

It is quite common practice in classical computer simulations not to attempt to represent intramolecular bonds by terms in the potential energy function, because these bonds have very high vibration frequencies (and arguably should be treated in a quantum mechanical way rather than in the classical approximation). Instead, the bonds are treated as being constrained to have fixed length. In classical mechanics, constraints are introduced through the Lagrangian or Hamiltonian formalisms. The general principle of the formalism is the Hamiltons variational principle:

Let L be the Lagrangian of the system; Hamiltons principle states that the physical trajectory taken by that system satisfies:

$$\partial \int_{t_1}^{t_2} L dt = 0$$

for any pair of times t_1, t_2, where variations are taken with respect to q and are fixed at the endpoints. This means that a trajectory followed by a mechanic system in the phase space is the one that minimize the integral $\int L dt$ where L is defined as L=K-V. Lagrangian formalism allow to treat constrains in a simply and direct way for system in which constrains are only on the position (holonomic) and can be written as $g_l(\{r_i\}) = 0$. Lagrangian equation becomes:

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{r}_i}\right) - \frac{\partial L}{\partial r_i} = \sum_{l=1}^{M} \lambda_l \frac{\partial g_l}{\partial r_i}$$

where M are Lagrange multiplier and λ_l are function of N coordinate and velocities. This means the right equation part could be considered as a generalized force that produces the same effects of imposed constrains. It is easy to derive an exact expression for the multiplier M from the above equations; However, this exact solution is not what we want: in practice, since the equations of motion are only solved approximately, in discrete time steps, the constraints will be increasingly violated as the simulation proceeds. The breakthrough in this area came with the proposal to determine the constraint forces in such a way that the constraints are satisfied exactly at the end of each time step. For the original Verlet algorithm, this scheme is called SHAKE, which calculates the constraint forces λgi necessary to ensure that the end-of-step positions r_i satisfy: $g_l(\{r_i\}) = 0$ An alternative constraint method, LINCS (Linear Constraint Solver) was developed in 1997 by Hess, Bekker, Berendsen and Fraaije(120)LINCS applies Lagrange multipliers to the constraint forces and solves for the multipliers by using a series expansion to approximate the inverse of the Jacobian. This approximation only works for matrices with Eigenvalues smaller than 1, making the LINCS algorithm suitable only for molecules with low connectivity. It is important to realize that a simulation of a system with rigidly constrained bond lengths, is not equivalent to a simulation with, for example, harmonic springs representing the bonds, even in the limit of very strong springs. A subtle, but crucial, difference lies in the distribution function for the other coordinates. If we obtain the configurational distribution function by integrating over the momenta, the difference arises because in one case a set of momenta is set to zero, and not integrated, while in the other integration is performed, which may lead to an extra term depending on particle coordinates. This is frequently called the metric tensor problem.

8.5.6 MD in NPT Ensemble

The ergodic hypothesis was introduced for a system described by the microcanonical distribution *NVE*. However, the conditions of constant volume V, number of particles N and total energy E do not fit

those in which experiments are usually made. Thus, it is necessary to define schemes allowing for the evolution of systems under conditions of constant volume and temperature (NVT), or constant pressure and temperature (NPT), corresponding to typical real-life situations. In this work, the simulations were performed in the NPT ensemble. To simulate the systems in such ensemble, thermostat and barostat algorithms are required to control the temperature and pressure during the MD run.

8.5.7 Nosé-Hoover thermostat

A way to sample the NVT ensemble within the framework of MD was introduced about twenty years ago by Nosé (219; 220) and reformulated by Hoover. This method modifies Newton equation of motion by adding two non physical variables, thus introducing the following non-Hamiltonian dynamical system where there is added a fictitious degree of freedom, with mass equal to Q. The new extended Hamiltonian has the form:

$$H* = \sum_{i=1}^{N} \frac{p_i^2}{2m_i} + \varphi(r_i) + \frac{Q}{2}\zeta^2 + gkT \ln S$$

where $\{r_i\}$,$\{p_i\}$ are coordinates and momenta of the N particles with masses m_i (as previously defined), and S and ζ are coordinates and momenta of fictitious atoms. If ϕ is the interaction potential and g are the degree of freedom, the new equation of motion are:

$$\begin{aligned} \dot{r}_i &= \tfrac{p_i}{m_i} \\ \dot{p}_i &= -\tfrac{d\varphi}{dt} - \zeta p_i \\ \tfrac{d\zeta}{dt} &= \tfrac{\sum \frac{p_i^2}{m_i} - g k_B T}{Q} \end{aligned}$$

These equations sample a microcanonical ensemble in the extended system, however, the energy of the real system is not constant. Nevertheless it can be shown, that the equations of motion sample a canonical ensemble in the real system. The parameter Q controls the strength of the coupling to the thermostat: high values result into a low coupling and viceversa. Although any finite (positive) mass is sufficient to guarantee in principle the generation of a canonical ensemble, if Q is too large, the canonical distribution will only be obtained after very long simulation times. On the other hand, too small values (tight coupling) may cause high-frequency temperature oscillations.

8.5.8 Berendsen thermostat

A weaker formulation of this approach is the Berendsen thermostat.(26; 27) To maintain the temperature the system is coupled to an external heat bath with fixed temperature T_0. The velocities are scaled at each step, such that the rate of change of temperature is proportional to the difference in temperature:

$$\frac{dT(t)}{dt} = \frac{1}{\tau}(T_0 - T(t))$$

where τ is the coupling parameter which determines how tightly the bath and the system are coupled together. This method gives an exponential decay of the system towards the desired temperature. The change in temperature between successive time steps is:

$$\Delta T = \frac{\delta t}{\tau}(T_0 - T(t))$$

Thus, the scaling factor for the velocities is:

$$\lambda^2 = 1 + \frac{\delta}{\tau}\left\{\frac{T_0}{T(t - \delta t/2}\right\}$$

In practice τ is used as an empirical parameter to adjust the strength of the coupling. Its value has to be chosen with care. In the limit $\tau \to \infty$ the Berendsen thermostat is inactive and the run is sampling a microcanonical ensemble. The temperature fluctuations will grow until they reach the appropriate value of a microcanonical ensemble. However, they will never reach the appropriate value for a canonical ensemble. On the other hand, too small values of τ will cause unrealistically low temperature fluctuations. If τ is chosen the same as the time step δt, the Berendsen thermostat is nothing else than the simple velocity scaling. Values of $\tau \approx 0$. 1ps are typically used in MD simulations of condensed-phase systems. The ensemble generated when using the Berendsen thermostat is not a canonical ensemble.

The Andersen method(6) was developed to adjust the pressure in a simulation of interacting particles. In the following description, only systems of pairwise interacting particles are treated. The method was later first extended to anisotropic coupling by Parrinello et al.(234) and later also to molecular systems by Nosé et al.(219; 220) Andersen proposed to replace the coordinates r_i by scaled coordinates ρ_i defined as:

$$\rho_i = r_i / V^{1/3}$$

Consider the new Lagrangian, in which a new variable Q appears:

$$L\left(\rho^N, \dot{\rho}^N, Q, \dot{Q}\right) = \frac{1}{2} Q^{2/3} \sum_{i=1}^{N} m_i \dot{\rho}_i^2 - \sum_{i<j}^{N} U(Q^{1/3} \rho_{ij}) + \frac{1}{2} M \dot{Q}^2 - p_0 Q$$

If we interpret Q as the volume V, the first two terms on the right are just the Lagrangian of the unscaled system. The third term is a kinetic energy for the motion of Q, and the fourth represent a potential energy associated with Q. Here p_0 and M are constants. A physical interpretation of the additional terms would be: Assume the system is simulated in a container and can be compressed by a piston. Thus, Q, whose value is the volume V, is the coordinate of the piston. p_0 V is the potential derived from an external pressure p_0 acting on the piston and M is the mass of the piston.

8.5.8.1 Parrinello-Rahman barostat

When simulating crystal structures, it is not sufficient only to scale the volume. Parinello and Rahman extended the method proposed by Andersen to let the simulation box also change it's shape.(234) Let us start with some notation: The cell can have an arbitrary shape, it's volume completely described by three vectors *a,b,c*. The vectors can have different lengths and arbitrary mutual orientations. An alternative description is obtained by arranging the vectors as $\{a, b, c\}$ to form a *3 x 3* matrix h whose columns are the latter vectors. The volume is given by:

$$V = \det h = a \cdot (b \times C)$$

The position r_i of a particle can be written in terms of h and a column of vector s_i, with components ξ_i, η_i and ζ_i as:

$$r_i = h s_i = \xi_i a + \eta_i b + \zeta_i c$$

with $0 \leq \xi_i, \eta_i, \zeta_i \leq 1$. The square of the distance between particle i and j is given by:

$$r_{ij}^2 = s_{ij}^T G s_{ij}$$

where the metric tensor G is $G = h^T h$. Using the latter notation, the Lagrangian can be written as:

$$L = \frac{1}{2} \sum m_i \dot{s}_i^T G \dot{s}_i - \sum \sum U(r_{ij}) + \frac{1}{2} M Tr(\dot{h}^T \dot{h}) = pV$$

Deriving the equations of motion is similar to the isotropic case from Andersen.

8.6 Ab Initio Molecular Dynamic: The electronic structure problem

Within the semi-classical approximation, we have arrived to:

$$\begin{aligned} i\hbar\tfrac{\partial\Psi}{\partial t} &= -\sum_i \tfrac{\hbar^2}{2m_e}\nabla^2_{r_i}\Psi + V(r,R(t))\Psi = H_e\Psi \\ M_I\ddot{R}_I(t) &= -\nabla_I \int dr\Psi^* H_e\Psi \end{aligned} \tag{8.21}$$

the knowledge of the electronic structure is needed to calculate forces acting on the nuclei. In the following, I describe the main methods used for solving the electronic structure problem.

8.6.1 Time-space separation

Within the TDSCF scheme derived in the previous section, it is relevant noticing that $\mathcal{H}_e$ depends on time only *parametrically* through the positions of nuclei. Whenever the Hamiltonian does not depend explicitly on time, it is possible to formally separate the variables and reduce to a time-independent eigenvalue problem. In fact, we can cast the electronic wavefunction Ψ as simple product:

$$\Psi(\{\mathbf{r}_i\};\{\mathbf{R}_I\},t) = \psi(\{\mathbf{r}_i\};\{\mathbf{R}_I\})\, f(t)\,, \tag{8.22}$$

where $\mathbf{R}_I$ are instantaneous positions at time t, and ψ and f satisfy the following set of equations (obtained substituting the above expression in eq. 8.21):

$$i\hbar\frac{d}{dt}f(t) = Ef(t) \tag{8.23}$$

$$\mathcal{H}_e\psi(\{\mathbf{r}_i\};\{\mathbf{R}_I\}) = E\psi(\{\mathbf{r}_i\};\{\mathbf{R}_I\})\,. \tag{8.24}$$

A *particular* solution of the Time-Dependent Schrödinger Equation (TDSE) is thus the product of a sinusoidal wave in time and a function satisfying the eigenvalue equation 8.24, which is called time-independent Schrödinger equation (TISE):

$$\Psi\left(\{\mathbf{r}_i\};\{\mathbf{R}_i\},t\right) = \psi\left(\{\mathbf{r}_i\};\{\mathbf{R}_i\}\right)\left[fe^{-iEt/\hbar}\right] \tag{8.25}$$

where $E = \langle\mathcal{H}_e\rangle$ is the energy of the electronic system in the nuclear configuration $\{\mathbf{R}_I\}$. At this point we have a *whole series* of solutions because generally there will be multiple values of E for which eq. 8.24 has solutions for ψ. As the time-dependent Schrödinger equation is linear in time, the general solution will be simply given by a linear combination of the various independent solutions:

$$\Psi = \sum_{k=0}^{\infty} f_k(t)\,\psi_k(\{\mathbf{r}_i\},\{\mathbf{R}_I\})\,, \tag{8.26}$$

where $f_k(t) \equiv f_k\, e^{-iE_k\, t/\hbar}$ and ψ_k is the eigenfunction corresponding to energy E_k. Thus, system time evolution is entirely described by the set of coefficients $f_k(t)$. In fact, the occupation of eigenstates at any time is given by $|f_k(t)|^2$ (with $\sum_k |f_k(t)|^2 = 1$), while transitions are described *via* the cross-terms $f_k f_{l\neq k}$.

8.6.2 Methods for solving Time Independent Schrödinger Equation

A common approach for solving Eq. 8.24 consist in writing the total electronic wavefunction as a product of *single-particle* wavefunctions. The work presented in this thesis has profited from Density Functional Theory (211) (DFT) and Hartree-Fock (HF) with Møller-Plesset (MP) 2^{nd} order corrections. Both of these methods include electronic correlation effects, and allows treatment of relatively large systems with a reasonable computational cost.

8.6.2.1 Hartree-Fock Methods

The Hartree-Fock method takes into account the Pauli principle for electrons writing the total electronic wavefunction as *single* (antisymmetric) Slater determinant of the spin-orbitals $\psi_i(\mathbf{x}) = \phi_i(\mathbf{r})\,\sigma(s)$ (232), where $\sigma(s) = \alpha(s)$ or $\beta(s)$:

$$\Psi_{HF} = \frac{1}{\sqrt{(N!)}} \begin{vmatrix} \psi_1(\mathbf{x}_1) & \psi_2(\mathbf{x}_1) & \cdots & \psi_N(\mathbf{x}_1) \\ \psi_1(\mathbf{x}_2) & \psi_2(\mathbf{x}_2) & \cdots & \psi_N(\mathbf{x}_2) \\ \vdots & \vdots & & \vdots \\ \psi_1(\mathbf{x}_N) & \psi_2(\mathbf{x}_N) & \cdots & \psi_N(\mathbf{x}_N) \end{vmatrix} \tag{8.27}$$

Minimizing the expectation value of the Hamiltonian $\mathcal{H}_e$ with respect to the set $\{\psi_i\}$ subject to the orthonormalization conditions $\langle\psi_i|\psi_j\rangle = \int d\mathbf{x}, \psi_i^*(\mathbf{x})\psi_j(\mathbf{x}) = \delta_{ij}$ gives, after diagonalization through a unitary operator $\mathbf{U}$, the canonical Hartree-Fock system of equations:

$$F_i\phi_i = \epsilon_i\phi_i\,, \tag{8.28}$$

with

$$F_i = \underbrace{-\frac{\nabla_i^2}{2} + \sum_I \frac{Z_I \mathrm{e}}{|\mathbf{r_i} - \mathbf{R_I}|}}_{h_i} + \sum_{j=1}^{N}(J_j - K_j)\,. \tag{8.29}$$

The so-called Fock operator F_i is an effective one-electron operator describing the kinetic energy of an electron, the attraction to all the nuclei, and the repulsion between electrons, through the Coulomb and exchange operators:

$$J_j(\mathbf{x}_1)\,\phi_i(\mathbf{x}_1) \equiv \sum_{k=1}^{N} \int d\mathbf{x}_2 \phi_k^*(\mathbf{x}_2)\phi_k(\mathbf{x}_2)\,\mathbf{g}_{12}\,\phi_i(\mathbf{x}_1)\,, \tag{8.30}$$

$$K_j(\mathbf{x}_1)\,\phi_i(\mathbf{x}_1) \equiv \sum_{k=1}^{N} \int d\mathbf{x}_2 \phi_k^*(\mathbf{x}_2)\phi_i(\mathbf{x}_2)\,\mathbf{g}_{12}\,\phi_k(\mathbf{x}_1)\,, \tag{8.31}$$

with

$$\mathbf{g}_{ij} \equiv \frac{1}{|\mathbf{r}_i - \mathbf{r}_j|}\,. \tag{8.32}$$

The exchange operator has a non-local character, and is the term that accounts for the exclusion principle of Pauli. The expectation value of the Fock operator

$$\varepsilon_l = \langle\phi_l|F_l|\phi_l\rangle \tag{8.33}$$

can be interpreted as the energy of the l-th MO, which in the limit of frozen orbitals is equal to minus the ionization energy I_l of the l-th electron (Koopmans' theorem (162)). The total energy

$$E = \sum_{i=1}^{N} \varepsilon_i - \frac{1}{2} \sum_{ij=1}^{N} (J_{ij} - K_{ij}) + V_{nn} \tag{8.34}$$

is not simply the sum of MO energies, because the Fock operator contains terms describing the repulsion of each MO to *all* other electrons, and thus the sum over ε_i counts the electron-electron repulsion twice, which have to be corrected by the second term. Moreover, this total energy cannot be exact, as the electron-electron repulsion is only accounted for in an average fashion, due to the approximation of a single Slater determinant as the trial wave function. The absence of correlation among electrons can be included within a perturbative scheme, like the one of Møller-Plesset.

8.6.2.2 Møller-Plesset perturbation theory

In the Møller-Plesset scheme (141; 211) the unperturbed Hamiltonian H_0 is taken to be a sum over Fock operators. As this sum counts twice the (average) electron-electron repulsion, the perturbation H_1 becomes the exact V_{ee} operator minus twice $\langle V_{ee} \rangle$ (also called *fluctuation potential*):

$$H_0 = \sum_{i=1}^{N} F_i = \sum_{i=1}^{N} \left(h_i + \sum_{j=1}^{N} (J_j - K_j) \right) = \sum_{i=1}^{N} h_i + 2 \langle V_{ee} \rangle \tag{8.35}$$

$$H_1 = V_{ee} - 2 \langle V_{ee} \rangle \tag{8.36}$$

The zero-order wave function is the HF determinant, while the first (MP1) order correction to the energy is given by the average electron-electron repulsion changed in sign. Electron correlation enters at the MP2 level, and involves only a sum over doubly excited determinants (if canonical HF orbitals are used) (141):

$$E^{\mathrm{MP2}} = \sum_{i<j}^{occ} \sum_{a<b}^{vir} \frac{\int d\mathbf{r}_1 d\mathbf{r}_2 \varphi_i(\mathbf{r}_1) \varphi_j(\mathbf{r}_2) \mathbf{g}_{12} [\varphi_a(\mathbf{r}_1) \varphi_b(\mathbf{r}_2) - \varphi_b(\mathbf{r}_1) \varphi_a(\mathbf{r}_2)]}{\varepsilon_i + \varepsilon_j - \varepsilon_a - \varepsilon_b} \tag{8.37}$$

The MP2 correction typically recovers 80-90% of the correlation energy, at a cost roughly twice as that for solving HF equations in practical calculations (this because only two-electron integrals corresponding to two combination of two occupied and two virtual MOs are required in eq. 8.37). Moreover, for "well-behaved" systems, MP2 usually gives better results than MP3 (141). Including higher terms in the perturbation is not very common, as other methods become competitive (141), like Configuration Interaction (CI), which has the advantage of being intrinsically multi-reference.

8.6.2.3 Density Functional Theory

Density Functional Theory (DFT) is a rigorous method to find the ground state of many particle system (122; 232). The main idea lies in assumption that the ground-state properties of a quantum system of N particles can be described starting from its density $\rho(\mathbf{r})$

$$\rho(\mathbf{r}) = N \int |\Psi(\mathbf{r}_1, \mathbf{r}_2, ...\mathbf{r}_N)|^2 d\mathbf{r}_2 d\mathbf{r}_3 ... d\mathbf{r}_N \, . \tag{8.38}$$

This has two main advantages over the other techniques:

- density is an observable that can be easily measured and visualized.
- the dimensionality of the problem is reduced from $3N$ to 3, as the density is a function of space.

The Hohenberg-Kohn theorems. The use of the density as fundamental quantity is based on the two Hohenberg-Kohn theorems, enunciated in the early sixties (122). The first theorem demonstrates that, given a Hamiltonian characterized by a general external potential V_{ext}, the ground-state density $\rho(\mathbf{r})$ associated to it is unique. As V_{ext} univocally determines the Hamiltonian of the system, it follows that the ground state wavefunction, and thus all the observables are functionals of the density ρ. The second theorem provides a variational principle for the ground state density: given any trial density $\bar{\rho} > 0$ for which $\int \bar{\rho}(\mathbf{r})d\mathbf{r} = N$, it follows that $E[\bar{\rho}] \geq E[\rho]$. From this result, one can get a variational equation to obtain the ground-state energy. Let apply the Hohenberg-Kohn theorem to a system of N electrons in which the external potential is due to the nuclei. The energy in terms of the electronic density reads:

$$E[\rho] = T_e[\rho] + V_{ee}[\rho] + V_{eN}[\rho] + V_{NN} = F[\rho] + \int d\mathbf{r}\rho(\mathbf{r})V_{ext} \tag{8.39}$$

where $F[\rho] = T_e[\rho] + V_{ee}[\rho] = \langle \psi \left| T_e + V_{ee} \right| \psi \rangle$ is a universal functional independent from the external potential $V_{ext}[\rho] = V_{eN}[\rho] + V_{NN}$.

Applying to ρ the stationary principle

$$\delta \left\{ E[\rho] - \mu \left[\int \rho(\mathbf{r})d\mathbf{r} - N \right] \right\} = 0, \tag{8.40}$$

we obtain the Euler-Lagrange equation for the multiplier μ:

$$\mu = V_{ext}(\mathbf{r}) + \frac{\partial F[\rho]}{\partial \rho} \tag{8.41}$$

Although DFT is formally a rigorous method, the application of the variational principle requires in practice an explicit form of the functional F. Kohn and Sham suggested to decompose it in parts whose only the most important need to be treated exactly (161).

Kohn-Sham equations. The main idea of the Kohn-Sham method lies in the possibility of mapping a system of N interacting particles into an equivalent one of non-interacting bodies, characterized by the same ground state density (161). For such systems, the density can be written as a summation over single-particle contributions:

$$\rho(\mathbf{r}) = \sum_{i=1}^{N} |\phi_i^{KS}(\mathbf{r})|^2 \tag{8.42}$$

and the kinetic energy functional has an analytical expression:

$$T_0[\rho] = \sum_{i=1}^{N} \left\langle \phi_i^{KS} \left| -\frac{1}{2}\nabla^2 \right| \phi_i^{KS} \right\rangle \tag{8.43}$$

The functional $F[\rho]$ can be rewritten as:

$$F[\rho] = T_0[\rho] + V_H[\rho] + E_{xc}[\rho] \tag{8.44}$$

where $V_H = \frac{1}{2}\int d\mathbf{r}' \frac{\rho(\mathbf{r})\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}$ is the classical part of the particle-particle interaction, and the 'exchange-correlation' functional E_{xc} is defined as:

$$E_{xc}[\rho] = T[\rho] - T_0[\rho] + V_{ee}[\rho] - V_H[\rho] \tag{8.45}$$

Thus all the unknowns of the problem are put into E_{xc}, which sums the corrections in the kinetic energy and of the non-classical part of the particle-particle interaction. The echange-correlation term describes the lowering in energy gained by a system of interacting electrons with respect to the Fermi gas, and has therefore a negative sign. Formally E_{xc} can be written in terms of an exchange-correlation energy per particle ε_{xc}, which is itself functional of the total density:

$$E_{xc}[\rho] = \int d\mathbf{r}\rho(\mathbf{r})\varepsilon_{xc}[\rho] \tag{8.46}$$

Equation 8.41, turns out to be:

$$\mu = V^{KS}(\mathbf{r}) + \frac{\partial T_0[\rho]}{\partial \rho} \tag{8.47}$$

with

$$V^{KS} = V_{ext}(\mathbf{r}) + \int d\mathbf{r}' \frac{\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} + V_{xc}[\rho] \tag{8.48}$$

where we have defined the exchange correlation potential

$$V_{xc}(\mathbf{r}) = \frac{\delta E_{xc}[\rho]}{\delta \rho} \tag{8.49}$$

Eq. 8.47 says that we can solve the original problem by finding the ground state energy for a system of non-interacting electrons in a effective potential. The single-particle orbitals describing these electrons solve the self-consistent Kohn-Sham (KS) equations:

$$[\frac{1}{2}\nabla^2 + V_{KS}(\mathbf{r})]\varphi_i^{KS} = \varepsilon_i \varphi_i^{KS} \qquad i = 1, ..., N \tag{8.50}$$

The total energy of the system is not the sum of KS eigenvalues, but can be expressed as

$$E = \sum_{i=1}^{N} \varepsilon_i - \frac{1}{2}\int d\mathbf{r}d\mathbf{r}' \frac{\rho(\mathbf{r})\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} + E_{xc}[\rho] - \int d\mathbf{r} V_{xc}(\mathbf{r})\ \rho(\mathbf{r}) \tag{8.51}$$

The method of Kohn and Sham shifts the complexity of the problem on finding a suitable analytical formulation of the exchange-correlation functional. In practice this is not possible, so once again approximated expressions have been derived for E_{xc}.

Exchange-Correlation functionals

Local Density Approximation. An approximation for the exchange-correlation functional has been proposed already in the original paper by Hohenberg and Kohn (122; 232). They recover the idea beyond Thomas-Fermi approximation of the kinetic energy for an homogeneous electron gas, and apply it to the evaluation of $E_{xc}[\rho]$. The exchange-correlation energy density in $\mathbf{r}$ is assumed to be *local*, i.e. only depends on the value of ρ in $\mathbf{r}$ itself (here the name Local Density Approximation, LDA): $\varepsilon_{xc}[\rho] = \varepsilon_{xc}(\rho(\mathbf{r}))$. In addition, ε_{xc} is approximated by that of an homogeneous gas of electrons of density $\rho^{hom} = \rho(\mathbf{r})$ (in a uniform background of positive charge). Thus

$$\varepsilon_{xc}^{LDA}[\rho] = \varepsilon_{xc}^{hom}(\rho(\mathbf{r})) \tag{8.52}$$

The simplification introduced by LDA becomes clear if one divides ε_{xc} into exchange and correlation contributions ε_x and ε_c. In fact, for a homogeneous electron gas ε_x is known exactly (46; 122; 232), and is proportional to the cubic square of the density:

$$\varepsilon_x^{LDA}[\rho] = -\frac{3}{4}\left(\frac{3}{\pi}\right)^{1/3}\rho^{1/3}(\mathbf{r}) \tag{8.53}$$

The situation is more complicated for the correlation term, which has been determined analytically in the high and low density limit (47; 48), and by Quantum Monte Carlo calculations for intermediate states (52). Suitable (approximate) analytical formulas have been derived by Vosko, Wilk and Nusair (VWN correlation functional) (317) and by Perdew and Wang (PW) (239). The main reason behind success of LDA is most probably a partial cancellation of errors. In fact, LDA typically underestimates E_c but overestimates E_x, resulting in unexpectedly good values of E_{xc}. However, for molecular systems ε_x is underestimated by a factor of 10, leading to errors larger than the whole correlation energy (overestimated by a factor ~ 2), and bond energies up to ~ 25 kcal/mol larger than experimental values (141). In addition, LDA exhibits heavy deficiencies in describing hydrogen-bonds, which are crucial for studies on biologically relevant systems (281; 308).

Generalized Gradient Approximation. The General Gradient Approximation (GGA) successfully improves the accuracy of DFT by introducing the gradient[1] of the density in the functional form of E_{xc}:

$$E_{xc}^{GGA}[\rho] = \int d(\mathbf{r}) f(\rho(\mathbf{r}), \nabla\rho(\mathbf{r})) \tag{8.54}$$

Quite generally, GGAs functionals give good results for all the main bond types (covalent, ionic, metallic and hydrogen bonds) (see for example (46)). For Van der Waals interactions, however, common GGAs and LDA fail. To treat these weak interactions more specialized approaches have been developed (7; 143) but I will not treat them in detail here. The GGA functionals used are typically derived by fitting parameters on the properties of sets of molecules.

8.6.3 Basis Set approximation

In the actual implementations of (Post)HF or DFT-Kohn-Sham schemes, the MOs are usually expanded in terms of M_b basis functions of well-known behavior

$$\varphi_i = \sum_{\alpha}^{M_b} c_{i\alpha}\chi_\alpha \tag{8.55}$$

The mathematical problem is thus transformed into that of solving a secular matrix equation, in which the matrix elements are calculated from arrays of integrals evaluated for the given basis functions. Taking HF equations 8.28 as example, and expanding the eigenfunctions as above, one obtain the famous Roothaan-Hall equations (for a closed shell system) (141):

$$\mathbf{FC} = \mathbf{SC}\varepsilon \tag{8.56}$$

[1]Notice here the difference from the gradient-expansion-approximation (GEA), where one tries to systematically calculate gradient-corrections to LDA of the form $|\nabla\rho|$, $\nabla^2\rho$, $|\nabla\rho|^2$. In contrast to GGA, GEA shown no improvement with respect to LDA, because of the loss of some important properties of the exchange-correlation hole (141).

where **S** is the matrix describing basis set functions overlap ($S_{\alpha\beta} = \langle\chi_\alpha|\chi_\beta\rangle$) and the matrix elements of the Fock operator are written as sum of one-electron integrals and products of a *density* matrix with two-electron integrals

$$F_{\alpha\beta} = \langle\chi_\alpha|\mathbf{F}|\chi_\beta\rangle = h_{\alpha\beta} + \sum_{\gamma\delta} G_{\alpha\beta\gamma\delta}\, D_{\gamma\delta} \tag{8.57}$$

$$h_{\alpha\beta} = \langle\chi_\alpha|\mathbf{h}|\chi_\beta\rangle = \int dr_1 \chi_\alpha(1) \frac{-\nabla^2}{2} \chi_\beta(1) + \sum_a^N \int dr_1 \chi_\alpha(1) \frac{Z_a}{|R_a - r_1|} \chi_\beta(1) \tag{8.58}$$

$$G_{\alpha\beta\gamma\delta} = \langle\chi_\alpha(1)\chi_\beta(2)|\mathbf{g}_{12}|\chi_\gamma(1)\chi_\delta(2)\rangle = \int d\mathbf{r}_1 \chi_\alpha(1)\chi_\beta(2) \frac{1}{\mathbf{r}_1 - \mathbf{r}_2} \chi_\beta(1)\chi_\delta(2) \tag{8.59}$$

$$D_{\gamma\delta} = \sum_j^{occ.MO} c_{\gamma j} c_{\delta j} \tag{8.60}$$

Essentially two philosophies exist for the construction of a basis set, one using localized atomic orbitals (AO), the other delocalized plane-waves (PW) for the expansion of MOs.

8.6.3.1 Localized basis sets

The basis sets are composed by localized and usually atomic-centered functions [1], firstly introduced by Slater, that resemble hydrogen-like eigenfunctions

$$\chi_{\zeta,n,l,m}(r,\theta,\varphi) = NY_{l,m}(\theta,\varphi) r^{n-1} \mathrm{e}^{-\zeta r} \tag{8.61}$$

Slater Type Orbitals (STO) are certainly the best suited for electronic structure calculations. However, the evaluation of three and four center two-electrons integrals is very expensive using such an exponential functions. So, for practical purposes STO are almost universally replaced by Gaussian Type Orbitals (GTO) (40)

$$\chi_{\zeta,n,l,m}(r,\theta,\varphi) = NY_{l,m}(\theta,\varphi) r^{2n-2-l} \mathrm{e}^{-\zeta r^2} \tag{8.62}$$

In fact, the product of two Gaussians located at different centers have the property of being a *Gaussian* centered at the intermediate position, that greatly improves the efficiency in calculating two-electrons integrals. Obviously a single GTO does not reproduce as well as an STO the proper behavior of the wavefunction (in particular near to the nucleus and for large r), so three times as many GTOs as STOs are roughly required to reach a given accuracy. This theoretical disadvantage is more than compensated by the overall gain in computational time. The quality of a calculation depends obviously on the number of functions used in the expansion. A double *zeta basis* (DZ) set, in which the number of basis functions is twice what is needed to contain all the electrons of neutral atoms, is considered "good" for organic molecules. Most often only valence orbitals are doubled, while core states are described with the smallest number of functions possible (minimum basis set), which gives the *double zeta valence split basis* (VDZ). Often polarization and diffuse functions are added to the basis set to improve the description of electronic correlation and polarization, and systems with loosely bound electrons (see Ref. (141) for further details). MOs are thus expanded as linear combination of a given number of

[1] For a recent and exhaustive description of AO type basis sets see e.g. the book by Jensen (141).

GTOs with different exponents ζ (*primitives*, PGTO). As these are determined by an energy-based variational procedure, most of them are "well-tuned" on core-states, which are energetically but not chemically relevant. To improve efficiency *contracted* basis sets have been introduced. The idea is to combine a given set of primitives into a smaller set of (contracted, CGTO) orbitals, each one being a linear combination of a given number of PGTO with *fixed* coefficients. The acronyms DZ, VDZ, etc. always refer to the number of contracted basis functions. Calculations reported in this thesis have been performed using the VDZ Pople-style k-nlmG basis set described in the next paragraph.

Pople-style *k-nlmG* basis sets. In this basis set the k indicates the number of PGTOs used for representing the core orbitals, while nlm indicates both how many functions the valence orbitals are split into, and how many PGTOs are used for their representation. Two values (nl) indicate a split valence, three (nlm) a triple split valence. Polarization functions are specified after the G. The most used basis set of this kind is the 6-31G (116), in which the core orbitals are a contraction of six PGTOs, the inner part of valence orbitals is contraction of three, and the outer part is represented by one PGTO.

8.6.3.2 Plane waves

Following the Bloch theorem (14) for periodic systems, a one-particle wave-function can be written as Fourier's series:

$$\varphi^{\mathbf{k}}(\mathbf{r}) = \frac{1}{\sqrt{V}} e^{i\mathbf{k}\cdot\mathbf{r}} \sum_{\mathbf{g}} c_j^{\mathbf{k}}(\mathbf{g}) e^{i\mathbf{g}\cdot\mathbf{r}} \tag{8.63}$$

where V is the volume of the cell, $\mathbf{k}$ vectors belong to the first Brillouin zone, $\mathbf{g}$ is a reciprocal lattice vector, c is the first Fourier component of the plane waves expansion, and the summation is extended to infinite lattice vectors. In the treatment of isolated clusters with a low symmetry, such as, organic molecules or the active sites of enzymes, the Γ-point approximation ($\mathbf{k}$=0) still guarantees a good accuracy, leading to a relevant reduction of the computational cost. The simulation of isolated clusters within a periodic boundary condition scheme needs some care, as self-interaction among replicas has to be cancelled.

Pseudopotentials. The greatest drawback in using a plane-wave basis-set comes from the impossibility, from a practical point of view, of describing core electrons within a reasonable computational expense. Indeed, the sharp spatial oscillations of their wave-functions near to the nuclei would require an extremely high number of plane-waves for an accurate characterization. On the other hand, the core levels are well separated in energy from valence electrons, and, at a first level of approximation, do not play any role in the chemical properties of molecular systems. Thus, the core electron orbitals can be frozen in the KS equations and only the valence electrons are described explicitly. The core-valence electron interactions are implicitly included into the nuclear potential, which becomes an "effective-potential" or "pseudopotential". Pseudopotentials are usually derived from all electron (AE) atomic calculations, and several recipes have been proposed to date. In the work presented in this thesis "norm-conserving" pseudopotentials derived from the Martins-Troullier (MT) method (305) have been used. Pseudopotentials have to satisfy the following conditions:

- The valence pseudo-wave-function should not contain any radial nodes.

- The valence AE and pseudopotential eigenvalues from the radial KS equations must be the same:

$$\varepsilon_\ell^{PP} = \varepsilon_\ell^{AE} \tag{8.64}$$

  where ℓ is the angular momentum.

- The pseudo and AE atomic radial wave-functions must be equal for r greater than a chosen cut-off distance r_{cut}.

These three conditions ensure that the pseudo-atom behaves like the real one in the region of interaction with other atoms while forming chemical bonds. Other conditions are the following:

- The integrated electron density within the cut-off radius for the two wave-functions must be the same. This requirement guarantees the transferability and the norm conserving rule of the MT pseudopotential.
- At $r = r_{cut}$, the pseudo wave-function and its first four derivatives should be continuous.
- The pseudopotentials should have zero curvature at the origin.

With these conditions, the general form for a pseudopotential wave-function is:

$$\varphi_\ell^{PP}(r) = \begin{cases} \varphi_\ell^{AE}(r) \ ; \ r > r_{cut} \\ r^\ell e^{p(r)}; \ r \leq r_{cut} \end{cases} \tag{8.65}$$

where $p(r) = c_0 + \sum_{i=1}^{6} c_i r^{2i}$, and the coefficients are obtained by imposing the first three conditions. The functional form of the pseudopotential is

$$V_{pseudo} = V_{val}(r) + \sum_{m,l} |Y_{l,m}\rangle V_l(r) \langle Y_{l,m}| \tag{8.66}$$

where $|Y_{l,m}\rangle$ are spherical harmonics. The "semilocality" of this functional form (local in the radial coordinate, non local in the angular ones), implies an increase in the computational cost. This difficulty can be overcome by using the method of Kleinman-Bylander (159), which implies addition and subtraction of an "ad-hoc" radial function $V_L(r)$ to the pseudopotential, leading to a new functional form, where the local and non-local parts can be completely separated.

8.7 Born-Oppenheimer approximation

We have seen that the Ehrenfest molecular dynamics scheme, eqs. 8.21, allows to propagate the electronic system by solving the time-dependent Schrödinger equation "on the fly", as the nuclear configuration changes under the force $\nabla_I \langle \mathcal{H}_e \rangle$. Unfortunately there is a major problem with practical implementation of the Ehrenfest scheme: the time scale and thus the time step used to integrate eqs. 8.21 simultaneously is dictated by the intrinsic dynamics of electrons. Now, typical vibrational and angular frequencies in biological systems rise up to $3000-4000$ cm^{-1} (for example bond frequencies in water are $\sim$3500 cm^{-1} (167)), which correspond to a timescale of $\tau_N \sim 10^{-14}$. This time interval is two order of magnitude larger than the maximum time step Δt_e^{max} necessary to integrate correctly electron dynamics. Thus, there is a bottleneck limiting the efficient implementation of such a simultaneous evolution of electronic and nuclear systems. A solution to this problem is the Born-Oppenheimer (BO) approximation, which was proposed in the early days of quantum mechanics (1927) (227). In the BO scheme the strong dynamical separation between electronic and nuclear motions is exploited to

increase the maximum time step used to propagate the nuclei. Since atoms are about three orders of magnitude heavier than electrons, the latter are supposed to follow *instantaneously* the motion of the nuclei, staying always in the same stationary state of the Hamiltonian. This stationary state will vary with the configuration of nuclei because of the Coulomb coupling between the two sets of degrees of freedom, but non-radiative transitions like those from *phonon-electron* interactions are negligible. This is obviously true only if the energy separation between the ground and the first excited state is larger than typical phonon energies. In this case, one can solve the time-independent Schrödinger equation at *fixed* positions of the nuclei, move them under the action of the effective electronic potential, and iterate the process. The equations describing the so-called Born-Oppenheimer approximation (for the ground state) are then [1].

$$\mathcal{H}_e \Psi_0 = E_0 \Psi_0 \tag{8.67}$$

$$M_I \ddot{\mathbf{R}}(t) = -\nabla_I \, min_\Psi \left\{ \langle \Psi_0 | \mathcal{H}_e | \Psi_0 \rangle \right\} \tag{8.68}$$

At opposite to Ehrenfest dynamics, now time dependence of the electronic structure is only *implicit* through the motion of nuclei. This allows for time steps $\Delta t_N^{max} \sim \tau_N/10$. However, the bottleneck of Born-Oppenheimer dynamics is that at each MD step the electronic wavefunction needs to be relaxed.

8.8 Car-Parrinello molecular dynamics

In 1985 Car and Parrinello developed a new scheme, based on the extended Lagrangian formalism and avoiding the optimization of the electronic wavefunction by introducing a second order *fictious* dynamics on the electrons. These latter are kept sufficiently close to the adiabatic surface, allowing for an increase of the time step by a factor ~ 10 with respect to Ehrenfest dynamics. The method is based on the observation that $\langle \Psi_0 | \mathcal{H} | \Psi_0 \rangle$ can be viewed not only as a function of $\{\mathbf{R}_I\}$, but also as a *functional* of the wavefunction Ψ_0, and thus of the set of one-electron orbitals $\{\psi_i\}$ used to build it. In this case, the force acting on these orbitals can be obtained from a functional derivative of a suitable Lagrangian containing $\langle \Psi_0 | \mathcal{H} | \Psi_0 \rangle$, like in classical mechanics for the nuclear motion. The Lagrangian $\mathcal{L}$ proposed by Car and Parrinello has the form

$$\mathcal{L} = \sum_{i=1}^{M} \frac{1}{2} M_I \dot{\mathbf{R}}_I^2 + \sum_{i=1}^{N} \frac{1}{2} \mu_i \left\langle \dot{\psi}_i \middle| \dot{\psi}_i \right\rangle - \langle \Psi_0 | \mathcal{H}_e | \Psi_0 \rangle + constraints \tag{8.69}$$

where the first term is the kinetic energy of nuclei, and $\mu_i = \mu$ are the "fictious masses" assigned to orbitals; the second term represents the fictious kinetic energy associated to them (the sum is on the occupied orbital only). The (holonomic) constraints act in general on both the orbitals (e.g. to guarantee orthonormality) and on the nuclei (e.g. if one would perform molecular dynamics with geometric restraints). The dynamics is described by the Euler-Lagrange equations associated to $\mathcal{L}$

$$\frac{d}{dt} \frac{\partial \mathcal{L}}{\partial \dot{\mathbf{R}}_I} = M_I \ddot{\mathbf{R}}_I = \frac{\partial \mathcal{L}}{\partial \mathbf{R}_I} = -\frac{\partial}{\partial \mathbf{R}_I} \langle \Psi_0 | \mathcal{H}_e | \Psi_0 \rangle + \frac{\partial}{\partial \mathbf{R}_I} \{constraints\} \tag{8.70}$$

$$\frac{d}{dt} \frac{\partial \mathcal{L}}{\partial \dot{\psi}_i^*} = \mu_i \ddot{\psi}_i^* = \frac{\partial \mathcal{L}}{\partial \psi_i^*} = -\frac{\partial}{\partial \psi_i^*} \langle \Psi_0 | \mathcal{H}_e | \Psi_0 \rangle + \frac{\partial}{\partial \psi_i^*} \{constraints\} \tag{8.71}$$

[1] Note that Born-Oppenheimer approximation slightly differs from the so-called "adiabatic" one for presence in the latter of a diagonal correction term containing the expectation value of the nuclear kinetic energy operator on the electronic wavefunction (141).

Note that if $|\mu\ddot{\psi}_i| \to 0$ eq. 8.71 reduces to a stationary problem, and the electronic system will stay on the Born-Oppenheimer surface (no forces acting on orbitals), corresponding to the true equilibrium dynamics. Higher is the fictious kinetic energy $T_e = \sum_{i=1}^{N} \frac{1}{2}\mu_i \left\langle \dot{\psi}_i \middle| \dot{\psi}_i \right\rangle$, more the electrons will be far from the minimum energy configuration. In particular, a ground state wavefunction optimized at time t_0 will stay close to its ground state if it is kept at sufficiently low temperature. The only quantity one can change to ensure this condition is μ, often called "adiabacity parameter" (204). If μ and τ are chosen consistently the energy flow between electronic and nuclear subsystems is slow enough to cause no drift in T_e [1], thus conserving the "physical" energy E_{phys}:

$$E_{phys} = E_{tot} - T_e = T_I + V_e = \sum_{i=1}^{M} \frac{1}{2} M_I \dot{\mathbf{R}}_I^2 + \langle \Psi_0 | \mathcal{H}_e | \Psi_0 \rangle \tag{8.72}$$

The choice of a reasonable fictious mass μ meets two opposite requirements. In fact, considering a simple harmonic model for the electronic system around the BO surface, (discrete) excitations frequencies are given by

$$\omega_{ij}^{e} = \sqrt{\frac{2(\varepsilon_i^* - \varepsilon_j)}{\mu}} \tag{8.73}$$

where ε^* and ε are energy levels of unoccupied and occupied orbitals, respectively. If ω_{max}^{n} is the maximum vibrational frequency of the nuclear system, in order to perform adiabatic dynamics it must be $\omega_{gap}^{e} >> \omega_{max}^{n}$. As the only tunable parameter is μ, one could decrease it arbitrarily to increase the frequency of the gap ω_{min}^{e}. However, decreasing μ stretches the entire spectrum $\{\omega_{ij}^{e}\}$ and in particular increases ω_{max}^{e}, which is inversely proportional to the maximum time step. Typical values of μ are in the range $500 - 1500$ a.u., which allow for a time step of about $5 - 10$ a.u. ($0.12 - 0.24$ fs). For calculations discussed here we used $\mu = 600$ a.u. and a time step of 5 a.u.

8.9 Hybrid Models

Pure quantum calculations are today restricted to the treatment of at most a few hundreds of atoms. classical molecular mechanics, on the other hand, can deal with systems containing $10^5 - 10^6$ atoms, but cannot take into account the quantum nature of chemical bonds. Since most of times the relevant chemistry of a biological process is restricted to a small subset of atoms, hybrid schemes have been developed that model different parts of the system at a different level of theory modeling (17; 89; 175; 323). These schemes allow to evaluate the effect of the biological environment on chemical processes, and represents thus an improvement over a quantum calculation *in vacuo*. In particular a widely adopted approach is to partition the system into two regions and to treat one at Quantum Mechanics (QM) and the other at Molecular Mechanics (MM) levels. Such approach, as implemented in the CPMD code (134), has been used in the works reported in this thesis, and is based on a single hybrid Hamiltonian:

[1] T_e actually performs two-frequency *bound* oscillations around a constant value. The first frequency is associated to the drag exerted by the nuclei, and it is in anti-phase with V_e oscillations, while the second is a small-amplitude high-frequency oscillation intrinsic to the fictious electronic dynamics. Note that having a nonvanishing masses, also the electrons dampen nuclear motion, causing a renormalization of the nuclear masses which can be important in the case of light atoms.

$$H = H_{QM} + H_{MM} + H_{QM/MM} \tag{8.74}$$

where H_{QM} is the quantum Hamiltonian, H_{MM} is the Molecular Mechanics Hamiltonian and $H_{QM/MM}$ is the Hamiltonian describing the interaction between the two subsystems. For the purpose of describing each term in eq. 8.74, let start by considering the total system (QM+MM) as described uniformly at quantum level and by (artificially) partitioning the system in the QM and MM regions. According to the Hohenberg-Kohn theorem, the total energy of the system is given by the following functional:

$$E[\rho] = T[\rho] + \int_\Omega V^{ext}(\mathbf{r})\rho(\mathbf{r})d\mathbf{r} + \frac{1}{2}\int\int \frac{\rho(\mathbf{r}_1)\rho(\mathbf{r}_2)}{r_{12}} d\mathbf{r}_1 d\mathbf{r}_2 + \frac{1}{2}\sum_{I,J}\frac{Z_I Z_J}{R_{IJ}} + E_{xc}[\rho] \tag{8.75}$$

where T and E_{xc} are the kinetic and exchange-correlation energy functionals, respectively; V^{ext} is the electrostatic potential of the nuclei, r_{12} and R_{IJ} the inter-electronic and internuclear distances and Z_I and Z_J the nuclear charge of atom I and J, respectively. By partitioning the total electronic density into the two contributions, $\rho_{QM+MM} = \rho_{QM} + \rho_{MM}$, the total energy can be rewritten as:

$$\begin{aligned} &E\left[\rho_{QM+MM}\right] = E[\rho_{QM}] + E[\rho_{MM}] + \int\int \tfrac{\rho_{QM}(\mathbf{r}_1)\rho_{MM}(\mathbf{r}_2)}{r_{12}} d\mathbf{r}_1 d\mathbf{r}_2 \\ &+ \sum_{\substack{I\in QM \\ J\in MM}} \tfrac{Z_I Z_J}{R_{IJ}} + E_{xc}^{NL} + T^{NL} \end{aligned} \tag{8.76}$$

where $E_{xc}^{NL} = E_{xc}[\rho_{QM} + \rho_{MM}] - E_{xc}[\rho_{QM}] - E_{xc}[\rho_{MM}]$ and $T^{NL} = T[\rho_{QM} + \rho_{MM}] - T[\rho_{QM}] - T[\rho_{MM}]$ arise from the nonlinearity of the kinetic and exchange and correlation functionals. In eq. 8.76, the term $E[\rho_{QM}]$ is treated at the quantum level, while each contribution to $E[\rho_{MM}]$ is approximated by using a force field, function of the nuclear coordinates only. In particular, as force fields are parametrized at a fixed value of the electronic density, the kinetic energy functional is an additive constant which can be neglected. In this context, the energy E_{xc} is approximated by a Lennard-Jones pair-additive potential:

$$E_{xc} \approx \sum_{I,J\in MM} 4 \in_{IJ} \left(\left(\frac{\sigma_{IJ}}{R_{IJ}}\right)^{12} - \left(\frac{\sigma_{IJ}}{R_{IJ}}\right)^{6} \right). \tag{8.77}$$

The remaining three term of eq. 8.75 describe the nuclear-electronic, electronic- electronic, and nuclear-nuclear charge densities electrostatic energies. In the force-field spirit, the total contribution to the energy is represented by the interaction energy among effective point charges located at the nuclear positions:

$$\int_\Omega V^{ext}(\mathbf{r})\rho(\mathbf{r})d\mathbf{r} + \frac{1}{2}\int\int \frac{\rho(\mathbf{r}_1)\rho(\mathbf{r}_2)}{r_{12}} d\mathbf{r}_1 d\mathbf{r}_2 + \frac{1}{2}\sum_{I,J\in MM}\frac{Z_I Z_J}{R_{IJ}} \approx \frac{1}{2}\sum_{I,J\in MM}\frac{q_I q_J}{R_{IJ}} \tag{8.78}$$

The chemical bonding cannot be described by solely Lennard-Jones and point charges electrostatic interaction energy, thus bonded terms have to be added to the MM energy. The most interesting part of eq. 8.76 concerns the interaction between the two subsystems. Using for T^{NL} and E_{xc}^{NL} the same approximation as above, we can express the interaction energy as:

$$
\begin{aligned}
E[\rho_{QM/MM}] &= \sum_{I\in MM} \int_\Omega \tfrac{q_I}{|\mathbf{R_I}-\mathbf{r}|}\rho_{QM}^{el+nuc}(\mathbf{r})d\mathbf{r} \\
&+ \sum_{\substack{I\in QM \\ J\in MM}} \in \left(\left(\tfrac{\sigma_{IJ}}{R_{IJ}}\right)^{12} - \left(\tfrac{\sigma_{IJ}}{R_{IJ}}\right)^{6}\right) \\
&+ \sum_{bonds} K_r(r-r_{eq})^2 + \sum_{angles} K_\vartheta(\vartheta-\vartheta_{eq})^2 \\
&+ \sum_{dihedrals} \tfrac{V_n}{2}[1+\cos(n\varphi-\gamma)]
\end{aligned}
\tag{8.79}
$$

where bonds, angles and dihedrals involve at least one QM atom. Notice that in eq. 8.79 the term T^{NL} has been neglected while for E_{xc}^{NL} the force-field approximation has been used. In this formulation the electrostatic potential provided by the effective classical point charges polarizes the QM electronic charge density. It should be noted that the presence of a discontinuous QM/MM interface introduces a series of artifacts. One of the most serious is the so-called *link atom problem*. When a chemical bond involves atoms on the two subsets, the QM system will contain by construction unsaturated valencies and has to be made chemically inert. Two approaches are mainly used to deal with this problem. The first consists of the use of a monovalent pseudopotential situated at the position of the MM involved in the bond crossing the QM/MM interface. In the CPMD code an analytical non-local pseudopotential of the Goedecker type (104) is used. The second approach introduces capping atoms (usually hydrogens) to saturate chemical bonds at the interface. It should be pointed out that the latter strategy introduces additional artifacts and a correction for the interactions between the "ghost" atoms and the classical environment is required. Furthermore the approximation of E_{xc}^{NL} by a purley classical term, i.e. not involving QM electronic degrees of freedom, results in the so-called *electron spill out* problem. Due to the fact that the MM region contains no electrons, those of the QM part are no longer repelled by closed-shell cores of the atoms belonging to the MM region. The effect of the missing Pauli repulsion is to artificially localize electrons on MM positive point charges. In order to avoid this artifact a pseudopotential-like approach can be applied by replacing the classical point charges Coulomb potential with a suitable function $v_I(|\mathbf{r}-\mathbf{R_I}|)$, which ensures the correct $1/r$ behaviour for large r and goes to a finite value for $r \to 0$. The first term of eq. 8.79 is thus replaced by the following quantity:

$$
E_{QM/MM}^{elec} = \sum_{I\in MM} q_I \int_\Omega \rho_{QM}^{el+nuc}(\mathbf{r}) v_I(|\mathbf{r}-\mathbf{R_I}|)d\mathbf{r}\,. \tag{8.80}
$$

In particular an appropriate form for $v_I(|\mathbf{r}-\mathbf{R_I}|)$ is:

$$
v_I(|\mathbf{r}-\mathbf{R_I}|) = \frac{r_{cI}^4 - r^4}{r_{cI}^5 - r^5}\,, \tag{8.81}
$$

where r_{cI} is the covalent radius of atom I. This functional form resembles that obtained by smearing the MM point charges into Gaussian charge distributions of finite width. In the context of plane-waves, the QM/MM scheme devised above cannot be used for practical purposes without an additional approximation. Indeed, the quantum charge distribution is distributed on a grid of $N_r \sim 100^3$ points, so that an exact evaluation of $E_{QM/MM}^{elec}$ would involve $N_r \times N_{MM}$ operations, with $N_{MM} \geq 10^5$. Therefore, this interaction term is split into a short and a long-range part, in a way reminishent of the Ewald method. The direct evaluation of the integral in eq. 8.80 is done only for a subset (NN) of MM atoms. The latter is defined in such a way as to include all non neutral atoms belonging to charge groups with at least one atom inside a shell of thickness R_c around any QM atom. The rest of MM atoms belong to the second shell. For those, the electrostatic interaction with the QM system is calculated using for the charge density of the QM system a multipolar expansion around

the geometrical center of the quantum system $\bar{\mathbf{r}}^\alpha = 1/N_{QM}\sum_I \mathbf{r}_I$ up to the quadrupole order. In particular, the electrostatic interaction Hamiltonian can be expressed as:

$$H_{elec} = \sum_{j\in NN} q_j \int d\mathbf{r}\rho(\mathbf{r})v_j(|\mathbf{r}-\mathbf{r}_j|) + H_{lr} \tag{8.82}$$

where H_{lr} is defined as:

$$H_{lr} = C\sum_{j\notin NN}\frac{q_j}{\tau_j} + \sum_\alpha D^\alpha \sum_{j\notin NN}\frac{q_j}{\tau_j^3}\tau_j^\alpha + \frac{1}{2}\sum_{\alpha\beta} Q^{\alpha\beta}\sum_{j\notin NN}\frac{q_j}{\tau_j^5}\tau_j^\alpha\tau_j^\beta \tag{8.83}$$

with $\tau_j^\alpha = r_j^\alpha - \bar{r}^\alpha$; C, D^α and $Q^{\alpha\beta}$ are the total charge, the dipole and the quadrupole of the electronic charge distribution, respectively. The potential entering into the Khon-Sham Hamiltonian is given by the functional derivative of H_{elec} with respect to the density ρ_{el}:

$$V(r) = \frac{\delta H_{elec}}{\delta\rho_{el}} = \sum_{j\notin NN}\frac{q_j}{\tau_j} + \sum_\alpha (r^\alpha - \bar{r}^\alpha)\sum_{j\notin NN}\frac{q_j}{\tau_j^3}\tau_j^\alpha + \frac{1}{2}\sum_{\alpha\beta} l[3(r^\alpha-\bar{r}^\alpha)(r^\beta-\bar{r}^\beta) - \delta^{\alpha\beta}|\mathbf{r}-\bar{\mathbf{r}}|^2]\sum_{j\notin NN}\frac{q_j}{\tau_j^5}\tau_j^\alpha\tau_j^\beta \tag{8.84}$$

The forces on the atom arising from H_{elec} are obtained by taking the derivatives with respect to the atomic positions. These are for the QM, NN and classical atoms not belonging to the NN set:

$$\begin{aligned} F_j^\gamma &= \frac{1}{N_{QM}} l[-\frac{5}{2}\sum_{\alpha\beta} Q^{\alpha\beta}\sum_{k\notin NN}\frac{q_k}{\tau_k^7}\tau_k^\alpha\tau_k^\beta\tau_k^\gamma + \sum_\alpha Q^{\alpha\gamma}\sum_{k\notin NN}\frac{q_k}{\tau_k^5}\tau_k^\alpha], \\ &\quad \text{for } j\in QM \end{aligned} \tag{8.85}$$

$$\begin{aligned} F_j^\gamma &= q_j\int d\mathbf{r}\rho(\mathbf{r})g_j(|\mathbf{r}-\mathbf{r}_j|)\frac{r^\gamma - r_j^\gamma}{|\mathbf{r}-\mathbf{r}_j|}, \\ &\quad \text{for } j\in NN \end{aligned} \tag{8.86}$$

$$\begin{aligned} F_j^\gamma &= -q_j\left[\left(-\frac{C}{\tau_j^3} - \frac{3}{\tau_j^5}\sum_\alpha D^\alpha\tau_j^\alpha - \frac{5}{2\tau_j^7}\sum_{\alpha\beta} Q^{\alpha\beta}\tau_j^\alpha\tau_j^\beta\right)\tau_j^\gamma + \frac{D^\gamma}{\tau_j^3} + \frac{1}{\tau_j^5}\sum_\alpha Q^{\alpha\gamma}\tau_j^\alpha\right] \\ &\quad \text{for } j\notin NN, QM \end{aligned} \tag{8.87}$$

where $g_j(r) = dv_j/dr$. This two level coupling scheme can also be refined introducing an intermediate third layer in which the charge density of the QM system is replaced by variational D-RESP charges. In the work exposed in the thesis I have used the implementation of QM/MM realized in the code CPMD (134), which has been interfaced to the AMBER (50) force field. The code is available for free with the CPMD package (www.cpmd.org).

8.10 Free Energy calculations

Several biological process, as some of those investigated here, occur in time scale which is much longer than that reachable by MD simulations. Enhanced sampled methods (49; 61; 73; 79; 140; 170; 173; 267; 301) allow to investigate rare events that may occur in these relatively long time scales. In some cases, one can identify few coordinates $\{s_j\}$, called *collective variables* (CVs) or *reaction coordinates*, which are believed to be relevant for the process under study. The free energy associated with such processes can be expressed then as a function of these CVs. Let us assume that we are working in the canonical ensemble NVT. The free energy then in the absolute and reduced representation reads:

$$\begin{array}{ccc} \mathbf{x} \equiv \{x_i\} \ , \ i = 1, ..., 3N & \leftrightarrow & F = -\frac{1}{\beta} \ln Z \\ \downarrow & & \downarrow \\ \mathbf{s} \equiv \{s_j\} \ , \ j = 1, ..., N_{CV} & \leftrightarrow & F(\mathbf{s}) = -\frac{1}{\beta} \ln P(\mathbf{s}) \end{array}$$

where $\beta = k_B T$, N is the number of particles, N_{CV} is the number of collective variables, $\mathcal{Z}$ is the canonical configuration partition function, and P is the probability to find the system in a state corresponding to a given value s of the reaction coordinates. For ergodic systems, this is equal to the distribution function along $\mathbf{s}$:

$$P(\mathbf{s}) = \lim_{t \to \infty} \frac{1}{t} \int_0^t dt \, \delta(\mathbf{s} - \mathbf{s}(t)) \equiv \rho(\mathbf{s}) = \frac{\int d\mathbf{x} \delta(\mathbf{s} - \mathbf{s}(\mathbf{x})) e^{-\beta U(\mathbf{x})}}{Z} \tag{8.88}$$

where the dependence of $\mathbf{s}$ on the coordinates $\mathbf{x}$ has been explicited.

8.10.1 Metadynamics

The main idea behind metadynamics (173) is to drive the evolution in the space of CVs by adding to the thermodynamic force, coming from the free energy $F(\mathbf{s})$, a force due to a history-dependent biasing potential $F_G(\mathbf{s}, t)$. The bias potential is constructed as a sum of Gaussians deposed along the trajectory of the CVs up to time t. The method can be seen as the finite temperature extension of the Wang and Landau algorithm (319), and is related in the spirit to taboo search (70), local elevation (130) and adaptive bias force (73). The most important property of metadynamics is that the biased trajectory proceeds by filling valleys in the free energy surface (173), so that the system tends to escape from every stable state. After a long enough time t the sum of the Gaussians deposed along the trajectory will counterbalance the free energy landscape, allowing to estimate the free energy itself:

$$\lim_{t \to \infty} F_G(\mathbf{s}, t) = F(\mathbf{s}) \tag{8.89}$$

It can be shown that the above relation is true under rather general assumptions (174). In real cases, the time needed for considering eq. 8.89 to be valid can be estimated by visual inspection of the trajectory of the CVs: when the $F_G(\mathbf{s}, t)$ counterbalances the free energy $F(\mathbf{s})$, the CVs have a diffusive behaviour. If one uses simultaneously 2 or more CVs is not necessary to know *a priori* the reaction path in metadynamics. The force due to the history-dependent potential naturally drives the system through the Lowest Free Energy Path (LFEP), i.e. the most likely reaction path (90); after the crossing of a barrier the system naturally goes towards a new and possibly unpredicted metastable state. For this reason the method has found a large use not only for predicting free energies, but also for accelerating rare events and investigating molecular mechanism of biological processes. Obviously, as for all the methods based on dimensional reduction, the chosen CVs must describe somewhat the process of interest. Nevertheless, many application have shown that choosing general and flexible CVs allows discovering unknown stable states and reaction mechanisms (135; 202).

Here I describe the "discrete" version of the algorithm, which has been introduced by Laio and Parrinello in 2002 (173). In discrete metadynamics the CVs are evolved step by step. A multidimensional Gaussian of width $\delta \mathbf{s} = (\delta \mathbf{s_1}, ..., \delta \mathbf{s_{N_{CV}}})$ and height w is deposed at position $\mathbf{s}(\mathbf{x}, t)$ every time *metastep* τ_G. Thus, at time t the free energy underlying the dynamics of the CVs is given by:

$$\tilde{F}(\mathbf{s}(\mathbf{x}), t) = F(\mathbf{s}(\mathbf{x})) + F_G(\mathbf{s}(\mathbf{x}), t) = F(\mathbf{s}(\mathbf{x})) + w \sum_{j=1}^{n_G^t} \prod_{i=1}^{N_{CV}} e^{-\left[\frac{s_i(\mathbf{x}) - s_i(\mathbf{x}(j\tau_G))}{\sqrt{2}\delta s_i}\right]^2} \tag{8.90}$$

where the number n_G^t of Gaussians deposed at time t is given by the integer closest to t/τ_G. In order to simplify matter the CVs space can be rendered approximately spherical by rescaling all the CVs s_i to their thermal fluctuations $\sqrt{(s_i - < s_i >)^2}$, which can be evaluated at the starting minimum through an unbiased dynamics. In this way spherical Gaussians can be used, for which the width is the same in all the directions, $\delta s_i = \delta s$, $\forall\, i$. From eq. 8.90 it can be seen that the dynamics of the CVs is driven by two forces:

- the thermodynamic force, evaluated at $\mathbf{s}^t = \mathbf{s}(\mathbf{x}(t))$:

$$f_i^{th}\Big|_t = -\frac{\partial}{\partial s_i^t} F(\mathbf{s})\,. \tag{8.91}$$

  Following Sprik and Ciccotti (282), these forces are estimated through the Constrained Reaction Coordinate Dynamics (CRCD) algorithm, adding to the normal Lagrangian of the system a restraining term $\sum_{i=1}^{N_{CV}} \lambda_i (s_i - s_i(\mathbf{x}(t))$. By averaging over the time, in the absence of inertial terms, the components of the thermodynamic force are given by $f_i^{th} =< \lambda_i >$. Thermodynamic forces are evaluated in the time between two subsequent hill depositions.

- the history-dependent force at time t, whose components are:

$$f_i^G\Big|_t = -\frac{\partial}{\partial s_i^t} w \sum_{j=1}^{n_G^t} \prod_{i=1}^{N_{CV}} e^{-\left[\frac{s_i(\mathbf{x}) - s_i(\mathbf{x}(j\tau_G))}{\sqrt{2}\delta s_i}\right]^2} \tag{8.92}$$

  which discourage the system to visit the same region in the CVs phase space.

The *metadynamics* of the walker in the CVs space thus is regulated by the following discrete equation of motion (I removed the superscript t for simplicity):

$$\begin{aligned} &\mathbf{s}(t+\tau_G) = \mathbf{s}(t) + \delta\mathbf{s} \cdot \frac{\widetilde{\mathbf{f}}}{|\widetilde{\mathbf{f}}|} \\ &\widetilde{\mathbf{f}} = \mathbf{f}^{th} + \mathbf{f}^G \end{aligned} \tag{8.93}$$

The equation was introduced firstly in Ref. (173). Subsequently, three corrections were applied in order to enhance accuracy and reduce sistematic errors:

1. To improve efficiency, the Gaussians are shifted with respect to the position of the walker. In fact, if the thermodynamic force at time t is evaluated at the same point where the Gaussian is placed, the total force felt by the walker will be the same as in the previous metastep. In order to compensate the thermodynamic force is better to depose the Gaussian at a distance $\delta\mathbf{s}$ from $\mathbf{s}(t)$ in the direction of the thermodynamic force.

2. To reduce the correlation induced by depositions with constant step, at every iteration the metastep is chosen randomly from a uniform distribution with two limiting values (e.g. $\delta\mathbf{s}$ and $1.5\,\delta\mathbf{s}$).

3. When the metadynamics is terminated F_G will present a bump in the region around the last hill; the spread of this bump depends on the correlation time of the metadynamics. In order to reduce these spatial correlations in the free energy the contributions of the Gaussians placed at the end of the dynamics are weighted less.

With the modification listed above, we arrive to:

$$\begin{aligned}
&\mathbf{s}(t+\tau_G) = \mathbf{s}(t) + \Delta\mathbf{s}\cdot\frac{\tilde{\mathbf{f}}}{|\tilde{\mathbf{f}}|},\ \Delta\mathbf{s}\in[\delta\mathbf{s},\ \ \alpha\delta\mathbf{s}],\ \alpha>1\\
&\tilde{f}_i = f_i^{th} - \frac{\partial}{\partial s_i^t} w \sum_{j=1}^{n_G^t}\prod_{i=1}^{N_{CV}} e^{-\left|\frac{s_i(\mathbf{x})-s_i(\mathbf{x}(j\tau_G))-\delta s_i\frac{\tilde{f}_i}{|\tilde{\mathbf{f}}|}}{\sqrt{2}\delta s_i}\right|^2}\\
&\tilde{F}(\mathbf{s}(\mathbf{x}),t) = F(\mathbf{s}(\mathbf{x})) + w\sum_{j=1}^{n_G^t}\prod_{i=1}^{N_{CV}}\tanh\left(\frac{(n_G^t-j)\tau_G}{\tau_c}\right) e^{-\left|\frac{s_i(\mathbf{x})-s_i(\mathbf{x}(j\tau_G))-\delta s_i\frac{\tilde{f}_i}{|\tilde{\mathbf{f}}|}}{\sqrt{2}\delta s_i}\right|^2}
\end{aligned} \tag{8.94}$$

Efficiency and Accuracy. It can be demonstrated (174) that the error on the calculated $F(\mathbf{s})$ is proportional to the height barrier w; furthermore, with the improvements achieved with eqs. 8.94, a single metadynamics run has shown to already give a very good estimate of the profile, with an error approximately constant in the region where the number of accumulated Gaussians is significant (conventionally this means that $F(\mathbf{s})/w \geq 5$. Obviously the width of the Gaussian also influences the efficiency and the accuracy of the method. In particular, using hills of width larger than typical thermal fluctuations can lead to "bury" some thermodynamic state corresponding to narrow minima. Finally, the accuracy in the evaluation of the thermodynamic force enters in the overall error.

In this thesis, we use Bias Exchange Metadynamics (247). A brief summary of the basic principles of the method is offered in the corresponding chapter).

8.10.2 Weighted Histogram Analysis Method

A single biased simulation is usually not enough to obtain a reliable $F(\mathbf{s})$ over the required range of $\mathbf{s}$. To cope with this problem, a number N_w of simulations can be performed each with different bias potentials, covering adjacent windows in the CVs space. The results from each window then need to be unbiased and glued together into a single PMF. Among the various algorithms proposed, the Weighted Hystogram Analysis Method (WHAM) (169; 267) has proven to be very efficient and almost free of information-loss. The main idea, which goes back to the hystogram method developed by Ferrenberg and Swensden (93), consist in constructing $\rho(\mathbf{s})$ as *weighted* sum of the unbiased distribution functions extracted from each window

$$\rho(\mathbf{s}) = A\sum_{i=1}^{N_w}\pi_i(\mathbf{s})\rho_i(\mathbf{s})\,. \tag{8.95}$$

The weigths π_i are functions of $\mathbf{s}$, and are chosen as to minimize $\sigma^2(\rho(\mathbf{s}))$, subject to normalization $\sum_i \pi_i(\mathbf{s}) = 1$. Thus they are determined using the Lagrange λ multiplier method:

$$\begin{aligned}
&\tfrac{\delta}{\delta\pi_j(\mathbf{s})}\left[\sigma^2(\rho(\mathbf{s})) - \lambda\left(\sum_i\pi_i(\mathbf{s})-1\right)\right] =\\
&\tfrac{\delta}{\delta\pi_j(\mathbf{s})}\left[A^2\sum_i\pi_i^2(\mathbf{s})\sigma^2(\rho_i(\mathbf{s})) - \lambda\left(\sum_i\pi_i(\mathbf{s})-1\right)\right]\\
&= 2A^2\pi_j(\mathbf{s})\sigma^2(\rho_j(\mathbf{s})) - \lambda = 0
\end{aligned} \tag{8.96}$$

which, after writing $\lambda/2A^2 = 1/\sum_i\left[\sigma^2(\rho_i(\mathbf{s}))\right]^{-1}$ (from normalization), gives:

$$\pi_j(\mathbf{s}) = \frac{\left[\sigma^2(\rho_i(\mathbf{s}))\right]^{-1}}{\sum_i\left[\sigma^2(\rho_i(\mathbf{s}))\right]^{-1}} \tag{8.97}$$

Thus, the most accurate ρ_j will have the largest weigth when composing the total distribution function. In a real simulation is clearly useful to express the weigths in terms of the known biasing potentials $U_w i$.

From eq. **??**, one has

$$\sigma^2(\rho_i(\mathbf{s})) = [e^{\beta[U_{wi}(\mathbf{s})-f_{wi}]}]^2\sigma^2(\rho_{wi}(\mathbf{s}))\,, \tag{8.98}$$

which by insertion of eq. **??** gives:

$$\pi_j(\mathbf{s}) = \frac{m_j e^{-\beta[U_{wj}(\mathbf{s})-f_{wj}]}}{\sum\limits_i m_i e^{-\beta[U_{wi}(\mathbf{s})-f_{wi}]}} \tag{8.99}$$

where m_j is the number of sampled points in the j-th window. Thus, the weigths depends on the parameters f_{wj}, that in turn are function of the π_j:

$$\begin{aligned} e^{-\beta f_{wj}} &= \tfrac{Z_{wj}}{Z} = \int d\mathbf{s}\,\rho_j(\mathbf{s}) e^{-\beta U_{wj}(\mathbf{s})} \\ &= A\int d\mathbf{s}\, e^{-\beta U_{wj}(\mathbf{s})} \sum_{k=1}^{N_w} \pi_k(\mathbf{s})\rho_k(\mathbf{s}) \\ &= A\int d\mathbf{s}\, e^{-\beta U_{wj}(\mathbf{s})} \frac{\sum\limits_{k=1}^{N_w} m_k \rho_{wk}(\mathbf{s})}{\sum\limits_i m_i e^{-\beta[U_{wi}(\mathbf{s})-f_{wi}]}} \end{aligned} \tag{8.100}$$

Appendix A

Structural Properties of Polyglutamine Aggregates Investigated via Molecular Dynamics Simulations

Supporting Material

A.0.3 Large monomeric models

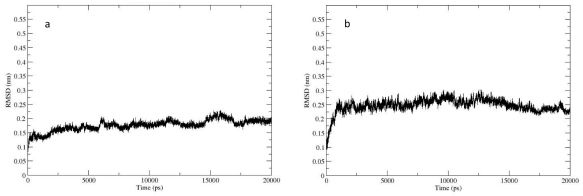


Figure A.1: RMSD - RMSD of P (a) and T (b) plotted as function of time.

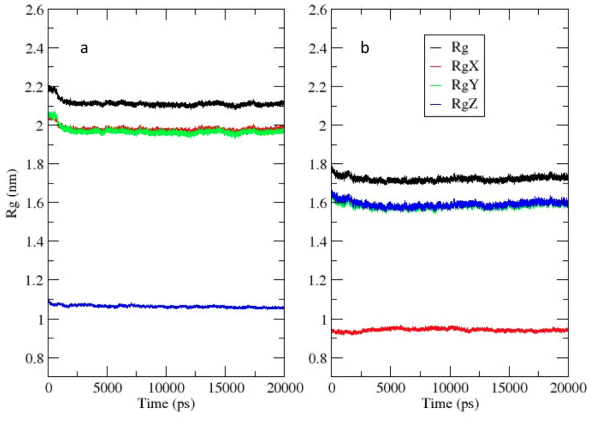


Figure A.2: Rg - Rg of P (a) and T (b) plotted as function of time. XYZ components of Rg for both systems.

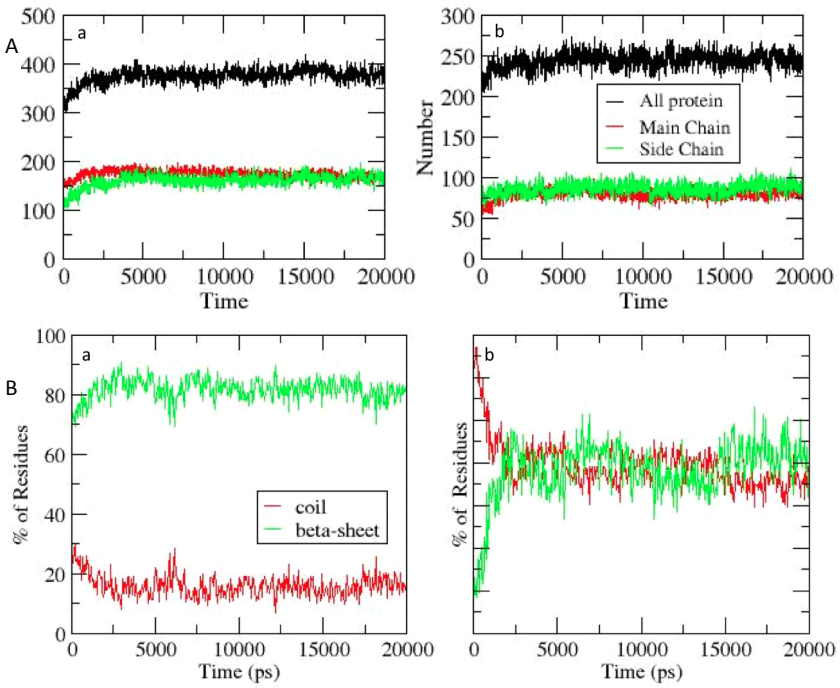


Figure A.3: Properties of P and T - A)Number of HB of P (a) and T (b) plotted as function of time. The overall contribution, as well as the single contributions of the main chains and of the side chains, are also reported. B)βSC and coil conformation of P (a) and T (b) plotted as a function of time.

RMSF

Flexibility. The root-mean-square fluctuations (RMSF) are relatively low, ranging between 0.05 and 0.25 for most residues (Fig. A.4). However, (i) the RMSF values of P are larger every 20 residues. This feature, which becomes even more evident in the final part of the structure, is caused by the fact that the 20_{th} Q of each turn must be rather flexible to allow the β-helix to turn. (ii) The RMSF of T exhibits a minima every 6 residues. In this case, each turn is an equilateral triangle with 6 Qs side: the residues in the vertices i.e. the 6_{th} in the sequence are the most constrained and the less flexible.

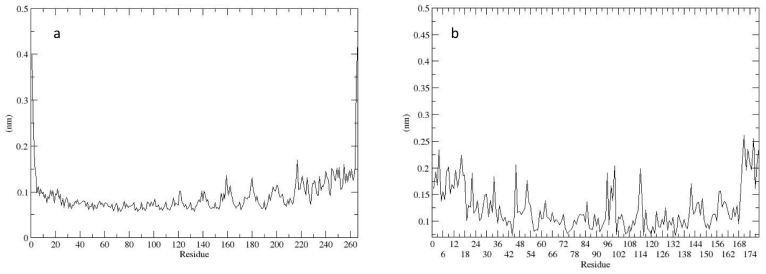


Figure A.4: RMSF - (a) RMSF of circular β-helix; (b) RMSF of triangular -helix.

A.0.4 Oligomeric and small monomeric models

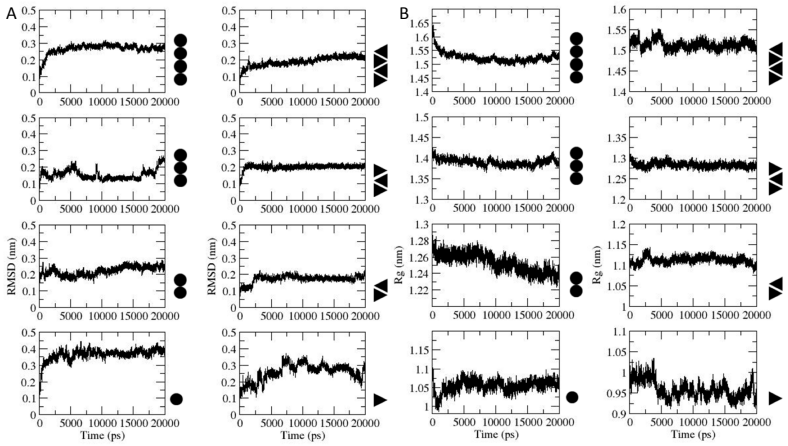


Figure A.5: Properties of oligomeric models - A)RMSD plotted as functions of time in the series of oligomers. B)Rg plotted as functions of time in the series of oligomers

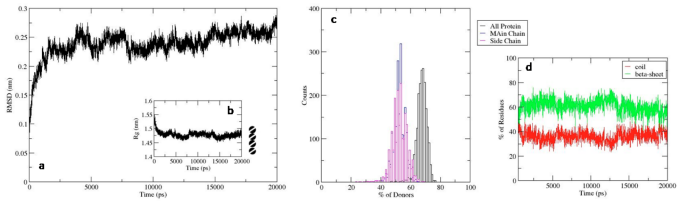


Figure A.6: Properties of P_{AH25} - Oligomeric model built with monomers of 25 Qs (P_{AH25}): a) RMSD vs time; b) Rg vs time; c) hydrogen bond distribution; d) Percentage of residues in random coil (red line) and β-sheet (green line).

A. STRUCTURAL PROPERTIES OF POLYGLUTAMINE AGGREGATES INVESTIGATED VIA MOLECULAR DYNAMICS SIMULATIONS

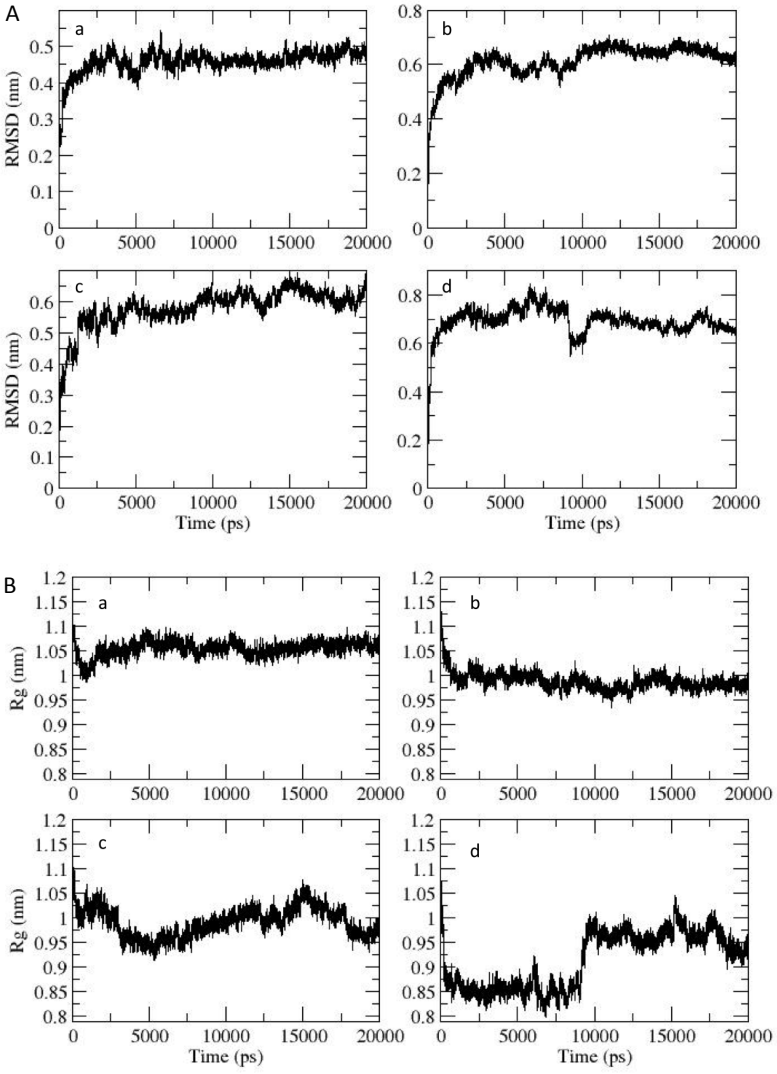


Figure A.7: Properties of Small Oligomeric Models - A) RMSD plotted as functions of time in the series of monomers: (a) monomer built with 40 Qs, (b) monomer built with 35 Qs, (c) monomer built with 30 Qs, (d) monomer built with 25 Qs. B)Rg plotted as functions of time for (a) monomer built with 40 Qs, (b) monomer built with 35 Qs, (c) monomer built with 30 Qs, (d) monomer built with 25 Qs.

A.0.5 Hess's Analysis

(120; 121)

System	cosine content λ_1	cosine content λ_2	cosine content λ_3
P	0.84	0.15	0.05
T	0.31	0.19	0.66
P_{AD}	0.71	0.68	0.42
P_{AC}	0.08	0.01	0.24
P_{AB}	0.89	0.02	0.04
P_A	0.18	0.08	0.03
T_{AD}	0.79	0.35	0.22
T_{AC}	0.90	0.14	0.07
T_{AB}	0.75	0.34	0.01
T_A	0.35	0.15	0.44
P_{AH25}	0.86	0.05	0.02
P_{40}	0.18	0.08	0.03
P_{35}	0.76	0.08	0.02
P_{30}	0.78	0.17	0.27
P_{25}	0.76	0.12	0.28

Table A.1: Cosine content of the first three eigenvalues for the systems studied.

A. STRUCTURAL PROPERTIES OF POLYGLUTAMINE AGGREGATES INVESTIGATED VIA MOLECULAR DYNAMICS SIMULATIONS

Appendix B

Hydrogen bonding cooperativity in polyQ β-sheets from first principle calculations.

Supporting Material

B.0.6 Glutamine Systems - Additional Figures, Schemes and Tables

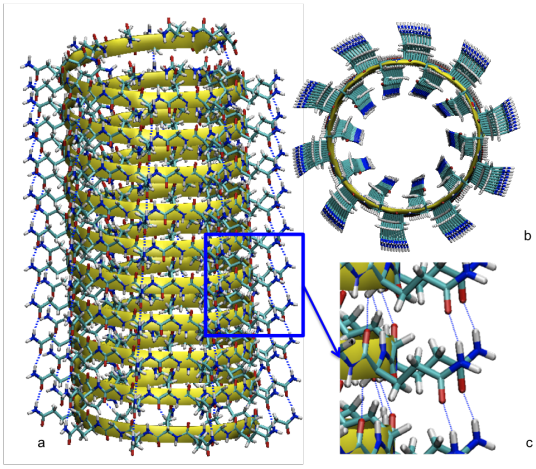


Figure B.1: Circular β-helix - a) Side; b) Front; c) Details of HB network in the structure. The structure is characterized by Q residues with ϕ and ψ angles of -162 and 159 degree. Its coordinates were kindly provided by Dr. A. Lesk.

B. HYDROGEN BONDING COOPERATIVITY IN POLYQ β-SHEETS FROM FIRST PRINCIPLE CALCULATIONS.

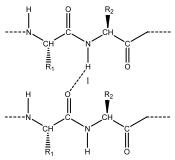
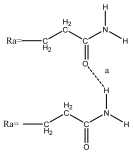
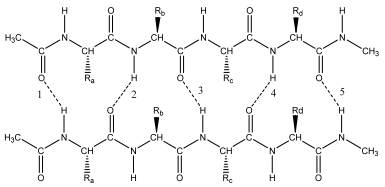
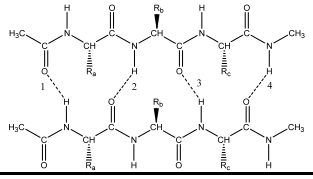
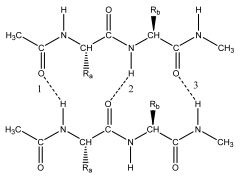
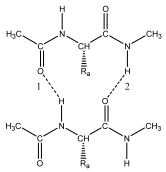
		Back Bone HB net 					Side Chain HB net 				Legend n=1:4 *number of Q in each strand* ***N***=1:4 *number of strands* *The series is define by the index* "n" The values are given in Angstrom
n		1	2	3	4	5	a	b	c	d	Visual Scheme on 2 x n (N=2; n=1:4)
Nx4	4x4	1.91	1.93	1.93	1.93	2.04	1.92	1.94	1.94	1.92	
		1.87	1.95	1.91	1.93	1.93	1.91	1.87	1.90	1.86	
		1.96	1.90	1.95	2.03	1.93	1.95	1.93	1.94	1.91	
	3x4	1.99	1.94	1.92	1.89	1.91	1.91	1.95	1.93	1.95	
		1.93	1.90	1.91	1.89	1.97	1.94	1.96	1.92	1.96	
	2x4	2.04	1.97	2.01	1.85	2.03	1.92	2.00	1.92	1.99	
Nx3	4x3	2.04	1.93	1.94	1.96		1.95	1.93	1.89		
		1.92	1.96	1.91	1.89		1.91	1.89	1.84		
		1.94	1.90	1.88	1.91		1.94	1.92	1.92		
	3x3	1.99	1.88	1.90	1.94		1.93	1.92	1.94		
		1.92	1.89	1.88	1.93		1.93	1.92	1.94		
	2x3	1.99	1.91	1.90	1.95		1.97	1.96	2.02		
Nx2	4x2	1.92	1.86	1.99			1.91	1.92			
		1.89	1.88	1.94			1.84	1.89			
		1.92	1.89	1.92			1.91	1.92			
	3x2	2.06	1.82	2.00			1.90	1.96			
		1.93	1.90	1.93			1.90	1.96			
	2x2	2.01	1.85	1.99			1.96	2.02			
Nx1	4x1	1.93	1.94				1.92				
		1.90	1.92				1.88				
		1.93	1.86				1.92				
	3x1	1.94	1.83				1.92				
		1.94	1.94				1.93				
	2x1	1.91	1.92				1.98				

Figure B.2: HBs lengths - HBs lengths ($\mathring{A}$) in the backbone and side chains for all the systems studied obtained from DFT calculations in vacuo

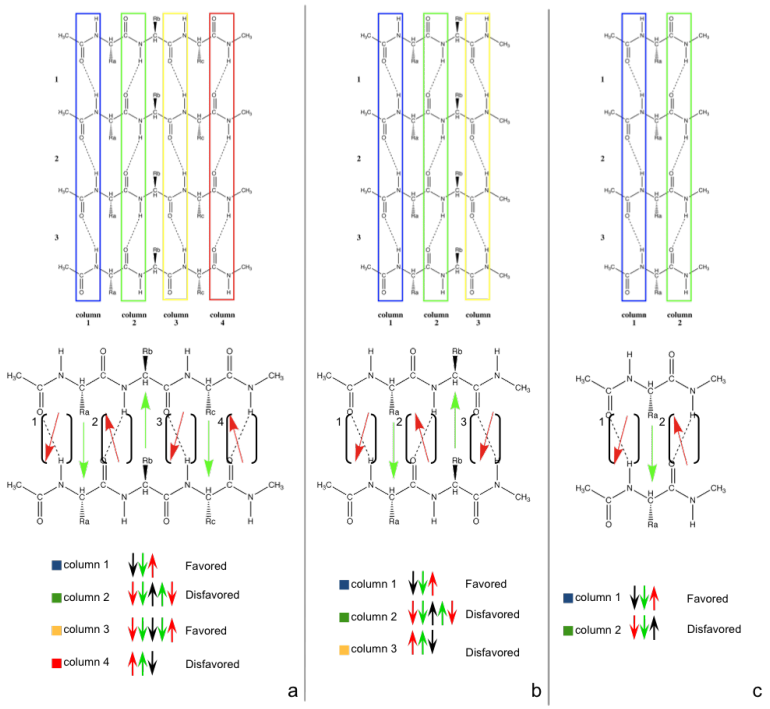


Figure B.3: Glutamine systems - HB dipole orientations in a) 4x3; b) 4x2; c) 4x1.

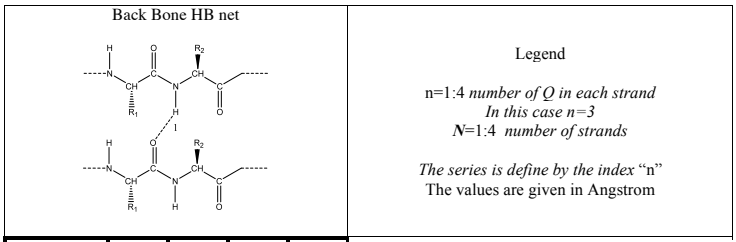
Series	E_{Nxn} (a.u.)	$\Delta E_N(N;n)$ kcal/mol	nHB	ΔE_{HB} kcal/mol
1x1	-132.172084	0	0	-
2x1	-264.368101	15.017957500	3	-5.01
3x1	-396.571458	34.641765000	6	-5.77
4x1	-528.773249	53.282907500	9	-5.92
1x2	-217.200614	0	0	-
2x2	-434.442737	26.046897500	5	-5.21
3x2	-651.696500	59.397895000	10	-5.94
4x2	-868.947322	90.903415000	15	-6.06
1x3	-302.228375	0	0	-
2x3	-604.518061	38.472652500	7	-5.50
3x3	-906.819121	84.082490000	14	-6.01
4x3	-1209.126081	133.394577500	21	-6.35
1x4	-387.256020	0	0	-
2x4	-774.597906	53.880915000	9	-5.99
3x4	-1161.950024	114.182410000	18	-6.34
4x4	-1549.304511	175.970452500	27	-6.52

Table B.1: DFT Energy - Energies obtained with DFT calculations in vacuo.

B.0.7 In vacuo DFT calculations.

The models are built with the modeling tool of HyperChem 8.0(**?**). Starting from one Q residue (1x1) model, (where the first number (N) refers to the number of strands, while the second (n) to the number of Q in each strands), we add progressively one Q in each polyQ chain and one chain in each system. Both N and n vary between 1 and 4.

The models were terminated by the addition of $-NCH_3$ and $-OCCH_3$ groups. Considering all possible Nxn combinations we have 16 models ranging from 29 to 320 atoms. We built also other two different Nx3 series: A) $Nx3_{SC}$ polyQ series where we varied the side chain conformations. B) $Nx3_{ALA}$. These are polyalanine systems.



Systems		1	2	3	4
$Nx3_{SC}$	4x3	2.01	1.93	1.93	1.96
		1.80	1.94	1.89	1.90
		1.85	1.95	1.86	1.96
	3x3	1.97	1.88	1.93	1.85
		1.95	1.91	1.92	1.97
	2x3	2.03	1.85	1.91	1.99
$Nx3_{ALA}$	4x3	1.97	2.12	2.13	2.01
		1.93	2.08	2.11	1.98
		1.97	2.14	2.13	2.03
	3x3	2.08	2.05	2.06	1.95
		1.96	2.00	2.00	1.98
	2x3	2.01	2.06	2.01	1.98

Figure B.4: HBs lengths - HBs lengths ($\mathring{A}$) in the backbone for the systems Nx3SC Nx3ALA studied from DFT calculations in vacuo

DFT calculations were carried out on all the models using the PBE exchange correlation functional,(24; 213; 238) that has been shown to give fairly good results in describing hydrogen bond networks, (138; 139) also with respect WFT method. (330; 332)

We have used a plane wave (PW) basis up to a kinetic energy cutoff of 70 Ry. Core/valence interactions were described using norm conserving pseudopotentials of the Martins-Troullier type.(305) Integration of the nonlocal parts of the pseudopotential was obtained via the Kleinman-Bylander scheme.(159) The models were inserted in an orthorhombic cell of 38, 52, 51 a.u. Isolated system conditions were applied.(18) The calculations were performed with the CPMD v3.11 program.(**?**) Geometries were relaxed by iterating geometry optimization runs (based on a conjugate gradient procedure) up convergence criteria of 10-4 a.u..

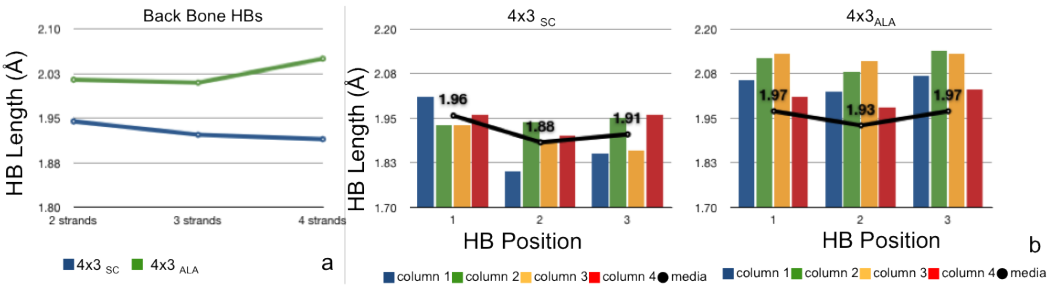


Figure B.5: Backbone CE - a) Backbone CE ($\perp$CE-effect a) in system $4x3_{SC}$ and system $4x3_{ALA}$. Mean values of HB lengths of the backbone atoms versus the number of strands for each series of n Q. b) Backbone CE ($\perp$CE-effect b) in the direction perpendicular to strand elongation: System $4x3_{SC}$, system $4x3_{ALA}$. In the histograms: HB length for each column (the position along the strand) versus the HB position (the position perpendicular to the strand direction); the black line represent the mean values over the rows.

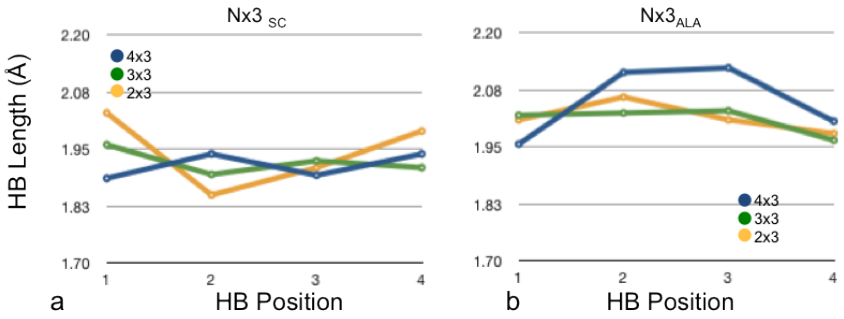


Figure B.6: Backbone $\|$CE - Series $Nx3_{SC}$, $Nx3_{ALA}$. HB length versus HB position.

series		DFT Energy E_{Nxn} (a.u.)	$\Delta E_N(N;n)$ kcal/mol
$Nx3_{SC}$	1x3	-302.228375	0
	2x3	-604.48601758	-18.3654064500061
	3x3	-906.76859604	-52.3780776000396
	4x3	-1209.04199332	-80.6295582999758
$Nx3_{ALA}$	1x3	-286.04906718	0
	2x3	-572.13470186	-22.9461062499578
	3x3	-858.21797862	-44.412617700053
	4x3	-1144.30326989	-67.1432341749221

Visual comparison between ΔE_N between $Nx3_{SC}$(GLN-SC), $Nx3_{ALA}$(ALA) and *Nx3(GLN)*.

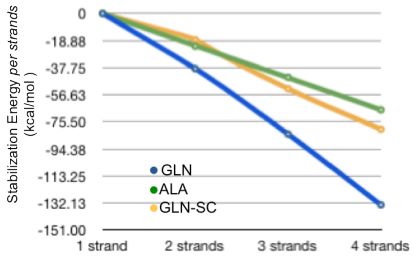


Figure B.7: DFT Energy - Energies obtained with DFT calculations in vacuo for the systems $Nx3_{SC}$, $Nx3_{ALA}$.

B.0.8 DFT/MM calculations.

DFT/MM calculations were performed on a β-helix nanotube(244) whose structure and stability were investigated in our previous work (Fig. B.1).(265) The structure is characterized by 8 turns of 20 Q residues with ϕ and ψ angles of -162 and 159 and its coordinates were kindly provided by Dr. A. Lesk. The β-helix was immersed in a rectangular box large enough to contain the protein and at least 12 $\mathring{A}$ of solvent molecules on each side of the solute.

The system underwent classical MD simulations similarly to our previous work.(265) The parm99 force field(65) and TIP3P water molecules(145) were used for the protein and for water, respectively. Constant temperature-pressure (T=298 K, P =1 bar) 5-ns molecular dynamics (MD) was then performed through the Nose-Hoover(126; 219; 220) and Andersen-Parrinello-Rahman coupling schemes.(6; 233) Periodic boundary conditions were applied. Long-range electrostatic interactions were treated with the particle mesh Ewald (PME) method,(71) using a grid with a spacing of 0.12 nm combined with a fourth-order B-spline interpolation to compute the potential and forces in between grid points. The cutoff radius for the Lenard-Jones interactions as well as for the real part of PME calculations was set to 0.9 nm. The pair list was updated every 2 steps, and the LINCS algorithm(120) was used to constrain all bond lengths involving hydrogen atoms allowing us to use a time step of 2 fs.

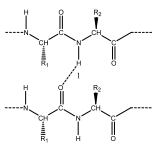
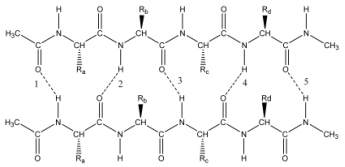
Mean values of HB lenghs (A)		Back Bone HB net 					Legend n=1:4 *number of Q in each strand* *N*=1:4 *number of strands* *The series is define by the index* "n" The values are given in Angstrom
Systems		1	2	3	4	5	visual scheme on 2 x n (N=2; n=1:4)
Nx4	$4x4_{MIX}$	2.21	2.15	2.17	2.03	2.19	
		1.99	1.95	1.98	1.98	1.96	
		1.85	2.22	2.29	2.31	1.99	
	$3x4_{MIX}$	1.89	2.08	1.95	1.86	1.87	
		1.97	1.88	2.16	1.86	2.03	
Nx3	$4x3_{MIX}$	2.17	2.10	2.08	2.11		
		2.00	2.12	2.03	1.92		
		1.99	2.07	1.98	1.94		

Figure B.8: HBs lengths in MIX sustems - HBs lengths ($\mathring{A}$) in the backbone obtained with DFT/MM MD.

The computational protocol was as follows: (i) Energy minimization of the solvent with Harmonic position constraints of 1000 kJmol-1nm-2 on solute atoms. ii) 2 ns of MD always with position constraints on solute allowing the equilibration of the solvent without distorting the solute structure. iii) Energy minimization of entire system; (iv) 1.2 ns MD in which temperature was gradually increased from 0 to 298 K, using simulated annealing(214; 218) (SA). The SA was performed by increasing the temperature from 0 to 298 K in 12 steps in which the temperature was increased by 25 K in 100 ps of MD. v) 5 ns of MD of the entire system. Three DFT/MM models were built based on the last MD snapshot. The QM part included the 4x4, 3x4 and 4x3 moieties. The corresponding mixed DFT/MM models are $4x4_{MIX}$, $3x4_{MIX}$, $4x3_{MIX}$. The QM moieties were cut at the C and C for the N and

C terminal part of the polypeptide chain, respectively and the valence of each carbon was saturated with the addition of capping hydrogen atoms. To perform the calculations we have applied the fully Hamiltonian coupling scheme(175) which couples the CPMD v3.11 code (**?**) with the GROMACS code. (28; 310)

The QM region was treated at DFT level with the same computational setup as in the in vacuo calculations. The only exception is on the dimension of the orthorhombic boxes for the QM part. These had dimensions of a = 50 $\AA$, b = 44 $\AA$, and c = 48 $\AA$; a = 41 $\AA$, b = 42 $\AA$, and c = 48 $\AA$ and a = 49 $\AA$, b = 42 $\AA$, and c = 45 $\AA$ for the $4x4_{MIX}$, $3x4_{MIX}$, $4x3_{MIX}$ models, respectively. As in the in vacuo calculations periodic images have been decupled.(203) The MM was treated exactly as above.

1000 steps of simulated annealing were first performed. Then, the systems were slowly heated to 300 K. A time step of 0.096 fs was used with a fictitious electronic mass of 400 a.u. NVT simulations were carried out by coupling the systems to a Nose-Hoover thermostat.(126; 219) The structures finally underwent 2 ps of DFT/MM MD.

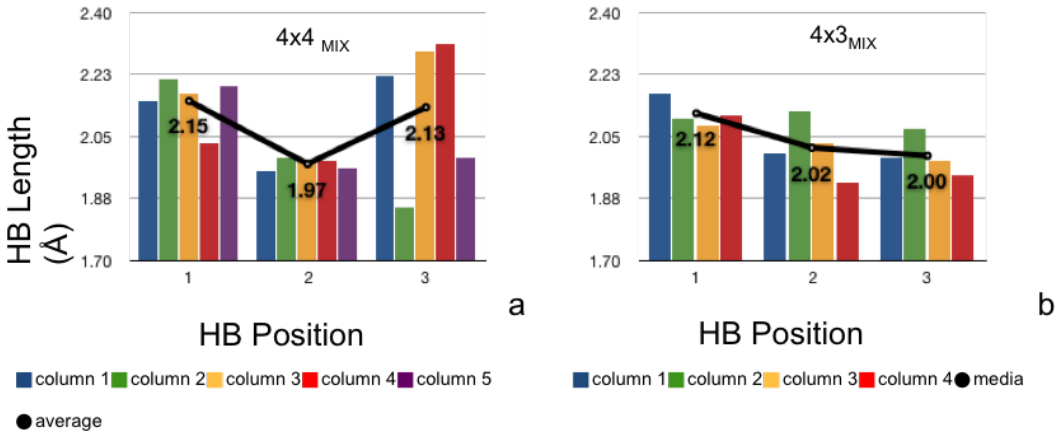


Figure B.9: Backbone CE in the direction perpendicular to strand elongation - a) System $4x4_{MIX}$, b) system $4x3_{MIX}$. In the histograms: HB length for each column (the position along the strand) versus the HB position (the position perpendicular to the strand direction); the black line represent the mean values over the rows.

Appendix C

Conformational ensemble of Huntingtin N-term in aqueous solution explored by atomistic simulations

C.1 Bias Exchange Metadynamics

C.1.1 Principles

In the standard metadynamics (173) approach, the dynamics of the system is biased by a history-dependent potential constructed as sum of Gaussians centered on the trajectory of a selected set of collective variables(CVs). After a transient time, the Gaussian potential compensates the free energy, allowing the system to efficiently explore the space defined by the CVs. This method allows an accurate free energy reconstruction in maximum three variables, as its performance deteriorates with the dimensionality (173), limiting its usefulness for studying protein folding.

Bias exchange (BE) metadynamics(247) consists on running in parallel several molecular dynamics simulations, each biased with a metadynamics potential acting on only one of the relevant CVs that describe the system. At fixed time intervals, swaps of the configurations between pairs of replicas are attempted, and the swap is accepted according to the Metropolis criterion. These swaps greatly enhance the convergence of the free energy estimates (200).

C.1.2 Definition of the collective variables

For N17 BE is performed using six replicas, each biasing one collective variable (CV). The six CVs are explicit functions of the atomic coordinates and have been selected as putative reaction coordinates to explore the different structural conformations of the peptide, following refs (200; 247; 248) :

- CV_1 counts number of $C_\gamma$contacts (hydrophobic contacts) definded as

$$\sum_{ij} \frac{1-(r_{ij}/r_0)^8}{1-(r_{ij}/r_0)^{10}} \tag{C.1}$$

were r_{ij} is the distance between atoms i, j, $r_o = 5\dot{A}$, and the sum runs over all the C_γ atoms.

- CV_2 counts the number of $C_\alpha$contacts, it is defined from Eq. 1, with $r_o = 6.5\dot{A}$ and the sum runs over all the C_α atoms.
- CV_3 counts number of backbone hydrogen bonds, it is defined from Eq. 1, with $r_o = 2\dot{A}$ and the sum runs over all the backbone H and O atoms.
- CV_4 counts total helical content, it is defined by counting the respective fraction of Ψ dihedrals belonging to the α region in the Ramachandran plot as $\sum_i = (1 - cos(\Psi_i - 45))/2$, were the sum runs over all the dihedral angles.
- CV_5 is the helical content of central part of the peptide, it is defined from Eq. 1 and the sum runs over the central dihedrals.
- CV_6 is the dihedral correlation which is a measure of the correlation between successive Ψ dihedrals $\sum_{i=1}^{N-1} = \sqrt{1 + cos^2(\Psi_i - \Psi_{i+1})}$, the sum runs over all the residues.

The results of the simulation are six low dimensional projections of the free energy over each CV. In order to obtain the free energy landscape in several dimensions, it is necessary to analyze the simulation with the technique described below.

C.1.3 Convergence Criteria

In BE the convergence of the bias potential (VG) is monitored like in standard metadynamics (172): after a transient time, VG reaches a stationary state in which it grows evenly fluctuating around an average that estimates the free energy. In our system this happens after 12 ns for each replica. An example of the form of the bias at different simulation times is provided in Fig. C.1. Only the part of the trajectories after 12 ns is retained for further analysis: indeed, after that time, the system stats defusing uniformaly throught CV space, and the bias becomes practically time independent (200).

C.1.4 Cluster Analysis and thermodynamic model

Using the approach of (200), the CV space is divided in a grid of clusters. Namely we group all the frames of the BE trajectories in sets (clusters) in which all the elements are close to each other in CV space.

The free energy of each cluster is estimated by a weighted-hystogram analysis approach (WHAM) (200). In this approach the effect of the bias is removed from the populations, and the estimates of the different walkers are combined in an optimal manner. This allows calculating the free energy of a approximately 2000 clusters of structures, that are representative of all the configurations explored by the system.

The analysis is here performed using variables: CV_2, CV_4 ,CV_5 and CV_6.

The choice of the number of CVs to use is done in order to find the minimum number of variables that allows providing an accurate description of the system. If the variables are too few, a cluster will contain structures that are very different from each other. On the other hand, performing the analysis in a very high dimensional CV space will lead to poor statistics.

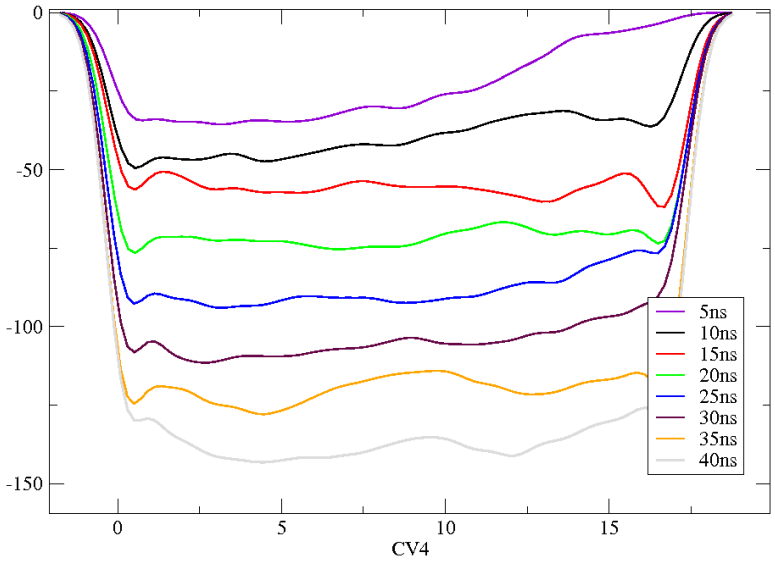


Figure C.1: Bias potential - Bias potential (VG) at different times for CV4. After 12ns, VG starts growing evenly.

The set of clusters was thus defined by partitioning this 4 dimensional CV space in small hyperectangles of sizes 9.2, 1.19, 0.6 and 1.27 respectively for each CV.

In order to show that convergence has been reached, in Fig. C.2 we plot the correlation between the free energies for each cluster estimated at two different filling times t_1=15 and t_2= 30.

C.1.5 Construction of the kinetic model

The transition rates between each cluster are valuated introducing a kinetic model. It has been constructed following ref (200), assuming transitions are possible only between nearest and second nearest neighbours.

The transition rate between two clusters (i,j) is assumed to depend only on the free energy difference ΔGij and on the diffusion matrix Dij in the space of Cvs (131). The diffusion matrix is estimated from a unbiased MD simulation of 10 ns by maximizing the likelihood (131; 200) of the trajectory within the kinetic model. This procedure is valid if on a suitably long time scale the dynamics is Markovian. For our system, the diffusion matrix becomes independent of the time lag after 5ns (200). The diffusion matrix is considered to be diagonal, and is equal to 0.01193, 0.00017, 0.00539, 0.00002.

The metastable states (kinetic clusters) of the network of bins have been found by diagonalizing the rate matrix(222). If this spectrum has a gap at the m^{th} eigenvalue the system is considered to be metastable, and the first times ($t_i \geq t_m$) represent the "slow" decay processes. The relaxation times of the system in descending order are shown in Fig C.3. A gap is found between the first three times and the rest, thus in our analysis we consider the first three as representative of the relevant relaxation processes of the system.

This leads to a partition of the system in four metastable basins.

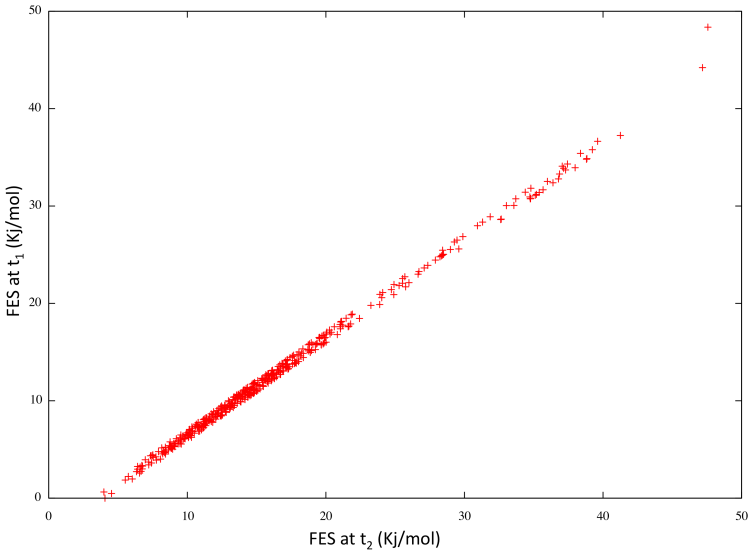


Figure C.2: Convergence - Correlation between free energies estimated using a filling time of 15ns with respect to those estimated at 30ns.

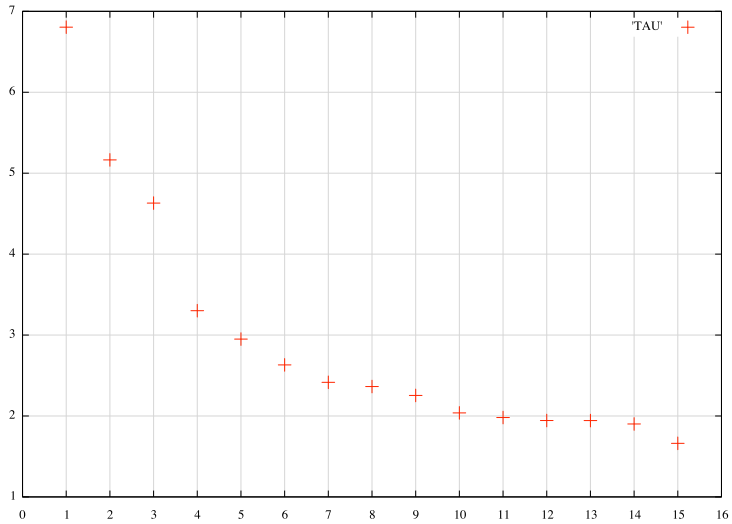


Figure C.3: Relaxation Times - The relaxation times of the system in descending order in ps^{-1}.

C.1.6 Cluster Analysis and thermodynamic model

The long-time scale dynamics of the system has been modeled on the network of clusters by generating a kinetic Monte Carlo (KMC) (39) trajectory of 200 ms.

For two clusters A and B with occupancy P_A and P_B, the rate constant to go from A to B was calculated counting the number of times N_{AB} that a trajectory goes from A to B without passing from any other cluster during the KMC simulation.

The rate to go from A to B was estimated as $k_{AB}\ N_{AB}/(P_A * t_{KMC})$. To minimize the number of recrossing, the KMC trajectory is assumed to visit a different basin any time it reaches a metastable attractor, other clusters that do not fall in this definition were considered as transition states.

C. CONFORMATIONAL ENSEMBLE OF HUNTINGTIN N-TERM IN AQUEOUS SOLUTION EXPLORED BY ATOMISTIC SIMULATIONS

Appendix D

Actin binding by Htt blocks intracellular aggregation.

Supporting Material

PDB ID	Chain ID	Structure Title	Exp. Method	Resolution Å	Macromolecule Name and Source	Expression Host	Citation
					Actin, alpha skeletal muscle,		
		Ternary complex of			Deoxyribonuclease-1		
2A3Z	ABC	the WH2 domain of WASP with Actin-DNAse I	X-RAY	2.08	(Oryctolagus cuniculus), Wiskott-Aldrich syndrome protein (Bos Taurus) Actin, alpha skeletal muscle	-	Chereau, et al. PNAS, 2005
					Deoxyribonuclease-1(Oryctolagus		
	ABC	Ternary complex of the WH2 domain of			cuniculus),	-	
2A40	DEF	WAVE with Actin-DNAse I	X-RAY	1.8	Wiskott-Aldrich syndrome protein family member 2 (Bos Taurus) Actin, alpha skeletal muscle		
					Deoxyribonuclease-1		
		Ternary complex of			(Oryctolagus		
2A41	ABC	the WH2 Domain of WIP with Actin-DNAse I	X-RAY	2.6	cuniculus), Wiskott-Aldrich syndrome protein interacting protein (Bos taurus)	-	
					Actin, alpha skeletal muscle		
		Ternary complex of			Deoxyribonuclease-1		
2D1K	ABC	the WH2 domain of mim with actin-dnase I	X-RAY	2.5	(Oryctolagus cuniculus), Metastasis suppressor protein 1 (Bos taurus)	-	Lee, S.H., et al, Structure 2007
					Actin, alpha skeletal muscle (Oryctolagus		
	ABDE	Cristal Structure of			cuniculus), Neuronal		Gaucher,
*2VCP	ABDE	N-Wasp VC Domain in complex with skeletal actin	X-RAY	3.2	Wiskott-Aldrich syndrome protein (homo sapiens)	E. coli	J.F., et al. To be Published

Table D.1: Structural Determinants - PDB structures used to predict the structural determinants of N17/actinb. After the modeling was carried out, a set of *actin-bound WH2 proteins of Spire X-Ray* structures has been made avaliable (PDBID 3MMV, 3MN5, 3MN6, 3MN7, 3MN9 (85)). Our model for N7/actin complex turns out to be very similar to those structures (Root main square deviations on the backbone ranging from 0.2 to 0.7 $\AA$), strongly suggesting that the results would not change if also these structures are included in the modeling.

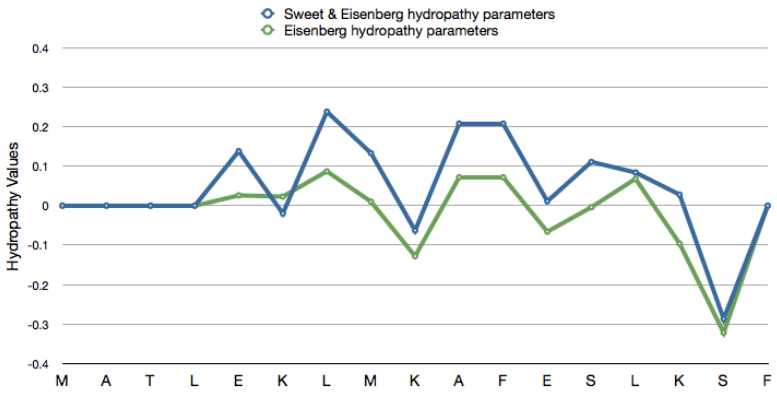


Figure D.1: Hydropathy Plot - Hydrophobicity values for each residues calculated including the contributions of the peptide bonds as well as the sidechains. They are based on experimentally determined values for transfer free energies of polypeptides. The plot was obtained using the EMBOSS serverer at EBI (http://www.ebi.ac.uk/Tools/emboss/pepinfo/).

D. ACTIN BINDING BY HTT BLOCKS INTRACELLULAR AGGREGATION.

D.1 F-ACTIN FILAMENT STRUCTURAL MODEL

Inspection of the F-actin X-Ray structure at 6.6 $\AA$ resolution of rabbit skeletal actin (PDPID: 3MFP (97)- SI of 63% with human F-Actin) shows that the actin hydrophobic pocket (AHP) is solvent accessible not only in G-actin, but also in F-actin(84). We further substantiate this finding by constructing a model of ten subunits of the F-Actin filament (FAF).

- Addition of loops and some extended chains in one human F-actin subunit (i.e. human G-actin), which were unresolved in the crystal structure (X-Ray resolution 12 $\AA$). The Modeller 9v8 (95; 201; 269) program was used.

- Construction of FAF by applying the roto-translations matrix provided in the PDB file of the structure (3BYH (100)). The UCSF Chimera 1.5.2 (246) package was used.

- Clash contacts in the FAF model were eliminated using the Swiss-Pdb Viewer (DeepView 4.0)(110) package.

- The model underwent 10 consecutive run of 1,000 steps of steepest-descent (77) energy minimization. 5000 kcal mol -1$\AA$ -2 harmonic position restraints were applied on the protein backbone atoms.

The resulting model turned out not to been too dissimilar from the F-actin X-Ray structure at 6.6 $\AA$ resolution of rabbit skeletal actin (PDPID: 3MFP (97)- SI of 63% with human F-Actin). The RMSD is 3.2 $\AA$.

Not unexpectedly, the AHP is completely solvent accessible also in our FAF model.

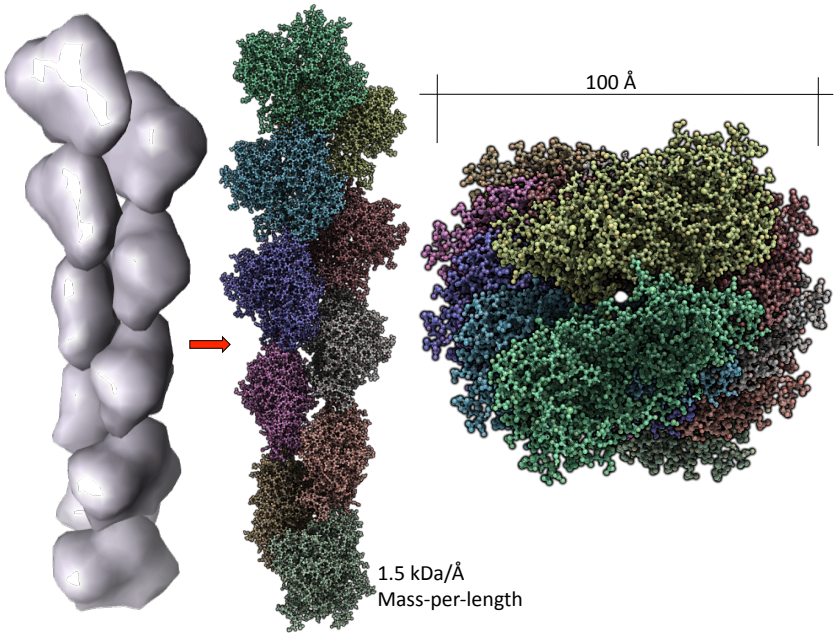


Figure D.2: F-actin model - Side (left) and front (right) view of our FAF model in CPK representation.

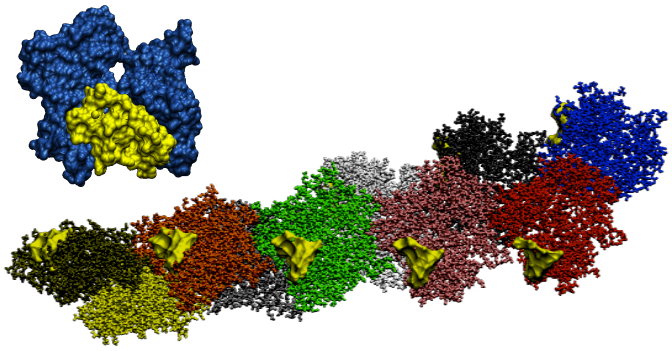


Figure D.3: The hydrophobic binding pocket - The putative N17-binding hydrophobic pocket in our model of FAF (below) and in human G-actin X-ray structure (above)(82). The pocket is shown as a yellow surface representation. The pocket consists of Y143, A144, G146, T148, G168, I341, I345, L346, L349, T351, M355 residues.

D. ACTIN BINDING BY HTT BLOCKS INTRACELLULAR AGGREGATION.

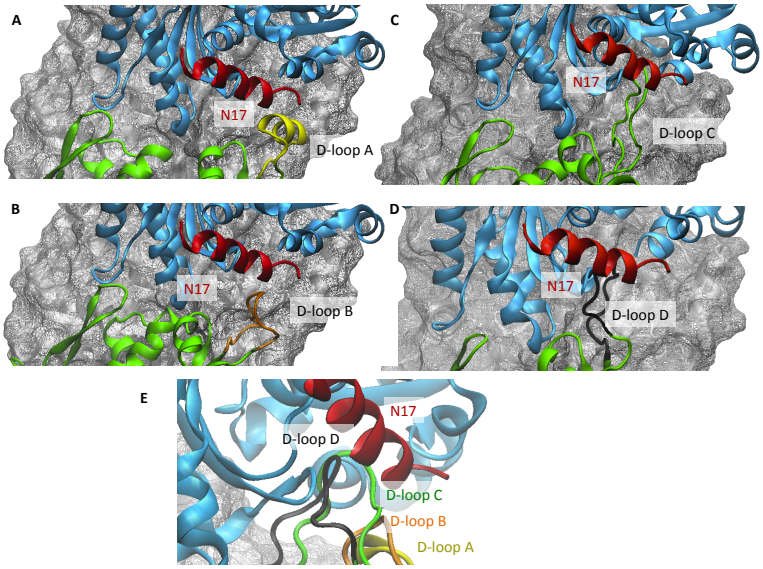


Figure D.4: D-loop - The so-called D-loop (DNase I binding loop) (102) may play a role in F-actin assembly. Hence it has been suggested to affect inderectly for WH2 binding (84). Inspection of the avaliable X-ray structure of uncomlexed actin (228) (A) and EM-based models of F-actin in the free state (97)(C) and in complex with fimbrin (100)(B) and alpha-actinin (D) (101), sheds some lights on this issue. Such inspection points to the flexibility of this loop depending on the oligomerization state, the species and the ligand bound(97; 100; 101; 223). We conclude that at present, it is very difficult to predict the effect of the presence of the D-loop on WH2 binding (E).

D.1.1 ORIENTATION OF N17 HELIX ONTO F-/G- ACTIN HYDROPHOBIC POCKET

Actin binding proteins bind to AHP either front-to-back (N-terminal part of the helix pointed toward the bulk of the actin), and from back-to-front (N-terminal part of the helix pointed opposite to the bulk of the actin monomer) (82).

We carrried out a structural prediction under the assumption that N17 binds to F-actin in a back-to-front orientation. The methods were the same as those for front-to-back. The back-to-front orientation creates less hydrophobic contacts than the front-to-back orientation(Tab. S2). Unlike the latter, it does not form H-bonds with actin. This allows us to suggest that the latter is the most likely binding mode of the ligand. We notice that predictions based on back-to-front are not consistent with several experimental facts (Tab. S3).

N17	Interaction	F-actin
L 4	Hydrophobic contact with	Y 169
K 6	Hydrophobic contact with	L 349
L 7	Hydrophobic contact with	L 349
L 7	Hydrophobic contact with	L 346
L 7	Hydrophobic contact with	Y 143
M 8	Hydrophobic contact with	G 146
A 10	Hydrophobic contact with	L 349
F 11	Hydrophobic contact with	A 144
F 11	Hydrophobic contact with	I 345
F 11	Hydrophobic contact with	I 341
K 15	Hydrophobic contact with	E 334

Table D.2: Alternative binding mode of N17 - Interactions between N17 and human F-Actin obtained by assuming the back-to-front orientation.

Mutation	Model	SI	Helix fold propensity	Aggregation Predicted	FRET Results
L7S	Stabilize helix fold But disrupt hydrophobic interactions with L346, L349 and Y169	Increase	Increase	?	Decrease
F11G	Replace hydrophobic interactions	Increase	Increase	Decrease	Decrease
K15C	Disrupt Salt Bridge with E334	Decrease	Decrease	Increase	Increase
K15A	Disrupt Salt Bridge with E334	Decrease	Decrease	Increase	Increase
F17V	Replace hydrophobic interactions	Increase	Equivalent	Decrease	Decrease

Table D.3: Mutations - Predicted effects of N17 mutations onto aggregation assuming back-to-front orientation.

D.1.2 SEQUENCE ALIGNMENT BETWEEN N17 AND AND ADF /COFILIN FAMILY CLASS

Most actin binding proteins can be grouped into conserved families. These include the Wiskot-tAldrich syndrome protein (WASP)-homology domain-2 (WH2), the actin-depolymerizing factor/cofilin (ADF/cofilin) domain, the gelsolin-homology domain, the calponin-homology (CH) domain, and the myosin motor domain (82). N17 has the highest Sequence Similarity (SS) with WH2 families (SS=63%), followed by ADF/cofilin family (SS=36%). We reported here the aligment for the second the latter one. We further notice that all the other families SS ¡ 30%. Thus, these families are not suitable for an modeling procedure (62).

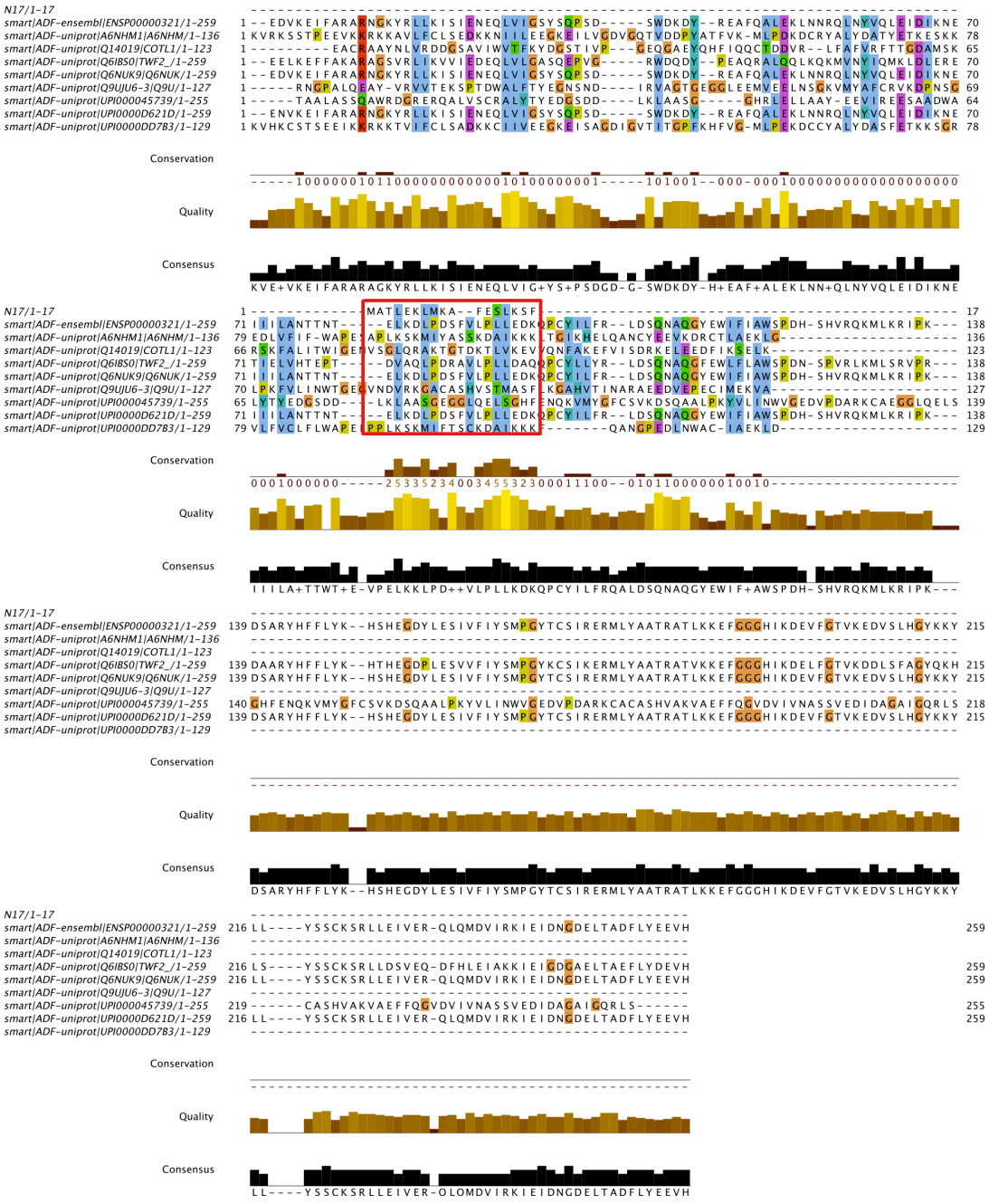


Figure D.5: Sequence Alignment. - Alignment of N17 (red square) on ADF domains in Eukaryotic Homo Sapiens proteins (SS=36%) using the CLUSTALW (300) server.

D. ACTIN BINDING BY HTT BLOCKS INTRACELLULAR AGGREGATION.

Bibliography

[1] J Aaqvist. Ion-water interaction potentials derived from free energy perturbation simulations. *J Phys Chem*, 94(21):8021–8024, 1990. 46

[2] AH Aguda, B Xue, E Irobi, T Preat, and RC Robinson. The structural basis of actin interaction with multiple wh2/beta-thymosin motif-containing proteins. *Structure*, 14(3):469–476, 2006. 54

[3] CT Aiken, JS Steffan, CM Guerrero, H Khashwji, T Lukacsovich, D Simmons, JM Purcell, K Menhaji, Y-Z Zhu, K Green, F Laferla, L Huang, LM Thompson, and JL Marsh. Phosphorylation of threonine 3: implications for huntingtin aggregation and neurotoxicity. *J Biol Chem*, 284(43):29427–36, 2009. 6, 42, 53, 57

[4] SF Altschul, W Gish, W Miller, EW Myers, and DJ Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–10, 1990. 58, 67

[5] SF Altschul, TL Madden, AA Schffer, J Zhang, Z Zhang, W Miller, and DJ Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–402, 1997. 68

[6] HC Andersen. Molecular dynamics simulations at constant pressure and/or temperature. *J Chem Phys*, 72(4):2384–2393, 1980. 89, 123

[7] Y Andersson, DC Langreth, and BI Lundqvist. van der waals interactions in density-functional theory. *Phys Rev Lett*, 76(1):102–105, 1996. 95

[8] MA Andrade and P Bork. Heat repeats in the huntington's disease protein. *Nat Genet*, 11(2):115–6, 1995. 7

[9] MA Andrade, C Petosa, SI O'Donoghue, CW Mller, and P Bork. Comparison of arm and heat protein repeats. *J Mol Biol*, 309(1):1–18, 2001. 7

[10] S Angeli, J Shao, and MI Diamond. F-actin binding regions on the androgen receptor and huntingtin increase aggregation and alter aggregate characteristics. *PLoS ONE*, 5(2):e9053, 2010. 3, 11, 42, 46, 50, 54, 57

[11] J Archie and K Karplus. Applying undertaker cost functions to model quality assessment. *Proteins*, 75(3):550–5, 2009. 73

[12] RS Armen, BM Bernard, R Day, DOV Alonso, and V Daggett. Characterization of a possible amyloidogenic precursor in glutamine-repeat neurodegenerative diseases. *Proc Nat Acad Sci USA*, 102(38):13433–13438, 2005. 9, 15, 16

[13] M Arrasate, S Mitra, ES Schweitzer, MR Segal, and S Finkbeiner. Inclusion body formation reduces levels of mutant huntingtin and the risk of neuronal death. *Nature*, 431(7010):805–10, 2004. 9

[14] NW Ashcroft and ND Mermin. *Solid State Physics*. Harcourt, Orlando, 1976. 97

[15] RS Atwal, J Xia, D Pinchev, J Taylor, RM Epand, and R Truant. Huntingtin has a membrane association signal that can modulate huntingtin aggregation, nuclear entry and toxicity. *Hum Mol Gen*, 16(21):2600–15, 2007. 2, 3, 6, 10, 11, 42, 44, 45, 50, 51, 57, 63

[16] J Bai, JH Hartwig, and N Perrimon. Sals, a wh2-domain-containing protein, promotes sarcomeric actin filament elongation from pointed ends during drosophila muscle growth. *Dev Cell*, 13(6):828–42, 2007. 51, 57

[17] D Bakowies and W Thiel. Hybrid models for combined quantum mechanical and molecular mechanical approaches. *J Phys Chem*, 100(25):10580–10594, 1996. 100

[18] RN Barnett and U Landman. Born-oppenheimer molecular-dynamics simulations of finite systems - structure and dynamics of (h2o)2. *Phys Rev B*, 48(4):2081–2097, 1993. 120

[19] G Bates. Huntingtin aggregation and toxicity in huntington's disease. *Lancet*, 361(9369):1642–4, 2003. 10

[20] GP Bates, L Mangiarini, and SW Davies. Transgenic mice in the study of polyglutamine repeat expansion diseases. *Brain Pathol*, 8(4):699–714, 1998. 6

[21] PO Bauer, HK Wong, F Oyama, A Goswami, M Okuno, Y Kino, H Miyazaki, and N Nukina. Inhibition of rho kinases enhances the degradation of mutant huntingtin. *J Biol Chem*, 284(19):13153–64, 2009. 3, 50

[22] T Beke, I Csizmadia, and A Perczel. Theoretical study on tertiary structural elements of -peptides: Nanotubes formed from parallel-sheet-derived assemblies of -peptides. *J Am Chem Soc*, 128(15):5158–5167, 2006. 32

[23] A Benchoua, Y Trioulier, D Zala, MC Gaillard, N Lefort, N Dufour, F Saudou, JM Elalouf, E Hirsch, P Hantraye, N Deglon, and E Brouillet. Involvement of mitochondrial complex ii defects in neuronal death produced by n-terminus fragment of mutated huntingtin. *Mol Biol Cell*, 17(4):1652–1663, 2006. 14

[24] NA Benedek, IK Snook, K Latham, and I Yarovsky. Application of numerical basis sets to hydrogen bonded systems: a density functional theory study. *J Chem Phys*, 122(14):144102, 2005. 32, 120

[25] MJ Bennett, KE Huey-Tubman, AB Herr, AP West, SA Ross, and PJ Bjorkman. A linear lattice model for polyglutamine in cag-expansion diseases. *Proc Nat Acad Sci USA*, 99(18):11634–11639, 2002. 1, 14, 15

BIBLIOGRAPHY

[26] HJC Berendsen, JR Grigera, and TP Straatsma. The missing term in effective pair potentials. *J Phys Chem*, 91(24):6269–6271, 1987. 30, 88

[27] HJC Berendsen, JPM Postma, WF Vangunsteren, A Dinola, and JR Haak. Molecular-dynamics with coupling to an external bath. *J Chem Phys*, 81(8):3684–3690, 1984. 30, 88

[28] HJC Berendsen, D van der Spoel, and R van Drunen. Gromacs: a message-passing parallel molecular dynamics implementation. *Comput Phys Commun*, 1995. 30, 32, 47, 82, 124

[29] HJC Berensen. *Transport Properties computed by Linear Response through weak coupling to a bath in: "Computer Simulation in Materials Science"*. Kluwer Academic Publishers, 1991. 30

[30] H Berman. The protein data bank: A retrospective and prospective. *Biophys J*, 78(1):267A–267A, 2000. 29

[31] M Berrera, S Pantano, and P Carloni. camp modulation of the cytoplasmic domain in the hcn2 channel investigated by molecular simulations. *Biophys J*, 90(10):3428–3433, 2006. 19

[32] RB Best and G Hummer. Coordinate-dependent diffusion in protein folding. *Proc Nat Acad Sci USA*, 107(3):1088–93, 2010. 43, 47

[33] AE Bevivino and PJ Loll. An expanded glutamine repeat destabilizes native ataxin-3 structure and mediates parallel beta-fibrils. *Proc Nat Acad Sci USA*, 98(21):11955–11960, 2001. 14

[34] I Bezprozvanny and MR Hayden. Deranged neuronal calcium signaling and huntington disease. *Biochem Biophys Res Commun*, 322(4):1310–1317, 2004. 14

[35] A Bhattacharyya, AK Thakur, VM Chellgren, G Thiagarajan, AD Williams, BW Chellgren, TP Creamer, and R Wetzel. Oligoproline effects on polyglutamine conformation and aggregation. *J Mol Biol*, 355(3):524–35, 2006. 3, 7, 11

[36] A Bhattacharyya, AK Thakur, and R Wetzel. Polyglutamine aggregation nucleation: Thermodynamics of a highly unfavorable protein folding reaction. *Proc Nat Acad Sci USA*, 102(43):15400–15405, 2005. 10, 11, 50

[37] N Bizat, JM Hermel, F Boyer, C Jacquard, C Creminon, S Ouary, C Escartin, P Hantraye, S Krajewski, and E Brouillet. Calpain is a major cell death effector in selective striatal degeneration induced in vivo by 3-nitropropionate: Implications for huntington's disease. *J Neurosci*, 23(12):5020–5030, 2003. 14

[38] P Borrell, M s, D Zala, S Humbert, and F Saudou. Huntingtons disease: from huntingtin function and dysfunction to therapeutic strategies. *Cell Mol Life Sci*, 63:2642–2660, 2006. 7, 13, 14, 41

[39] A BORTZ, M KALOS, and J LEBOWITZ. New algorithm for monte-carlo simulation of ising spin systems. *J Comput Phys*, 17(1):10–18, 1975. 129

[40] SF Boys. Electronic wave functions. i. a general method of calculation for the stationary states of any molecular system. *Proc R Soc (London) A Mathematical and Physical Sciences*, 200(1063):542–554, 1950. 96

[41] D Breitsprecher, AK Kiesewetter, J Linkner, C Urbanke, GP Resch, JV Small, and J Faix. Clustering of vasp actively drives processive, wh2 domain-mediated actin filament elongation. *Embo J*, 27(22):2943–2954, 2008. 54

[42] B Brooks, R Bruccoleri, B Olafson, D States, S Swaminathan, and M Karplus. Charmm - a program for macromolecular energy, minimization, and dynamics calculations. *J Comp Chem*, 4(2):187–217, 1983. 71

[43] D Bulone, L Masino, DJ Thomas, PL San Biagio, and A Pastore. The interplay between polyq and protein context delays aggregation by forming a reservoir of protofibrils. *PLoS ONE*, 1(1):e111, 2006. 9

[44] BG Burnett, J Andrews, S Ranganathan, KH Fischbeck, and NA Di Prospero. Expression of expanded polyglutamine targets profilin for degradation and alters actin dynamics. *Neurobiol Dis*, 30(3):365–74, 2008. 3, 50

[45] AA Canutescu, AA Shelenkov, and RL Dunbrack. A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci*, 12(9):2001–14, 2003. 71

[46] K Capelle. A bird's-eye view of density-functional theory. *arxiv:cond-mat/0211443*, 2006. 95

[47] W CARR. Energy, specific heat, and magnetic properties of low-density electron gas. *Phys Rev*, 122(5):1437–, 1961. 95

[48] W CARR and A MARADUDIN. Ground-state energy of high-density electron gas. *Phys Rev A-Gen Phys*, 133(2A):A371–, 1964. 95

[49] EA Carter, G Ciccotti, JT Hynes, and R Kapral. Constrained reaction coordinate dynamics for the simulation of rare events. *Chem Phys Lett*, 156(5):472–477, 1989. 103

[50] DA Case, TE Cheatham, T Darden, H Gohlke, R Luo, KM Merz, A Onufriev, C Simmerling, B Wang, and RJ Woods. The amber biomolecular simulation programs. *J Comput Chem*, 26(16):1668–88, 2005. 28, 46, 71, 72, 103

[51] JP Caviston, JL Ross, SM Antony, M Tokito, and ELF Holzbaur. Huntingtin facilitates dynein/dynactin-mediated vesicle transport. *Proc Natl Acad Sci Usa*, 104(24):10045–50, 2007. 6

[52] DM Ceperley and BJ Alder. Ground state of the electron gas by a stochastic method. *Phys Rev Lett*, 45(7):566–569, 1980. 95

[53] J-HJ Cha. Transcriptional signatures in huntington's disease. *Prog Neurobiol*, 83(4):228–48, 2007. 7

[54] G Chang, CB Roth, CL Reyes, O Pornillos, Y-J Chen, and AP Chen. Retraction. *Science*, 314(5807):1875, 2006. 71

[55] S Chen, V Berthelier, W Yang, and R Wetzel. Polyglutamine aggregation behavior in vitro supports a recruitment mechanism of cytotoxicity. *J Mol Biol*, 311(1):173–182, 2001. 1, 14, 15, 27

[56] SM Chen, V Berthelier, JB Hamilton, B O'Nuallain, and R Wetzel. Amyloid-like features of polyglutamine aggregates and their assembly kinetics. *Biochem*, 41(23):7391–7399, 2002. 14, 15, 27, 31, 49

[57] SM Chen, FA Ferrone, and R Wetzel. Huntington's disease age-of-onset linked to polyglutamine aggregation nucleation. *Proc Nat Acad Sci USA*, 99(18):11884–11889, 2002. 15, 27, 31

[58] D Chereau and R Dominguez. Understanding the role of the g-actin-binding domain of ena/vasp in actin assembly. *J Struct Biol*, 155(2):195–201, 2006. 54

[59] D Chereau, F Kerff, P Graceffa, Z Grabarek, K Langsetmo, and R Dominguez. Actin-bound structures of wiskott-aldrich syndrome protein (wasp)-homology domain 2 and the implications for filament assembly. *Proc Nat Acad Sci USA*, 102(46):16644–9, 2005. 51, 54, 58

[60] G Chinea, G Padron, RW Hooft, C Sander, and G Vriend. The use of position-specific rotamers in model building by homology. *Proteins*, 23(3):415–21, 1995. 70

[61] C Chipot and A Pohorille. *Free energy calculations:Theory and applications in chemistry and biology.* Chemical Physics. Springer Berlin Heidelberg, New York, 2007. 42, 103

[62] C Chothia and AM Lesk. The relation between the divergence of sequence and structure in proteins. *Embo J*, 5(4):823–6, 1986. 58, 66, 68, 69, 138

[63] DW Colby, Y Chu, J Cassady, M Duennwald, H Zazulak, J Webster, A Messer, S Lindquist, V Ingram, and K Wittrup. Potent inhibition of huntingtin and cytotoxicity by a disulfide bond-free single-domain intracellular antibody. *Proc Nat Acad Sci USA*, 101(51):17616–17621, 2004. 3, 11, 42, 46, 50, 51

[64] JK Cooper, G Schilling, MF Peters, WJ Herring, AH Sharp, Z Kaminsky, J Masone, FA Khan, M Delanoy, DR Borchelt, VL Dawson, TM Dawson, and CA Ross. Truncated n-terminal fragments of huntingtin with expanded glutamine repeats form nuclear and cytoplasmic aggregates in cell culture. *Hum Mol Gen*, 7(5):783–790, 1998. 14, 49

[65] WD Cornell, P Cieplak, CI Bayly, IR Gould, KM Merz, DM Ferguson, DC Spellmeyer, T Fox, JW Caldwell, and PA Kollman. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J Am Chem Soc*, 117(19):5179–5197, 1995. 123

[66] P Cossio, F Marinelli, A Laio, and F Pietrucci. Optimizing the performance of bias-exchange metadynamics: Folding a 48-residue lysm domain using a coarse-grained model. *J Phys Chem B*, 114(9):3259–3265, 2010. 47

[67] D Cozzetto, A Kryshtafovych, M Ceriani, and A Tramontano. Assessment of predictions in the model quality assessment category. *Proteins*, 69 Suppl 8:175–83, 2007. 71

[68] D Cozzetto, A Kryshtafovych, and A Tramontano. Evaluation of casp8 model quality predictions. *Proteins*, 77 Suppl 9:157–66, 2009. 71

[69] SL Crick, M Jayaraman, C Frieden, R Wetzel, and RV Pappu. Fluorescence correlation spectroscopy shows that monomeric polyglutamine molecules form collapsed structures in aqueous solutions. *P Natl Acad Sci Usa*, 103(45):16764–9, 2006. 11, 50

[70] D Cvijovic and J Klinowski. Taboo search - an approach to the multiple minima problem. *Science*, 267(5198):664–666, 1995. 104

[71] T Darden, D York, and L Pedersen. Particle mesh ewald: An n-log(n) method for ewald sums in large systems. *J Chem Phys*, 98(12):10089–10092, 1993. 30, 46, 84, 123

[72] G Darnell, JPRO Orgel, R Pahl, and SC Meredith. Flanking polyproline sequences inhibit beta-sheet structure in polyglutamine segments by inducing ppii-like helix structure. *J Mol Biol*, 374(3):688–704, 2007. 11

[73] E Darve and A Pohorille. Calculating free energies using average force. *J Chem Phys*, 115(20):9169–9183, 2001. 103, 104

[74] R Das and D Baker. Macromolecular modeling with rosetta. *Annu Rev Biochem*, 77:363–82, 2008. 73

[75] SW Davies, M Turmaine, BA Cozens, M DiFiglia, AH Sharp, CA Ross, E Scherzinger, EE Wanker, L Mangiarini, and GP Bates. Formation of neuronal intranuclear inclusions underlies the neurological dysfunction in mice transgenic for the hd mutation. *Cell*, 90(3):537–548, 1997. 6, 14, 15, 31, 41, 49, 61

[76] MO Dayhoff and RV Eck. *A Model of Evolutionary Change in Proteins.*, volume 3, pages 33–41. National Biomedical Research Foundation, Washington, D.C., dayhoff, m.o. edition, 1968. 68

[77] P Debye. Nä herungsformeln fü r die zylinderfunktionen fü r groe werte des arguments und unbeschrä nkt verä nderliche werte des index. *Mathematische Annalen*, 67(4):535–558, 1909. 134

[78] B Dehay and A Bertolotti. Critical role of the proline-rich region in huntingtin for aggregation and cytotoxicity in yeast. *J Biol Chem*, 281(47):35608–15, 2006. 10, 11

[79] C Dellago and PG Bolhuis. Transition path sampling and other advanced simulation techniques for rare events. *Adv Polym Sci*, 221:167–233, 2009. 42, 103

[80] UA Desai, J Pallos, AAK Ma, BR Stockwell, LM Thompson, JL Marsh, and MI Diamond. Biologically active molecules that reduce polyglutamine aggregation and toxicity. *Hum Mol Genet*, 15(13):2114–24, 2006. 54, 64

[81] M DiFiglia, E Sapp, KO Chase, SW Davies, GP Bates, JP Vonsattel, and N Aronin. Aggregation of huntingtin in neuronal intranuclear inclusions and dystrophic neurites in brain. *Science*, 277(5334):1990–1993, 1997. 2, 9, 13, 14

BIBLIOGRAPHY

[82] R Dominguez. Actin-binding proteinsa unifying hypothesis. *Trends Biochem Sci*, 29(11):572–578, 2004. 51, 135, 137, 138

[83] R Dominguez. The beta-thymosin/wh2 fold: multifunctionality and structure. *Ann NY Acad Sci*, 1112:86–94, 2007. 54

[84] R Dominguez and CH Kenneth. Actin structure and function. *Annu Rev Biophys*, 40:169–86, 2011. 134, 136

[85] AM Ducka, P Joel, GM Popowicz, KM Trybus, M Schleicher, AA Noegel, R Huber, TA Holak, and T Sitar. Structures of actin-bound wiskott-aldrich syndrome protein homology 2 (wh2) domains of spire and the implication for filament nucleation. *P Natl Acad Sci Usa*, 107(26):11757–62, 2010. 132

[86] ML Duennwald, S Jagadish, PJ Muchowski, and S Lindquist. Flanking sequences profoundly alter polyglutamine toxicity in yeast. *Proc Nat Acad Sci USA*, 103(29):11045–11050, 2006. 10, 42

[87] RL Dunbrack and M Karplus. Backbone-dependent rotamer library for proteins. application to side-chain prediction. *Journal of Molecular Biology*, 230(2):543–74, 1993. 70

[88] RC Edgar. Muscle: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, 5:113, 2004. 69

[89] M Eichinger, P Tavan, J Hutter, and M Parrinello. A hybrid method for solutes in complex solvents: Density functional theory combined with empirical force fields. *J Phys Chem*, 110(21):10452–10467, 1999. 100

[90] B Ensing, A Laio, M Parrinello, and ML Klein. A recipe for the computation of the free energy barrier and the lowest free energy path of concerted reactions. *J Phys Chem B*, 109(14):6676–6687, 2005. 104

[91] L Esposito, A Paladino, C Pedone, and L Vitagliano. Insights into structure, stability, and toxicity of monomeric and aggregated polyglutamine models from molecular dynamics simulations. *Biophys J*, 94(10):4031–4040, 2008. 9, 15, 16

[92] U Essmann, L Perera, ML Berkowitz, T Darden, H Lee, and LG Pedersen. A smooth particle mesh ewald method. *J Chem Phys*, 103(19):8577–8593, 1995. 30, 46, 84

[93] AM Ferrenberg and RH Swendsen. New monte carlo technique for studying phase transitions. *Phys Rev Lett*, 61(23):2635–2638, 1988. 106

[94] J Finke, M Cheung, and J Onuchic. A structural model of polyglutamine determined from a host-guest method combining experiments and landscape theory. *Biophys J*, 87(3):1900–1918, 2004. 9, 15, 16

[95] A Fiser, R Do, and A Sali. Modeling of loops in protein structures. *Protein Sci*, 9(9):1753–1773, 2000. 58, 71, 134

[96] D Frenkel and B Smit. *Understanding Molecular Simulation, Second Edition: From Algorithms to Applications (Computational Science Series, Vol 1)*. Academic Press, 2001. 75

[97] T Fujii, AH Iwane, T Yanagida, and K Namba. Direct visualization of secondary structures of f-actin by electron cryomicroscopy. *Nature*, 467(7316):724–728, 2010. 134, 136

[98] J Gafni and LM Ellerby. Calpain activation in huntington's disease. *J Neurosci*, 22(12):4842–9, 2002. 7

[99] J Gafni, E Hermel, JE Young, CL Wellington, MR Hayden, and LM Ellerby. Inhibition of calpain cleavage of huntingtin reduces toxicity: accumulation of calpain/caspase fragments in the nucleus. *J Biol Chem*, 279(19):20211–20, 2004. 7

[100] VE Galkin, A Orlova, O Cherepanova, M-C Lebart, and EH Egelman. High-resolution cryo-em structure of the f-actin-fimbrin/plastin abd2 complex. *Proc Nat Acad Sci USA*, 105(5):1494–1498, 2008. 58, 134, 136

[101] VE Galkin, A Orlova, A Salmazo, K Djinovic-Carugo, and EH Egelman. Opening of tandem calponin homology domains regulates their affinity for f-actin. *Nat Struct Mol Biol*, 17(5):614–6, 2010. 136

[102] VE Galkin, A Orlova, GF Schrder, and EH Egelman. Structural polymorphism in f-actin. *Nat Struct Mol Biol*, 17(11):1318–1323, 2010. 136

[103] Y Georgalis, EB Starikov, B Hollenbach, R Lurz, E Scherzinger, W Saenger, H Lehrach, and EE Wanker. Huntingtin aggregation monitored by dynamic light scattering. *Proc Nat Acad Sci USA*, 95(11):6118–6121, 1998. 15

[104] S Goedecker, M Teter, and J Hutter. Separable dual-space gaussian pseudopotentials. *Phys Rev B*, 54(3):1703–1710, 1996. 102

[105] H Goehler, M Lalowski, U Stelzl, S Waelter, M Stroedicke, U Worm, A Droege, KS Lindenberg, M Knoblich, C Haenig, M Herbst, J Suopanki, E Scherzinger, C Abraham, B Bauer, R Hasenbank, A Fritzsche, AH Ludewig, K Bssow, K Buessow, SH Coleman, C-A Gutekunst, BG Landwehrmeyer, H Lehrach, and EE Wanker. A protein interaction network links git1, an enhancer of huntingtin aggregation, to huntington's disease. *Mol Cell*, 15(6):853–65, 2004.

[106] RK Graham, Y Deng, EJ Slow, B Haigh, N Bissada, G Lu, J Pearson, J Shehadeh, L Bertram, Z Murphy, SC Warby, CN Doty, S Roy, CL Wellington, BR Leavitt, LA Raymond, DW Nicholson, and MR Hayden. Cleavage at the caspase-6 site is required for neuronal dysfunction and degeneration due to mutant huntingtin. *Cell*, 125(6):1179–91, 2006. 7

[107] H Grubmuller. Predicting slow structural transitions in macromolecular systems: Conformational flooding. *Phys Rev E*, 52(3):2893–2906, 1995. 46

[108] T Grunewald and MF Beal. Bioenergetics in huntington's disease. *Ann NY Acad Sci*, 893(1):203–213, 1999. 7, 14, 41

[109] X Gu, ER Greiner, R Mishra, R Kodali, A Osmand, S Finkbeiner, JS Steffan, LM Thompson, R Wetzel, and XW Yang. Serines 13 and 16 are critical determinants of full-length human mutant huntingtin induced disease pathogenesis in hd mice. *Neuron*, 64(6):828–40, 2009. 6, 53, 54

[110] N Guex and MC Peitsch. Swiss-model and the swiss-pdbviewer: an environment for comparative protein modeling. *Electrophoresis*, 18(15):2714–23, 1997. 134

[111] S Gunawardena, L-S Her, RG Brusch, RA Laymon, IR Niesman, B Gordesky-Gold, L Sintasath, NM Bonini, and LSB Goldstein. Disruption of axonal transport by loss of huntingtin or expression of pathogenic polyq proteins in drosophila. *Neuron*, 40(1):25–40, 2003. 9

[112] JF Gusella and ME MacDonald. Molecular genetics: unmasking polyglutamine triggers in neurodegenerative disease. *Nat Rev Neurosci*, 1(2):109–15, 2000. 1, 5, 61

[113] CA Gutekunst, SH Li, H Yi, JS Mulroy, S Kuemmerle, R Jones, D Rye, RJ Ferrante, SM Hersch, and XJ Li. Nuclear and neuropil aggregates in huntington's disease: Relationship to neuropathology. *J Neurosci*, 19(7):2522–2534, 1999. 9, 14

[114] D Hanein, N Volkmann, S Goldsmith, AM Michon, W Lehman, R Craig, D DeRosier, S Almo, and P Matsudaira. An atomic model of fimbrin binding to f-actin and its implications for filament crosslinking and regulation. *Nat Struct Biol*, 5(9):787–92, 1998. 51

[115] P Harjes and EE Wanker. The hunt for huntingtin function: interaction partners tell many different stories. *Trends Biochem Sci*, 28(8):425–33, 2003. 3, 7, 11, 42, 50

[116] WJ Hehre, R Ditchfield, and JA Pople. Self-consistent molecular orbital methods. xii. further extensions of gaussian–type basis sets for use in molecular orbital studies of organic molecules. *J Chem Phys*, 56(5):2257–2261, 1972. 97

[117] S Henikoff and JG Henikoff. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA*, 89(22):10915–9, 1992. 68

[118] E Hermel, J Gafni, SS Propp, BR Leavitt, CL Wellington, JE Young, AS Hackam, AV Logvinova, AL Peel, SF Chen, V Hook, R Singaraja, S Krajewski, PC Goldsmith, HM Ellerby, MR Hayden, and DE Bredesen. Specific caspase interactions and amplification are involved in selective neuronal vulnerability in huntington's disease. *Cell Death Differ*, 11(4):424–38, 2004. 7

[119] M Hertzog, C van Heijenoort, D Didry, M Gaudier, J Coutant, B Gigant, G Didelot, T Preat, M Knossow, E Guittet, and MF Carlier. The beta-thymosin/wh2 domain: Structural basis for the switch from inhibition to promotion of actin assembly. *Cell*, 117(5):611–623, 2004. 54

[120] B Hess, H Bekker, HJC Berendsen, and JGEM Fraaije. Lincs: A linear constraint solver for molecular simulations. *J Comput Chem*, 18(12):1463–1472, 1997. 30, 46, 87, 115, 123

[121] B Hess, C Kutzner, D van der Spoel, and E Lindahl. Gromacs 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J Chem Theory Comput*, 4(3):435–447, 2008. 82, 115

[122] P Hohenberg and W Kohn. Inhomogeneous electron gas. *Phys Rev B*, 136:864–871, 1964. 92, 93, 94, 95

[123] WG Hol, LM Halie, and C Sander. Dipoles of the alpha-helix and beta-sheet: their role in protein folding. *Nature*, 294(5841):532–6, 1981. 37

[124] B Hollenbach, E Scherzinger, K Schweiger, R Lurz, H Lehrach, and EE Wanker. Aggregation of truncated gst-hd exon 1 fusion proteins containing normal range and expanded glutamine repeats. *Philos Trans R Soc London [Biol]*, 354(1386):991–994, 1999. 15

[125] KC Holmes, I Angert, FJ Kull, W Jahn, and RR Schrder. Electron cryo-microscopy shows how strong binding of myosin to actin releases nucleotide. *Nature*, 425(6956):423–7, 2003. 51

[126] W Hoover. Canonical dynamics: Equilibrium phase-space distributions. *Phys Rev A*, 31(3):1695–1697, 1985. 46, 123, 124

[127] V Horvath, Z Varga, and A Kovacs. Long-range effects in oligopeptides. a theoretical study of the beta-sheet structure of gly(n) (n=2-10). *J Phys Chem A*, 108(33):6869–6873, 2004. 32, 38

[128] V Horvath, Z Varga, and A Kovacs. Substituent effects on long-range interactions in the -sheet structure of oligopeptides. *J Mol Struct (Theochem.)*, 755(1-3):247–251, 2005. 32

[129] D Housman. Gain of glutamines, gain of function? *Nat Genet*, 1995. 7

[130] T Huber, AE Torda, and WF van Gusteren. Local elevation - a method for improving the searching properties of molecular-dynamics simulation. *J Comput Aided Mol Des*, 8(6):695–708, 1994. 104

[131] G Hummer. Position-dependent diffusion coefficients and free energies from bayesian analysis of equilibrium and replica molecular dynamics simulations. *New J Phys*, 7(1):34, 2005. 43, 47, 127

[132] G Huntington. On chorea. *The Medical and Surgical Reporter*, 26(15):317–321, 1872. 1, 5

[133] C Husson, F-X Cantrelle, P Roblin, D Didry, KHD Le, J Perez, E Guittet, C van Heijenoort, L Renault, and M-F Carlier. Multifunctionality of the beta-thymosin/wh2 module: G-actin sequestration, actin filament growth, nucleation, and severing. *Ann NY Acad Sci*, 1194:44–52, 2010. 51, 57

[134] J Hutter, A Alavi, T Deutsch, P Ballone, M Bernasconi, P Focher, S Goedecker, M Tuckerman, and M Parrinello. Cpmd, 1995. 100, 103

[135] M Iannuzzi, A Laio, and M Parrinello. Efficient exploration of reactive potential energy surfaces using car-parrinello molecular dynamics. *Phys Rev Lett*, 90(23):238302, 2003. 104

[136] Z Ignatova, AK Thakur, R Wetzel, and LM Gierasch. In-cell aggregation of a polyglutamine-containing chimera is a multistep process initiated by the flanking sequence. *J Biol Chem*, 282(50):36736–43, 2007. 3, 10, 11, 42, 50

BIBLIOGRAPHY

[137] S Imarisio, J Carmichael, V Korolchuk, C-W Chen, S Saiki, C Rose, G Krishna, Janet E Davies, E Ttofi, Benjamin R Underwood, and David C Rubinsztein. Huntington's disease: from pathology and genetics to potential therapies. *Biochem J*, 412(2):191, 2008. 6

[138] R Improta. Assessing the reliability of density functional methods in the conformational study of polypeptides: The treatment of intraresidue nonbonding interactions. *J Comput Chem*, 25(11):1333–1341, 2004. 120

[139] R Improta, V Barone, KN Kudin, and GE Scuseria. The conformational behavior of polyglycine as predicted by a density functional model with periodic boundary conditions. *J Chem Phys*, 114(6):2541–2549, 2001. 32, 120

[140] C Jarzynsky. Nonequilibrium equality for free energy differences. *Phys Rev Lett*, 78:2690–2693, 1997. 103

[141] F Jensen. *Introduction to Computational Chemistry*. Wiley, 2006. 92, 95, 96, 99

[142] CD Johnson and BL Davidson. Huntington's disease: progress toward effective disease-modifying treatments and a cure. *Hum Mol Genet*, 19(R1):R98–R102, 2010. 7, 41

[143] ER Johnson and AD Becke. A post-hartree–fock model of intermolecular interactions. *J Chem Phys*, 123(2):024101, 2005. 95

[144] DT Jones, WR Taylor, and JM Thornton. A new approach to protein fold recognition. *Nature*, 358(6381):86–9, 1992. 68

[145] W Jorgensen, J Chandrasekhar, J Madura, R Impey, and M Klein. Comparison of simple potential functions for simulating liquid water. *J Chem Phys*, 79(2):926–935, 1983. 123

[146] W Kabsch and C Sander. Dictionary of protein secondary structure - pattern-recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, 1983. 19

[147] M Kalman and N Ben-Tal. Quality assessment of protein model-structures using evolutionary conservation. *Bioinformatics*, 26(10):1299–307, 2010. 71, 73

[148] LS Kaltenbach, E Romero, RR Becklin, R Chettier, R Bell, A Phansalkar, A Strand, C Torcassi, J Savage, A Hurlburt, G-H Cha, L Ukani, CL Chepanoske, Y Zhen, S Sahasrabudhe, J Olson, C Kurschner, LM Ellerby, JM Peltier, J Botas, and RE Hughes. Huntingtin interacting proteins are genetic modifiers of neurodegeneration. *PLoS Genet*, 3(5):e82, 2007. 3, 11, 15, 50

[149] K Kar, M Jayaraman, B Sahoo, R Kodali, and R Wetzel. Critical nucleus size for disease-related polyglutamine aggregation is repeat-length dependent. *Nat Struct Mol Biol*, 2011. 11

[150] R Kayed, E Head, JL Thompson, TM McIntire, SC Milton, CW Cotman, and CG Glabe. Common structure of soluble amyloid oligomers implies common mechanism of pathogenesis. *Science*, 300(5618):486–489, 2003. 1, 14

[151] KB Kegel, AR Meloni, Y Yi, YJ Kim, E Doyle, BG Cuiffo, E Sapp, Y Wang, Z-H Qin, JD Chen, JR Nevins, N Aronin, and M DiFiglia. Huntingtin is present in the nucleus, interacts with the transcriptional corepressor c-terminal binding protein, and represses transcription. *J Biol Chem*, 277(9):7466–76, 2002. 6

[152] KB Kegel, E Sapp, J Yoder, B Cuiffo, L Sobin, YJ Kim, Z-H Qin, MR Hayden, N Aronin, DL Scott, G Isenberg, WH Goldmann, and M DiFiglia. Huntingtin associates with acidic phospholipids at the plasma membrane. *J Biol Chem*, 280(43):36464–73, 2005. 6, 64

[153] SD Khare, F Ding, KN Gwanmesia, and NV Dokholyan. Molecular origin of polyglutamine aggregation in neurodegenerative diseases. *PLoS Comp Biol*, 1(3):230–5, 2005. 9, 15, 16, 17

[154] A Khoshnan, J Ko, and PH Patterson. Effects of intracellular expression of anti-huntingtin antibodies of various specificities on mutant huntingtin aggregation and toxicity. *Proc Natl Acad Sci Usa*, 99(2):1002–7, 2002. 3

[155] MW Kim, Y Chelliah, SW Kim, Z Otwinowski, and I Bezprozvanny. Secondary structure of huntingtin amino-terminal region. *Structure/Folding and Design*, 17(9):1205–1212, 2009. 3, 7, 10, 11, 42, 50, 51

[156] S Kim, K Shilagardi, S Zhang, SN Hong, KL Sens, J Bo, GA Gonzalez, and EH Chen. A critical function for the actin cytoskeleton in targeted exocytosis of prefusion vesicles during myoblast fusion. *Dev Cell*, 12(4):571–86, 2007. 51, 57

[157] YJ Kim, Y Yi, E Sapp, Y Wang, B Cuiffo, KB Kegel, ZH Qin, N Aronin, and M DiFiglia. Caspase 3-cleaved n-terminal fragments of wild-type and mutant huntingtin are present in normal and huntington's disease brains, associate with membranes, and undergo calpain-dependent proteolysis. *P Natl Acad Sci Usa*, 98(22):12784–9, 2001. 7

[158] F Klein, A Pastore, L Masino, G Zederlutz, H Nierengarten, M Ouladabdelghani, D Altschuh, J Mandel, and Y Trottier. Pathogenic and non-pathogenic polyglutamine tracts have similar structural properties: Towards a length-dependent toxicity gradient. *J Mol Biol*, 371(1):235–244, 2007. 2, 10, 15, 24, 28, 31

[159] L Kleinman and DM Bylander. Efficacious form for model pseudopotentials. *Phys Rev Lett*, 48(20):1425–1428, 1982. 98, 120

[160] O Koch, M Bocola, and G Klebe. Cooperative effects in hydrogen-bonding of protein secondary structure elements: A systematic analysis of crystal data using secbase. *Proteins*, 61(2):310–317, 2005. 32, 34, 35

[161] W Kohn and LJ Sham. Self-consistent equations including exchange and correlation effects. *Phys Rev*, 140:A1133–A1138, 1965. 93

[162] T Koopmans. über die zuordnung von wellenfunktionen und eigenwerten zu den einzelnen elektronen eines atoms. *Physica*, 1:104, 1934. 92

[163] T Kortemme, M Ramirez-Alvarado, and L Serrano. Design of a 20-amino acid, three-stranded beta-sheet protein. *Science*, 281(5374):253–256, 1998. 37

[164] E Krieger, T Darden, SB Nabuurs, A Finkelstein, and G Vriend. Making optimal use of empirical energy functions: Force-field parameterization in crystal space. *Proteins*, 57(4):678–683, 2004. 71, 72

[165] E Krieger, G Koraimann, and G Vriend. Increasing the precision of comparative models with yasara nova–a self-parameterizing force field. *Proteins*, 47(3):393–402, 2002. 71, 72

[166] E Krieger, SB Nabuurs, and G Vriend. *Homology Modeling*. Structural Bioinformatics. John Wiley Sons, Inc., 2005. 72

[167] MF Kropman and HJ Bakker. Dynamics of water molecules in aqueous solvation shells. *Science*, 291(5511):2118–2120, 2001. 98

[168] S Kuemmerle, CA Gutekunst, AM Klein, XJ Li, SH Li, MF Beal, SM Hersch, and RJ Ferrante. Huntington aggregates may not predict neuronal death in huntington's disease. *Ann Neurol*, 46(6):842–9, 1999. 9

[169] S Kumar, D Bouzida, R Swendsen, PA Kollman, and J ROSENBERG. The weighted histogram analysis method for free-energy calculations on biomolecules .1. the method. *J Comput Chem*, 13(8):1011–1021, 1992. 47, 106

[170] S Kumar, JM Rosenberg, D Bouzida, RH Swendsen, and PA Kollman. Multidimensional free-energy calculations using the weighted histogram analysis method. *J Comp Chem*, 16(11):1339–1350, 1995. 103

[171] E Lacroix, AR Viguera, and L Serrano. Elucidating the folding problem of alpha-helices: local motifs, long-range electrostatics, ionic-strength dependence and prediction of nmr parameters. *J Mol Biol*, 284(1):173–91, 1998. 51

[172] A Laio and FL Gervasio. Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science. *Rep Prog Phys*, 71(12):126601, 2008. 126

[173] A Laio and M Parrinello. Escaping free-energy minima. *Proc Natl Acad Sci Usa*, 99(20):12562–6, 2002. 103, 104, 105, 125

[174] A Laio, A Rodriguez-Fortea, FL Gervasio, M Ceccarelli, and M Parrinello. Assessing the accuracy of metadynamics. *J Phys Chem B*, 109(14):6714–6721, 2005. 104, 106

[175] A Laio, J VandeVondele, and U Rothlisberger. A hamiltonian electrostatic coupling scheme for hybrid car–parrinello molecular dynamics simulations. *J Chem Phys*, 116(16):6941–6947, 2002. 32, 100, 124

[176] VV Lakhani, F Ding, and NV Dokholyan. Polyglutamine induced misfolding of huntingtin exon1 is modulated by the flanking sequences. *PLoS Comput Biol*, 6(4):e1000772, 2010. 3, 11, 42, 50

[177] RA Laskowski, MW MacArthur, DS Moss, and JM Thornton. Procheck: a program to check the stereochemical quality of protein structures. *J Appl Crystallogr*, 26(2):283–291, 1993. 73

[178] BR Leavitt, JA Guttman, JG Hodgson, GH Kimel, R Singaraja, AW Vogl, and MR Hayden. Wild-type huntingtin reduces the cellular toxicity of mutant huntingtin in vivo. *Am J Hum Genet*, 68(2):313–24, 2001. 7

[179] BR Leavitt, CL Wellington, and MR Hayden. Recent insights into the molecular pathogenesis of huntington disease. *Semin Neurol*, 19(4):385–95, 1999. 7

[180] JM Lecerf, TL Shirley, Q Zhu, A Kazantsev, P Amersdorfer, DE Housman, A Messer, and JS Huston. Human single-chain fv intrabodies counteract in situ huntingtin aggregation in cellular models of huntington's disease. *Proc Natl Acad Sci Usa*, 98(8):4764–9, 2001. 3

[181] SH Lee, F Kerff, D Chereau, F Ferron, A Klug, and R Dominguez. Structural basis for the actin-binding function of missing-in-metastasis. *Structure*, 15(2):145–55, 2007. 58

[182] W-CM Lee, M Yoshihara, and JT Littleton. Cytoplasmic aggregates trap polyglutamine-containing proteins and block axonal transport in a drosophila model of huntington's disease. *Proc Natl Acad Sci Usa*, 101(9):3224–9, 2004. 9

[183] J Legleiter, E Mitchell, GP Lotz, E Sapp, C Ng, M DiFiglia, LM Thompson, and PJ Muchowski. Mutant huntingtin fragments form oligomers in a polyglutamine length-dependent manner in vitro and in vivo. *J Biol Chem*, 285(19):14777–90, 2010. 9

[184] V Leone, F Marinelli, P Carloni, and M Parrinello. Targeting biomolecular flexibility with metadynamics. *Curr Opin Struct Biol*, 20(2):148–154, 2010. 42

[185] H Li, SH Li, H Johnston, PF Shelbourne, and XJ Li. Amino-terminal fragments of mutant huntingtin show selective accumulation in striatal neurons and synaptic toxicity. *Nature Genet*, 25(4):385–389, 2000. 14

[186] H Li, SH Li, ZX Yu, P Shelbourne, and XJ Li. Huntingtin aggregate-associated axonal degeneration is an early pathological event in huntington's disease mice. *J Neurosci*, 21(21):8473–81, 2001. 9

[187] P Li, KE Huey-Tubman, T Gao, X Li, AP West, MJ Bennett, and PJ Bjorkman. The structure of a polyq-anti-polyq complex reveals binding according to a linear lattice model. *Nat Struct Mol Biol*, 14(5):381–7, 2007. 3

[188] W Li, LC Serpell, WJ Carter, DC Rubinsztein, and JA Huntington. Expression and characterization of full-length human huntingtin, an elongated heat repeat protein. *J Biol Chem*, 281(23):15916–22, 2006. 6, 7

[189] XJ Li. The early cellular pathology of huntington's disease. *Mol Neurobiol*, 20(2-3):111–124, 1999. 14

[190] AM Liapunov. *Stability of motion*. Academic Press, New York, 1966. 75

[191] PA Loomis, AE Kelly, L Zheng, B Changyaleket, G Sekerkov, E Mugnaini, A Ferreira, RD Mullins, and JR Bartles. Targeted wild-type and jerker espins reveal a novel, wh2-domain-dependent way to make actin bundles in cells. *J Cell Sci*, 119(Pt 8):1655–65, 2006. 51, 54, 57

BIBLIOGRAPHY

[192] SC Lovell, JM Word, JS Richardson, and DC Richardson. The penultimate rotamer library. *Proteins*, 40(3):389–408, 2000. 70

[193] R Ludwig. Cooperative hydrogen bonding in amides and peptides. *J Mol Liq*, 84(1):65–75, 2000. 31, 62

[194] A Lunkes, KS Lindenberg, L Ben-Haiem, C Weber, D Devys, GB Landwehrmeyer, JL Mandel, and Y Trottier. Proteases acting on mutant huntingtin generate cleaved products that differentially build up cytoplasmic and nuclear inclusions. *Mol Cell*, 10(2):259–269, 2002. 7, 14, 49

[195] M MacDonald, C Ambrose, M Duyao, R Myers, C Lin, L Srinidhi, G Barnes, S Taylor, M James, N Groot, H MacFarlane, B Jenkins, M Anderson, N Wexler, J Gusella, G Bates, S Baxendale, H Hummerich, S Kirby, M North, S Youngman, R Mott, G Zehetner, Z Sedlacek, A Poustka, A-M Frischauf, H Lehrach, A Buckler, D Church, L Doucette-Stamm, M O'Donovan, L Riba-Ramirez, M Shah, V Stanton, S Strobel, K Draths, J Wales, P Dervan, D Housman, M Altherr, R Shiang, L Thompson, T Fielder, J Wasmuth, D Tagle, J Valdes, L Elmer, M Allard, L Castilla, M Swaroop, K Blanchard, F Collins, R Snell, T Holloway, K Gillespie, N Datson, D Shaw, and P Harper. A novel gene containing a trinucleotide repeat that is expanded and unstable on huntington's disease chromosomes. *Cell*, 72(6):971–983, 1993. 1, 5, 6, 41, 49

[196] L Mangiarini, K Sathasivam, M Seller, B Cozens, A Harper, C Hetherington, M Lawton, Y Trottier, H Lehrach, SW Davies, and GP Bates. Exon 1 of the hd gene with an expanded cag repeat is sufficient to cause a progressive neurological phenotype in transgenic mice. *Cell*, 87(3):493–506, 1996. 1, 6, 7, 14, 61

[197] AJ Marchut and CK Hall. Side-chain interactions determine amyloid formation by model polyglutamine peptides in molecular dynamics simulations. *Biophys J*, 90(12):4574–4584, 2006. 9, 15, 16

[198] AJ Marchut and CK Hall. Spontaneous formation of annular structures observed in molecular dynamics simulations of polyglutamine peptides. *Comput Biol Chem*, 30(3):215–218, 2006. 9, 15, 16

[199] AJ Marchut and CK Hall. Effects of chain length on the aggregation of model polyglutamine peptides: Molecular dynamics simulations. *Proteins Struct Funct Bioinf*, 66(1):96–109, 2007. 9, 15, 16

[200] F Marinelli, F Pietrucci, A Laio, and S Piana. A kinetic model of trp-cage folding from multiple biased molecular dynamics simulations. *PLoS Comp Biol*, 5(8):e1000452, 2009. 42, 43, 47, 125, 126, 127

[201] M Marti-Renom, A Stuart, A Fiser, R Sanchez, F Melo, and A Sali. Comparative protein structure modeling of genes and genomes. *Annu Rev Bioph Biom*, 29:291–325, 2000. 57, 58, 69, 134

[202] R Martonak, A Laio, and M Parrinello. Predicting crystal structures: The parrinello-rahman method revisited. *Phys Rev Lett*, 90(7):075503, 2003. 104

[203] GJ Martyna and ME Tuckerman. A reciprocal space based method for treating long range interactions in ab initio and force-field-based calculations in clusters. *J Chem Phys*, 110(6):2810–2821, 1999. 124

[204] D Marx and J Hütter. Modern methods and algorithms of quantum chemistry, 2000. 100

[205] L Masino and A Pastore. A structural approach to trinucleotide expansion diseases. *Brain Res Bull*, 56(3-4):183–189, 2001. 31

[206] L Masino and A Pastore. Glutarnine repeats: structural hypotheses and neurodegeneration. *Biochem Soc Trans*, 30:548–551, 2002. 1, 14, 15

[207] AB Meriin, X Zhang, IM Alexandrov, AB Salnikova, MD Ter-Avanesian, YO Chernoff, and MY Sherman. Endocytosis machinery is involved in aggregation of proteins with expanded polyglutamine domains. *Faseb J*, 21(8):1915–25, 2007. 3, 42, 50, 51, 58

[208] AB Meriin, X Zhang, NB Miliaras, A Kazantsev, YO Chernoff, JM McCaffery, B Wendland, and MY Sherman. Aggregation of expanded polyglutamine domain in yeast leads to defects in endocytosis. *Mol Cell Biol*, 23(21):7554–65, 2003. 3, 50, 51, 58

[209] A Merlino, L Esposito, and L Vitagliano. Polyglutamine repeats and beta-helix structure: Molecular dynamics study. *Proteins Struct Funct Bioinf*, 63(4):918–927, 2006. 9, 15, 16

[210] E Michalsky, A Goede, and R Preissner. Loops in proteins (lip)–a comprehensive loop database for homology modelling. *Protein Eng*, 16(12):979–85, 2003. 70

[211] C Moller and MS Plesset. Note on an approximation treatment for many-electron systems. *Phys Rev*, 46(7):618–622, 1934. 91, 92

[212] JF Morley, HR Brignull, JJ Weyers, and RI Morimoto. The threshold for polyglutamine-expansion protein aggregation and cellular toxicity is dynamic and influenced by aging in caenorhabditis elegans. *Proc Nat Acad Sci USA*, 99(16):10417–10422, 2002. 15

[213] AV Morozov, T Kortemme, K Tsemekhman, and D Baker. Close agreement between the orientation dependence of hydrogen bonds observed in protein structures and quantum mechanical calculations. *Proc Natl Acad Sci USA*, 101(18):6946–51, 2004. 32, 120

[214] JW Moskowitz, KE Schmidt, SR Wilson, and W Cui. The application of simulated annealing to problems of molecular mechanics. *Int J Quantum Chem*, pages 611–617, 1988. 30, 123

[215] R Myers. Huntington's disease genetics. *NeuroRX*, 1(2):255–262, 2004. 5

[216] Y Nagai, T Inui, HA Popiel, N Fujikake, K Hasegawa, Y Urade, Y Goto, H Naiki, and T Toda. A toxic monomeric conformer of the polyglutamine protein. *Nat Struct Mol Biol*, 14(4):332–340, 2007. 9

[217] AF Neuwald and T Hirano. Heat repeats associated with condensins, cohesins, and other complexes involved in chromosome-related functions. *Genome Res*, 10(10):1445–52, 2000. 7

[218] M Nilges, GM Clore, and AM Gronenborn. Determination of 3-dimensional structures of proteins from interproton distance data by dynamical simulated annealing from a random array of atoms - circumventing problems associated with folding. *Febs Lett*, 239(1):129–136, 1988. 30, 123

[219] S Nose. A molecular dynamics method for simulations in the canonical ensemble. *Mol Phys*, 52(2):255 – 268, 1984. 46, 88, 89, 123, 124

[220] S Nose and M Klein. Constant pressure molecular dynamics for molecular systems. *Mol Phys*, 50(5):1055 – 1076, 1983. 46, 88, 89, 123

[221] C Notredame, DG Higgins, and J Heringa. T-coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol*, 302(1):205–17, 2000. 69

[222] F No, I Horenko, C Schtte, and JC Smith. Hierarchical analysis of conformational dynamics in biomolecules: transition networks of metastable states. *J Chem Phys*, 126(15):155102, 2007. 45, 127

[223] T Oda, M Iwasa, T Aihara, Y Mada, and A Narita. The nature of the globular- to fibrous-actin transition. *Nature*, 457(7228):441–5, 2009. 136

[224] H Ogawa, M Nakano, H Watanabe, EB Starikov, SM Rothstein, and S Tanaka. Molecular dynamics simulation study on the structural stabilities of polyglutamine peptides. *Comput Biol Chem*, 32(2):102–110, 2008. 9, 15, 16, 17, 27, 28

[225] MA Olshina, LM Angley, YM Ramdzan, J Tang, MF Bailey, AF Hill, and DM Hatters. Tracking mutant huntingtin aggregation kinetics in cells reveals three major populations that include an invariant oligomer pool. *J Biol Chem*, 285(28):21807–16, 2010. 9

[226] Y Oma, Y Kino, N Sasagawa, and S Ishiura. Intracellular localization of homopolymeric amino acid-containing proteins expressed in mammalian cells. *J Biol Chem*, 279(20):21217–22, 2004. 9

[227] MBR Oppenheimer. Zur quantentheorie der molekeln (on the quantum theory of molecules). *Annalen der Physik*, 84:457–484, 1927. 98

[228] L Otterbein, P Graceffa, and R Dominguez. The crystal structure of uncomplexed actin in the adp state. *Science*, 293(5530):708–711, 2001. 136

[229] GA Palidwor, S Shcherbinin, MR Huska, T Rasko, U Stelzl, A Arumughan, R Foulle, P Porras, L Sanchez-Pulido, EE Wanker, and MA Andrade-Navarro. Detection of alpha-rod protein repeats using a neural network and application to huntingtin. *PLoS Comp Biol*, 5(3):e1000304, 2009. 7

[230] AV Panov, CA Gutekunst, BR Leavitt, MR Hayden, JR Burke, WJ Strittmatter, and JT Greenamyre. Early mitochondrial calcium defects in huntington's disease are a direct effect of polyglutamines. *Nature Neurosci*, 5(8):731–736, 2002. 14

[231] JA Parker, JB Connolly, C Wellington, M Hayden, J Dausset, and C Neri. Expanded polyglutamines in caenorhabditis elegans cause axonal abnormalities and severe dysfunction of plm mechanosensory neurons without cell death. *Proc Natl Acad Sci Usa*, 98(23):13318–23, 2001. 9

[232] RG Parr and W Yang. *Density-Functional Theory of Atoms and Molecules*. Oxford University Press US, 1989. 91, 92, 94, 95

[233] M Parrinello and A Rahman. Crystal-structure and pair potentials - a molecular-dynamics study. *Phys Rev Lett*, 45(14):1196–1199, 1980. 123

[234] M Parrinello and A Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *J Appl Phys*, 52(12):7182–7190, 1981. 46, 82, 89

[235] E Paunola, P Mattila, and P Lappalainen. Wh2 domain: a small, versatile adapter for actin monomers. *FEBS Letters*, 513(1):92–97, 2002. 50, 51, 54, 58

[236] WR Pearson. Rapid and sensitive sequence comparison with fastp and fasta. *Meth Enzymol*, 183:63–98, 1990. 67

[237] A Perczel, Z Gaspari, and IG Csizmadia. Structure and stability of beta-pleated sheets. *J Comput Chem*, 26(11):1155–1168, 2005. 32

[238] JP Perdew, K Burke, and M Ernzerhof. Generalized gradient approximation made simple. *Phys Rev Lett*, 77(18):3865–3868, 1996. 32, 120

[239] JP Perdew and Y Wang. Accurate and simple analytic representation of the electron-gas correlation energy. *Phys Rev B*, 45(23):13244–13249, 1992. 95

[240] J Perry and N Kleckner. The atrs, atms, and tors are giant heat repeat proteins. *Cell*, 112(2):151–5, 2003. 7

[241] M Perutz. Polar zippers - their role in human-disease. *Protein Sci*, 3(10):1629–1637, 1994. 2, 7, 9, 10, 15, 31, 61, 62

[242] M Perutz, B Pope, D Owen, EE Wanker, and E Scherzinger. Aggregation of proteins with expanded glutamine and alanine repeats of the glutamine-rich and asparagine-rich domains of sup35 and of the amyloid beta-peptide of amyloid plaques. *Proc Nat Acad Sci USA*, 99(8):5596–5600, 2002. 10, 16

[243] MF Perutz. Glutamine repeats and neurodegenerative diseases: molecular aspects. *Trends Biochem Sci*, 24(2):58–63, 1999. 6, 15

[244] MF Perutz, JT Finch, J Berriman, and A Lesk. Amyloid fibers are water-filled nanotubes. *Proc Nat Acad Sci USA*, 99(8):5591–5595, 2002. 9, 15, 19, 29, 32, 123

[245] MF Perutz and AH Windle. Cause of neural death in neurodegenerative diseases attributable to expansion of glutamine repeats. *Nature*, 412(6843):143–144, 2001. 2, 31

[246] E Pettersen, T Goddard, C Huang, G Couch, D Greenblatt, E Meng, and T Ferrin. Ucsf chimeraa visualization system for exploratory research and analysis. *J Comput Chem*, 25(13):1605–1612, 2004. 134

BIBLIOGRAPHY

[247] S Piana and A Laio. A bias-exchange approach to protein folding. *J Phys Chem B*, 111(17):4553–9, 2007. 42, 47, 106, 125

[248] S Piana, A Laio, F Marinelli, M Van Troys, D Bourry, C Ampe, and JC Martins. Predicting the effect of a point mutation on a protein fold: the villin and advillin headpieces and their pro62ala mutants. *J Mol Biol*, 375(2):460–70, 2008. 42, 125

[249] S Piana, D Sebastiani, P Carloni, and M Parrinello. Ab initio molecular dynamics-based assignment of the protonation state of pepstatin a/hiv-1 protease cleavage site. *J Am Chem Soc*, 123(36):8730–7, 2001. 39

[250] SK Pollitt, J Pallos, J Shao, UA Desai, AAK Ma, LM Thompson, JL Marsh, and MI Diamond. A rapid cellular fret assay of polyglutamine aggregation identifies a novel inhibitor. *Neuron*, 40(4):685–94, 2003. 3, 50, 54, 58

[251] CRH Raetz and SL Roderick. A left-handed parallel beta helix in the structure of udp-n-acetylglucosamine acyltransferase. *Science*, 270(5238):997–1000, 1995. 10, 16, 17, 29

[252] YM Ramdzan, RM Nisbet, J Miller, S Finkbeiner, AF Hill, and DM Hatters. Conformation sensors that distinguish monomeric proteins from oligomers in live cells. *Chem Biol*, 17(4):371–9, 2010. 9

[253] G Rebowski, M Boczkowska, DB Hayes, L Guo, TC Irving, and R Dominguez. X-ray scattering study of actin polymerization nuclei assembled by tandem w domains. *Proc Nat Acad Sci USA*, 105(31):10785–90, 2008. 51

[254] G Rebowski, S Namgoong, M Boczkowska, PC Leavis, J Navaza, and R Dominguez. Structure of a longitudinal actin dimer assembled by tandem w domains: implications for actin filament nucleation. *J Mol Biol*, 403(1):11–23, 2010. 57

[255] A Reiner, I Dragatsis, S Zeitlin, and D Goldowitz. Wild-type huntingtin plays a role in brain development and neuronal survival. *Mol Neurobiol*, 28(3):259–76, 2003. 6

[256] J Riedl, AH Crevenna, K Kessenbrock, JH Yu, D Neukirchen, M Bista, F Bradke, D Jenne, TA Holak, Z Werb, M Sixt, and R Wedlich-Soldner. Lifeact: a versatile marker to visualize f-actin. *Nat Meth*, 5(7):605–607, 2008. 54

[257] E Rockabrand, N Slepko, A Pantalone, VN Nukala, AG Kazantsev, JL Marsh, PG Sullivan, JS Steffan, SL Sensi, and LM Thompson. The first 17 amino acids of huntingtin modulate its sub-cellular localization, aggregation and effects on calcium homeostasis. *Hum Mol Gen*, 16(1):61–77, 2007. 3, 6, 11, 42, 50, 57

[258] A Rosenblatt, K-Y Liang, H Zhou, MH Abbott, LM Gourley, RL Margolis, J Brandt, and CA Ross. The association of cag repeat length with clinical progression in huntington disease. *Neurology*, 66(7):1016–20, 2006. 1

[259] C Ross. When more is less: Pathogenesis of glutamine repeat neurodegenerative diseases. *Neuron*, 15(3):493–496, 1995. 1, 5, 61

[260] C Ross and M Poirier. What is the role of protein aggregation in neurodegeneration? *Nat Rev Mol Cell Biol*, 6(11):891–898, 2005. 9

[261] CA Ross. Huntington's disease: new paths to pathogenesis. *Cell*, 118(1):4–7, 2004. 10

[262] CA Ross and MA Poirier. Protein aggregation and neurodegenerative disease. *Nature Med*, pages S10–S17, 2004. 1, 9, 14

[263] CA Ross, MA Poirier, EE Wanker, and M Amzel. Polyglutamine fibrillogenesis: The pathway unfolds. *Proc Nat Acad Sci USA*, 100(1):1–3, 2003. 46

[264] CA Ross and SJ Tabrizi. Huntington's disease: from molecular pathogenesis to clinical treatment. *Lancet Neurol*, 10(1):83–98, 2011. 1, 8, 9, 49

[265] G Rossetti, A Magistrato, A Pastore, F Persichetti, and P Carloni. Structural properties of polyglutamine aggregates investigated via molecular dynamics simulations. *J Phys Chem B*, 112(51):16843–50, 2008. 32, 123

[266] B Rost. Twilight zone of protein sequence alignments. *Protein Eng*, 12(2):85–94, 1999. iv, 66, 68

[267] B Roux. The calculation of the potential of mean force using computer-simulations. *Comput Phys Comm*, 91(1-3):275–282, 1995. 103, 106

[268] DC Rubinsztein, J Leggo, R Coles, E Almqvist, V Biancalana, JJ Cassiman, K Chotai, M Connarty, D Crauford, A Curtis, D Curtis, MJ Davidson, AM Differ, C Dode, A Dodge, M Frontali, NG Ranen, OC Stine, M Sherr, MH Abbott, ML Franz, CA Graham, PS Harper, JC Hedreen, MR Hayden, and et al. Phenotypic characterization of individuals with 30-40 cag repeats in the huntington disease (hd) gene reveals hd cases with 36 repeats and apparently normal elderly individuals with 36-39 repeats. *Am J Hum Genet*, 59(1):16–22, 1996. 15

[269] A Sali and T Blundell. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol*, 234(3):779–815, 1993. 58, 70, 71, 134

[270] C Sander and R Schneider. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, 9(1):56–68, 1991. 66

[271] E Sapp, J Penney, A Young, N Aronin, JP Vonsattel, and M DiFiglia. Axonal transport of n-terminal huntingtin suggests early pathology of corticostriatal projections in huntington disease. *J Neuropathol Exp Neurol*, 58(2):165–173, 1999. 7, 13, 14, 41

[272] F Saudou, S Finkbeiner, D Devys, and ME Greenberg. Huntingtin acts in the nucleus to induce apoptosis but death does not correlate with the formation of intranuclear inclusions. *Cell*, 95(1):55–66, 1998. 9

[273] HM Saunders and SP Bottomley. Multi-domain misfolding: understanding the aggregation pathway of polyglutamine proteins. *Protein Eng Des Sel*, 22(8):447–51, 2009. 42

[274] S Scheiner and T Kar. Effect of solvent upon ch...o hydrogen bonds with implications for protein folding. *J Phys Chem B*, 109(8):3681–3689, 2005. 32

[275] E Scherzinger, R Lurz, M Turmaine, L Mangiarini, B Hollenbach, R Hasenbank, GP Bates, SW Davies, H Lehrach, and EE Wanker. Huntingtin-encoded polyglutamine expansions form amyloid-like protein aggregates in vitro and in vivo. *Cell*, 90(3):549–558, 1997. 14, 15, 49

[276] E Scherzinger, A Sittler, K Schweiger, V Heiser, R Lurz, R Hasenbank, GP Bates, H Lehrach, and EE Wanker. Self-assembly of polyglutamine-containing huntingtin fragments into amyloid-like fibrils: Implications for huntington's disease pathology. *Proc Nat Acad Sci USA*, 96(8):4604–4609, 1999. 14, 15, 31, 49

[277] A Sechi and J Wehland. Ena/vasp proteins: Multifunctional regulators of actin cytoskeleton dynamics. *Front Biosci*, 9:1294–1310, 2004. 54

[278] J Shao, WJ Welch, NA DiProspero, and MI Diamond. Phosphorylation of profilin by rock1 regulates polyglutamine aggregation. *Mol Cell Biol*, 28(17):5196–5208, 2008. 3, 11, 50, 54

[279] D Sharma, LM Shinchuk, H Inouye, R Wetzel, and DA Kirschner. Polyglutamine homopolymers having 8-45 residues form slablike beta-crystallite assemblies. *Proteins Struct Funct Bioinf*, 61(2):398–411, 2005. 9, 15

[280] P Sikorski and E Atkins. New model for crystalline polyglutamine assemblies and their connection with amyloid fibrils. *Biomacromolecules*, 6(1):425–432, 2005. 9, 10, 15, 16, 32

[281] F Sim, A St. Amant, I Papai, and DR Salahub. Gaussian density functional calculations on hydrogen-bonded systems. *J Am Chem Soc*, 114(11):4391–4400, 1992. 95

[282] M Sprik and G Ciccotti. Free energy from constrained molecular dynamics. *J Chem Phys*, 109(18):7737–7744, 1998. 105

[283] N Srinivasan and TL Blundell. An evaluation of the performance of an automated procedure for comparative modelling of protein tertiary structure. *Protein Eng*, 6(5):501–12, 1993. 58

[284] JS Steffan, N Agrawal, J Pallos, E Rockabrand, LC Trotman, N Slepko, K Illes, T Lukacsovich, Y-Z Zhu, E Cattaneo, PP Pandolfi, LM Thompson, and JL Marsh. Sumo modification of huntingtin and huntington's disease pathology. *Science*, 304(5667):100–4, 2004. 3, 11, 42, 46, 50, 57

[285] OC Stine, N Pleasant, ML Franz, and MH Abbott. Correlation between the onset age of huntington's disease and length of the trinucleotide repeat in . *Hum Mol Gen*, 1993. 14

[286] M Stork, A Giese, HA Kretzschmar, and P Tavan. Molecular dynamics simulations indicate a possible role of parallel beta- helices in seeded aggregation of polygln. *Biophys J*, 88(4):2442–2451, 2005. 9, 10, 15, 16, 17

[287] ANT Strehlow, JZ Li, and RM Myers. Wild-type huntingtin participates in protein trafficking between the golgi and the extracellular space. *Hum Mol Genet*, 16(4):391–409, 2007. 6

[288] K Sugaya, S Matsubara, Y Kagamihara, A Kawata, and H Hayashi. Polyglutamine expansion mutation yields a pathological epitope linked to nucleation of protein aggregate: Determinant of huntington's disease onset. *PLoS ONE*, 2(7):e635, 2007. 15

[289] ST Suhr, MC Senut, JP Whitelegge, KF Faull, DB Cuizon, and FH Gage. Identities of sequestered proteins in aggregates from cells with induced polyglutamine expression. *J Cell Biol*, 153(2):283–94, 2001. 3, 50

[290] M Sunde and C Blake. The structure of amyloid fibrils by electron microscopy and x-ray diffraction. *Adv Protein Chem*, 50:123–159, 1997. 9, 15

[291] M Sunde, LC Serpell, M Bartlam, PE Fraser, MB Pepys, and CCF Blake. Common core structure of amyloid fibrils by synchrotron x-ray diffraction. *J Mol Biol*, 273(3):729–739, 1997. 9, 15

[292] MJ Sutcliffe, I Haneef, D Carney, and TL Blundell. Knowledge based modelling of homologous proteins, part i: Three-dimensional frameworks derived from the simultaneous superposition of multiple structures. *Protein Eng*, 1(5):377–84, 1987. 58

[293] R Snchez and A Sali. Evaluation of comparative protein structure modeling by modeller-3. *Proteins*, Suppl 1:50–8, 1997. 46, 58

[294] S Tam, C Spiess, W Auyeung, L Joachimiak, B Chen, MA Poirier, and J Frydman. The chaperonin tric blocks a huntingtin sequence element that promotes the conformational switch to aggregation. *Nat Struct Mol Biol*, 16(12):1279–1285, 2009. 3, 11, 42, 46, 50, 51

[295] M Tanaka, I Morishima, T Akagi, T Hashikawa, and N Nukina. Intra- and intermolecular beta-pleated sheet formation in glutamine-repeat inserted myoglobin as a model for polyglutamine diseases. *J Biol Chem*, 276(48):45470–45475, 2001. 14, 15, 49

[296] M Tartari, C Gissi, V Lo Sardo, C Zuccato, E Picardi, G Pesole, and E Cattaneo. Phylogenetic comparison of huntingtin homologues reveals the appearance of a primitive polyq in sea urchin. *Mol Biol Evol*, 25(2):330–8, 2008. 6, 7, 11

[297] PA Temussi, L Masino, and A Pastore. From alzheimer to huntington: why is a structural understanding so difficult? *Embo J*, 22(3):355–361, 2003. 1, 14, 27

[298] AK Thakur, M Jayaraman, R Mishra, M Thakur, VM Chellgren, I-JL Byeon, DH Anjum, R Kodali, TP Creamer, JF Conway, AM Gronenborn, and R Wetzel. Polyglutamine disruption of the huntingtin exon 1 n terminus triggers a complex aggregation mechanism. *Nat Struct Mol Biol*, 16(4):380–389, 2009. 3, 6, 11, 42, 44, 45, 46, 50, 51, 57, 63

[299] AK Thakur and R Wetzel. Mutational analysis of the structural organization of polyglutamine aggregates. *Proc Natl Acad Sci Usa*, 99(26):17014–9, 2002. 3, 11, 42, 50

[300] J Thompson, Higgins, D., and T Gibson. Clustalw: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specic gap penalties and weight matrix choice. *Nucleic Acids Res*, 22:4673–4690, 1994. 58, 68, 69, 139

BIBLIOGRAPHY

[301] GM Torrie and JP Valleau. Nonphysical sampling distributions in monte carlo free-energy estimation: Umbrella sampling. *J Chem Phys*, 23:187–199, 1977. 103

[302] J Toshima, JY Toshima, AC Martin, and DG Drubin. Phosphoregulation of arp2/3-dependent actin assembly during receptor-mediated endocytosis. *Nat Cell Biol*, 7(3):246–54, 2005. 51, 54, 57

[303] F Trettel, D Rigamonti, P Hilditch-Maguire, VC Wheeler, AH Sharp, F Persichetti, E Cattaneo, and ME MacDonald. Dominant phenotypes produced by the hd mutation in sthdh(q111) striatal cells. *Hum Mol Gen*, 9(19):2799–2809, 2000. 14, 18, 19

[304] Y Trottier, Y Lutz, G Stevanin, G Imbert, D Devys, G Cancel, F Saudou, C Weber, G David, L Tora, Y Agid, A Brice, and JL Mandel. Polyglutamine expansion as a pathological epitope in huntingtons-disease and 4 dominant cerebellar ataxias. *Nature*, 378(6555):403–406, 1995. 9, 15

[305] N Troullier and JL Martins. Efficient pseudopotentials for plane-wave calculations. *Phys Rev B*, 43(3):1993–2006, 1991. 97, 120

[306] R Truant, RS Atwal, C Desmond, L Munsie, and T Tran. Huntington's disease: revisiting the aggregation hypothesis in polyglutamine neurodegenerative diseases. *Febs J*, 275(17):4252–4262, 2008. 3, 7, 9, 10, 11, 42, 57, 61, 62

[307] K Tsemekhman, L Goldschmidt, D Eisenberg, and D Baker. Cooperative hydrogen bonding in amyloid formation. *Protein Sci*, 16(4):761–4, 2007. 32

[308] C Tuma, AD Boese, and NC Handy. Predicting the binding energies of h-bonded complexes: A comparative dft study. *Phys Chem Chem Phys*, 1:3939–3947, 1999. 95

[309] MW van der Kamp, KE Shaw, CJ Woods, and AJ Mulholland. Biomolecular simulation and modelling: status, progress and prospects. *J R Soc Interface*, 5:S173–S190, 2008. 65, 77

[310] D van der Spoel, E Lindahl, B Hess, G Groenhof, AE Mark, and HJ Berendsen. Gromacs: Fast, flexible, and free. *J Comput Chem*, 26(16):1701–1718, 2005. 32, 124

[311] D van der Spoel, R van Drunen, and HJC Berendsen. *GROningen MAchine for Chemical Simulation*. Department of Biophysical Chemistry, BIOSON Research Institute, Nijenborgh 4 NL-9717 AG Groningen, 1994. 30

[312] WF van Gunsteren, SR Billeter, AA Eising, PH Hunenberger, P Kruger, AE Mark, WRP Scott, and IG Tironi. *Biomolecular simulation: the GROMOS96 manual and user guide*. Zurich: Hochschulverlag AG an der ETH, 1996. 30

[313] Z Varga and A Kovacs. Hydrogen bonding in peptide secondary structures. *Int J Quantum Chem*, 105(4):302–312, 2005. 32

[314] J Velier, M Kim, C Schwarz, TW Kim, E Sapp, K Chase, N Aronin, and M DiFiglia. Wild-type and mutant huntingtins function in vesicle trafficking in the secretory and endocytic pathways. *Exp Neurol*, 152(1):34–40, 1998. 14

[315] H Venselaar, RP Joosten, B Vroling, CAB Baakman, ML Hekkelman, E Krieger, and G Vriend. Homology modelling and spectroscopy, a never-ending love story. *Eur Biophys J*, 39(4):551–63, 2010. 66, 71

[316] R Viswanathan, A Asensio, and JJ Dannenberg. Cooperative hydrogen-bonding in models of antiparallel -sheets. *J Phys Chem A*, 108(42):9205–9212, 2004. 32

[317] SH Vosko, L Wilk, and M Nusair. Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis. *Can J Phys*, 58:1200–1211, 1980. 95

[318] FO Walker. Huntington's disease. *Lancet*, 369(9557):218–28, 2007. 1

[319] F Wang and DP Landau. Efficient, multiple-range random walk algorithm to calculate the density of states. *Phys Rev Lett*, 86(10):2050–2053, 2001. 104

[320] J Wang, RM Wolf, JW Caldwell, PA Kollman, and DA Case. Development and testing of a general amber force field. *J Comput Chem*, 25(9):1157–74, 2004. 46

[321] JM Wang, P Cieplak, and PA Kollman. How well does a restrained electrostatic potential (resp) model perform in calculating conformational energies of organic and biological molecules? *J Comput Chem*, 21(12):1049–1074, 2000. 28

[322] EE Wanker. Protein aggregation and pathogenesis of huntington's disease: mechanisms and correlations. *Biol Chem*, 381(9-10):937–42, 2000. 10

[323] A Warshel and M Levitt. Theoretical studies of enzymic reactions - dielectric, electrostatic and steric stabilization of carbonium-ion in reaction of lysozyme. *J Mol Biol*, 103(2):227–249, 1976. 100

[324] WJ Welch and MI Diamond. Glucocorticoid modulation of androgen receptor nuclear aggregation and cellular toxicity is associated with distinct forms of soluble expanded polyglutamine protein. *Hum Mol Gen*, 10(26):3063–3074, 2001. 14

[325] CL Wellington, LM Ellerby, C-A Gutekunst, D Rogers, S Warby, RK Graham, O Loubser, J van Raamsdonk, R Singaraja, Y-Z Yang, J Gafni, D Bredesen, SM Hersch, BR Leavitt, S Roy, DW Nicholson, and MR Hayden. Caspase cleavage of mutant huntingtin precedes neurodegeneration in huntington's disease. *J Neurosci*, 22(18):7862–72, 2002. 7

[326] CL Wellington, LM Ellerby, AS Hackam, RL Margolis, MA Trifiro, R Singaraja, K McCutcheon, GS Salvesen, SS Propp, M Bromm, KJ Rowland, TQ Zhang, D Rasper, S Roy, N Thornberry, L Pinsky, A Kakizuka, CA Ross, DW Nicholson, DE Bredesen, and MR Hayden. Caspase cleavage of gene products associated with triplet expansion disorders generates truncated fragments containing the polyglutamine tract. *J Biol Chem*, 273(15):9158–9167, 1998. 14

[327] R Wieczorek and JJ Dannenberg. H-bonding cooperativity and energetics of alpha-helix formation of five 17-amino acid peptides. *J Am Chem Soc*, 125(27):8124–9, 2003. 32

[328] TE Williamson, A Vitalis, SL Crick, and RV Pappu. Modulation of polyglutamine conformations and dimer formation by the n-terminus of huntingtin. *J Mol Biol*, 396(5):1295–309, 2010. 3, 11, 42, 44, 45, 50

[329] Z Xiang, CS Soto, and B Honig. Evaluating conformational free energies: the colony energy and its application to the problem of loop prediction. *P Natl Acad Sci Usa*, 99(11):7432–7, 2002. 70

[330] W Yu, L Liang, Z Lin, S Ling, M Haranczyk, and M Gutowski. Comparison of some representative density functional theory and wave function theory methods for the studies of amino acids. *J Comput Chem*, 30(4):589–600, 2009. 120

[331] D Zanuy, K Gunasekaran, AM Lesk, and R Nussinov. Computational study of the fibril organization of polyglutamine repeats reveals a common motif identified in beta-helices. *J. Mol. Biol.*, 358(1):330–345, 2006. 9, 15, 16, 32

[332] Y Zhao and D Truhlar. Benchmark databases for nonbonded interactions and their use to test density functional theory. *J Chem Theory Comput*, 1(3):415–432, 2005. 120

[333] YL Zhao and YD Wu. A theoretical study of beta-sheet models: is the formation of hydrogen-bond networks cooperative? *J Am Chem Soc*, 124(8):1570–1, 2002. 32, 34, 37

[334] HY Zoghbi and HT Orr. Glutamine repeats and neurodegeneration. *Annu Rev Neurosci*, 23:217–247, 2000. 7, 14, 15

[335] C Zuccato, M Valenza, and E Cattaneo. Molecular mechanisms and potential therapeutical targets in huntington's disease. *Physiol Rev*, 90(3):905–981, 2010. 6, 7, 9, 10, 62

Declaration

I herewith declare that I have produced this paper without the prohibited assistance of third parties and without making use of aids other than those specified; notions taken over directly or indirectly from other sources have been identified as such. This thesis has not previously been presented in identical or similar form to any other German or foreign examination board.

Publications related to the Thesis

1. *Submitted* Rossetti, G., Magistrato, A., Diamond, D., Carloni, P., *F-Actin modulates Huntingtin Exon 1 fibrillation*
2. *Submitted* Rossetti, G., Cossio, P., Laio, A., Carloni, P., *NT-Htt folding ensembles: implications in Htt aggregation and pathogenicity.*
3. Rossetti, G., Magistrato, A., Pastore, A. Carloni, P. (2010). *Hydrogen Bonding Cooperativity in polyQ β-Sheets from First Principle Calculations.* J Chem Theory Comput 6, 1777-1782.
4. Rossetti, G., Magistrato, A., Pastore, A., Persichetti, F. Carloni, P. (2008). *Structural properties of polyglutamine aggregates investigated via molecular dynamics simulations.* J Phys Chem B 112, 16843-50.

Other Publications

1. Rossetti, G., Cong, X., Caliandro, R., Carloni, P., *Common structural traits across disease-linked variants of the human prion protein and their implications for familial prion diseases.*
2. Rossetti, G., Giachin, G., Legname, G. Carloni, P. (2010). *Structural facets of disease-linked human prion protein mutants: A molecular dynamic study.* Proteins 78, 3270-3280.
3. Kranjc, A., Bongarzone, S., Rossetti, G., Biarnes, X., Cavalli, A., Bolognesi, M. L., Roberti, M., Legname, G. Carloni, P. (2009). *Docking Ligands on Protein Surfaces: The Case Study of Prion Protein.* J Chem Theory Comput 5, 2565-2573
4. Flock, D., Rossetti, G., Daidone, I., Amadei, A. Di Nola, A. (2006). *Aggregation of small peptides studied by molecular dynamics simulations.* Proteins 65, 914-21.