ABSTRACT

ZHANG, YUAN. Structural Characterization of Various Peptides and Nucleotides Associated with Neurodegenerative Diseases. (Under the direction of Dr. Celeste Sagui and Dr. Christopher Roland.)

Eukaryotic proteomes are rich in “low-complexity” sequences, which lack well-defined three-dimensional structures and are known as intrinsically disordered proteins (IDPs) or proteins with intrinsically disordered regions (IDRs). These disordered proteins have been shown to be functionally important to the cell, and the impairment of their physiological functions results in a wide variety of disease including cancers, neurodegenerative diseases, diabetes and prion diseases. Since the secondary structure of IDPs are typically transient, weakly populated, and confined to short stretches of the proteins sequence, the structural characterization of these proteins is often difficult. From a computational point of view, molecular dynamic (MD) simulations show varying results due to different force fields and sometimes inadequate sampling in such disordered proteins. Also, the use of different methods to characterize the secondary structure motifs can further complicate comparisons and interpretation of the data. Here we show several MD simulations of disordered peptides, specifically these including \( \text{gp41}_{659-671} \) (ELLELDKWASLWN), the homo-peptide polyasparagine (\( N_{18} \)), and polyasparagine dimers (\( N_8-N_8 \)) in order to present how a proper combination of computational tools can be used to characterize challenging disordered proteins and to unify seemingly contradictory experimental result.

Sequences rich in glutamine (Q) and asparagine (N) are intrinsically disordered in monomeric form but can aggregate into highly ordered amyloids, as seen in Q/N-rich prion domains (PrDs). Pure polyglutamine (polyQ) stretches in humans are associated with at least nine late-onset progressive neurodegenerative diseases, which are caused by polyQ expansions greater than a given threshold in proteins with little or no similarity except for the polyQ region. Although each disease has a different pathology, they all share some common feature such as the formation of polyQ aggregates, and eventual neuronal death. Here, we consider a variety of N/Q-rich peptides, including sequences found in the yeast Sup35 PrD, in parallel and antiparallel \( \beta \)-sheet aggregates, and probe via fully atomistic molecular dynamics simulations all their possible steric-zipper interfaces to determine their aprotofibril structure and their relative stability. Our results show that polyglutamine aggregates are more stable than polyasparagine aggregates. Enthalpic
contributions to the free energy favor the formation of polyQ protofibrils, while entropic contributions favor the formation of polyN protofibrils. The considerably larger phase space that disordered polyQ must sample on its way to aggregation probably is at the root of the associated slower kinetics observed experimentally. When other amino acids are present, such as in the Sup35 PrD, their shorter side chains favor steric-zipper formation for N but not Q, as they preclude the in-register association of the long Q side chains.

Besides, we also have studied nucleotide repeat disorder structures, which associate with two neurodegenerative diseases, frontotemporal dementia (FTD) and amyotrophic lateral sclerosis (ALS). A (GGGGCC) hexanucleotide repeat (HR) expansion in the C9ORF72 gene, and its associated antisense (CCCCGG) expansion, are considered the primary cause behind FTD and ALS. We have performed molecular dynamics simulations to characterize the conformation and dynamics of the twelve duplexes that result from the three different reading frames in sense and antisense HRs for both DNA and RNA. Our results show that G-rich double helices are more stable than C-rich helices due to better stacking and hydrogen bonds (HBs) of G-G mismatches. The inner G-G mismatches stay inside the helix in G\textsubscript{syn}-G\textsubscript{anti} conformations and form two HBs, O6-N1H and N7-N2H in DNA form and one additional HB OP2-N2H within G\textsubscript{syn} base in RNA form. Inner G-G mismatches cause local unwinding of the helix. C-rich helix conformations vary wildly. C mismatches flip out of the helix in DNA but not in RNA. Least (most) stable C-rich RNA and DNA helices have single (double) mismatches separated by two (four) Watson-Crick basepairs. The most stable DNA structure displays an “e-motif” where mismatched bases flip towards the minor groove pointing in the 5’ direction. There are two RNA conformations, where the orientation and HB pattern of the mismatches are coupled to bending of the helix. Ion distributions and ion bridges that stabilize structures are described.
Structural Characterization of Various Peptides and Nucleotides Associated with Neurodegenerative Diseases

by

Yuan Zhang

A dissertation submitted to the Graduate Faculty of North Carolina State University in partial fulfillment of the requirements for the Degree of Doctor of Philosophy

Physics

Raleigh, North Carolina

2016

APPROVED BY:

Dr. Keith Weninger

Dr. Robert B. Rose

Dr. Celeste Sagui
Co-chair of Advisory Committee

Dr. Christopher Roland
Co-chair of Advisory Committee
DEDICATION

To my beloved parents,

Yuling Sui and Ruiqi Zhang,

for all their love and support and putting me through the best education possible.
I appreciate your sacrifices and I wouldn’t have been able to get to this stage without you.

To my dear husband,

Bin Xiao,

for his unending support, level-headedness and love.

I am truly thankful for having you in my life.
BIOGRAPHY

Born in a small city of Yinchuan located northwest of China, on January 12, 1987, to Ruiqi and Yuling Sui, Yuan Zhang is the only child in the family. Even though her parents are not highly educated people, they encouraged her to pursue science and mathematics from an early age. In 2005, Zhang enrolled in one of the most prestigious universities in China, University of Science and Technology of China (USTC), and got her bachelor degree in physics in 2009.

In the last two years in USTC, Zhang worked in a computational physics laboratory on a theoretical study of electronic structure of graphene under the advice of Dr. Bicai Pan and Dr. Haiyan He. In 2010, Zhang entered the doctoral program in physics at North Carolina State University (NCSU), where she joined the computational biophysics group under the advice of Dr. Celeste Sagui and Dr. Christopher Roland. To date, her research has been focused on structural and dynamic characterization of various proteins and nucleotides associated with neurodegenerative diseases by using all-atom molecular dynamics.
ACKNOWLEDGEMENTS

The completion of this project could not have been possible without the support, patience and generosity of many people.

Firstly, I would like to express my sincere gratitude to my advisor Prof. Celeste Sagui. I appreciate all her contributions of time and ideas to make my Ph.D. experience productive and stimulating. The joy and enthusiasm she has for her research were contagious and motivational for me, especially during tough times in the Ph.D. pursuit. She has provided me an excellent example of a successful woman physicist and professor. I also want to thank my advisor Prof. Christopher Roland for the continuous support of my Ph.D. study and related research, for his patience, encouragement, and immense knowledge. I could not have imagined having better advisors and mentors for my Ph.D. study.

Besides, I would like to thank the rest of my thesis committee: Prof. Keith Weninger and Prof. Robert Rose, for their insightful comments.

My sincere thanks also go to my current and former colleagues: Dr. Volodymyr Babin, Dr. Jung-Goo Lee, Dr. Mahmoud Moradi, Dr. Veit Man, and Feng Pan for the stimulating discussion, and for all the fun we have had in the last several years.
## TABLE OF CONTENTS

### LIST OF TABLES

<table>
<thead>
<tr>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>viii</td>
</tr>
</tbody>
</table>

### LIST OF FIGURES

<table>
<thead>
<tr>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ix</td>
</tr>
</tbody>
</table>

### Chapter 1 Introduction

1. Computational approaches for the characterization of challenging intrinsically disordered peptides ......................................................... 2
2. Amyloid protofibril structures of polyQ and polyN ........................................... 6
3. DNA/RNA duplex obtained from the GGGGCC and CCCCGG hexanucleotide repeats ................................................................. 8
4. References ........................................................................................................... 10

### Chapter 2 The gp41_{659−671} HIV-1 Antibody Epitope: A Structurally Challenging Small Peptide

1. Introduction ........................................................................................................ 22
2. Computational methods ...................................................................................... 23
3. Implicit solvent simulations ............................................................................. 24
4. Explicit solvent simulations ............................................................................. 25
5. Results ................................................................................................................. 25
6. Discussion .......................................................................................................... 29
7. Comparison to previously published experimental data .................................. 29
8. Comparison to previous simulations ................................................................. 31
9. Comparison between the force fields ............................................................... 32
10. Conformational equilibrium in stretched conformations ............................... 32
11. Conclusion ......................................................................................................... 33
12. References ........................................................................................................ 33

### Chapter 3 Secondary structure assignment for conformationally irregular peptides: Comparison between DSSP, STRIDE and KAKSI

1. Introduction ........................................................................................................ 35
2. Method ................................................................................................................ 36
3. DSSP .................................................................................................................. 37
4. STRIDE .............................................................................................................. 37
5. KAKSI ............................................................................................................... 38
6. SOV ................................................................................................................... 38
7. Simulation details .............................................................................................. 39
8. gp41_{659−671} peptide ....................................................................................... 39
9. Polyasparagine ................................................................................................. 39
10. Polyasparagine dimers .................................................................................... 39
Chapter 4 Amyloid properties of asparagine and glutamine in prion-like proteins

Introduction .................................................................................. 50
Results and discussion ................................................................. 51
   Initial structures ........................................................................ 51
   Monomeric polyQ β-hairpins are more stable than monomeric polyN β-hairpins ................................................................ 53
   N-rich oligomers are most stable in “class 1” steric zippers with 2-by-2 interdigitation .............................................................. 54
   PolyQ oligomers are most stable in antiparallel-stranded β-sheets with 1-by-1 steric zippers .......................................................... 55
   The most stable polyQ structures show higher stability than the most stable polyN structures ......................................................... 56
   The enthalpic contributions to the free energy favor polyQ over polyN ................................................................................ 57
   The entropic free energy difference for the formation of steric zippers is higher for polyQ than for polyN ............................................ 58
   Other interspersed amino acids create more disruption in polyQ zippers ................................................................. 58
Conclusions .................................................................................. 59
Methods ....................................................................................... 60
   Molecular Dynamics Simulations .................................................. 60
   Laser Melting Simulations ............................................................ 60
References ..................................................................................... 60
Supporting information ................................................................. 62
   MD Simulations ......................................................................... 62
   Laser Simulations ....................................................................... 64
   Analysis ...................................................................................... 65

Chapter 5 Structure and dynamics of DNA and RNA double helices obtained from the GGGGCC and CCCCCG hexanucleotide repeats that are the hallmark of C9FTD/ALS diseases ........................................................................ 74
Introduction .................................................................................. 75
Results .......................................................................................... 77
   Initial structures ........................................................................ 77
#### Chapter 6 Conclusions

References ........................................... 133

References ........................................... 136


### LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 2.1</td>
<td>Secondary structure population based on STRIDE for ( \text{gp}41_{659-671} ) and the epitope ELDKWA.</td>
<td>29</td>
</tr>
<tr>
<td>Table 3.1</td>
<td>Secondary structure content (%) of ( \text{gp}41_{659-671} ), ( N_{18} ) and dimers ( N_8-N_8 ) in parallel (P) and anti-parallel (AP) orientation, with a restrained (R) or unrestrained (U) backbone.</td>
<td>41</td>
</tr>
<tr>
<td>Table 3.2</td>
<td>Average SOV measures (%) between KAKSI and DSSP/STRIDE based on helix (including “H”, “G”, “I” from DSSP and STRIDE), ( \beta ) sheet (including “B”, “E” and “b” from DSSP and STRIDE) and coil (including “T” and “C” from DSSP and STRIDE). KAKSI is chosen as reference.</td>
<td>41</td>
</tr>
<tr>
<td>Table 3.3</td>
<td>Same as Table 2, except that DSSP is taken as reference.</td>
<td>41</td>
</tr>
<tr>
<td>Table 3.4</td>
<td>Average SOV measures (%) between KAKSI and DSSP/STRIDE based on helix (including “H”, “G”, “I” from DSSP and STRIDE), ( \beta ) sheet (including “B”, “E” and “b” from DSSP and STRIDE) and coil (including “T” and “C” from DSSP and STRIDE). KAKSI is chosen as reference.</td>
<td>41</td>
</tr>
<tr>
<td>Table 3.5</td>
<td>C3 scores (%) between KAKSI, DSSP and STRIDE on ( \text{gp}41_{659-671} ) and ( N_{18} ).</td>
<td>42</td>
</tr>
<tr>
<td>Table 4.1</td>
<td>Characterization of the initial conformations.</td>
<td>52</td>
</tr>
<tr>
<td>Table 5.1</td>
<td>Characterization of the initial conformations</td>
<td>78</td>
</tr>
<tr>
<td>Table 5.S1</td>
<td>Hydrogen bond population of G-G mismatches in double (D) and single (S) models for DNA and RNA.</td>
<td>112</td>
</tr>
<tr>
<td>Table 5.S2</td>
<td>Hydrogen bond population of C-C mismatches in double (D) (G6-G19 and G7-G18) and single (S) (G4-G22, G7-G19, G10-G16) models for DNA and RNA.</td>
<td>113</td>
</tr>
<tr>
<td>Table 5.S3</td>
<td>Structure prediction for 10 repeats of the GGGGCC and CCCCGG hexanucleotide sequences and the corresponding Gibbs energy as calculated by the Mfold web server [23]. Restraints have been imposed in order to obtain the duplex models considered in this paper. The restraint position indicates the pair base where a restraint force is applied on. The folding temperature is fixed at 310 K.</td>
<td>114</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

Figure 2.1 Conformations of the gp41 11-mer peptide \( \text{gp41}_{659-671} \) minus the terminal residues, i.e., LLELDKWASLW) from the 1TJH crystal structure and the \( \text{gp41}_{659-671} \) 13-mer peptide from the 1LCX NMR solution structure. The flanking residues E662 and A667 of the epitope ELDKWA are shown. The color scheme for the secondary structure (generated by VMD) is as follows: \( 3_{10} \) helix: blue; turn: cyan; coil: white. ................................. 24

Figure 2.2 Snapshot of chosen initial conformations for the unrestrained MD simulations under the ff99SB force field. The top of the peptide corresponds to E659 and the bottom to W671. This conformations all evolved from a fully extended initial structure. .................. 25

Figure 2.3 Free energy profiles as a function of the end-to-end distance for \( \text{gp41}_{659-671} \). Results for AMBER ff99SB (top) and ff12SB (bottom) in implicit (black) and explicit (red) solvent are shown. ............ 26

Figure 2.4 Ramachandran plots for each residue in ELELDKWASLWN for the unrestrained MD simulations. Results for the AMBER ff99SB are presented in blue and for ff12SB in red/yellow. For each force field, the top two rows show the plots for implicit solvent, while the bottom two rows show the plots for the explicit solvent. ............... 27

Figure 2.5 Secondary structure map, as obtained with STRIDE, of the unrestrained MD simulations under implicit ff99SB (top) and explicit ff99SB (bottom). The red, blue, yellow, green and white colors correspond to \( \alpha \) helix, \( 3_{10} \) helix, \( \pi \) helix, turn, and “other” secondary structural motifs, respectively. ................................. 28

Figure 2.6 Secondary structure map, as obtained with STRIDE, of the unrestrained MD simulations under implicit ff12SB (top) and explicit ff12SB (bottom). The red, blue, yellow, green and white colors correspond to \( \alpha \) helix, \( 3_{10} \) helix, \( \pi \) helix, turn, and “other” secondary structural motifs, respectively. ................................. 28

Figure 2.7 Average secondary structure map, as obtained with STRIDE, of the unrestrained MD simulation under implicit ff99SB, explicit ff99SB, implicit ff12SB and explicit ff12SB (from left to right). Secondary structural motifs are: \( \alpha \) helix (red), \( 3_{10} \) helix (blue), \( \pi \) helix (yellow), turn (green), and “other” (white). ................................. 28

Figure 2.8 Scatter plot of the RMSD values of the epitope ELDKWA with respect to the corresponding sequence motifs of the peptide in the 1TJG crystal structure (x-axis) and in the 1LCX NMR structure (y-axis). The coordinate data are from the unrestrained MD simulation under explicit ff99SB (top) and explicit ff12SB (bottom). .......... 29
Figure 2.9 RMSD of predicted chemical shifts with respect to experimental chemical shifts for HN (top) and HA (bottom). The experimental data corresponds to 1LCX and the predicted data is from the simulations under explicit ff99SB (red) and explicit ff12SB (blue).

Figure 2.10 Secondary structure calculated with STRIDE for the ff99SB umbrella sampling runs. The y-axis shows the residue and the x-axis shows the end-to-end distance used as restraint in the umbrella sampling runs. Each small window represents an average for the given residue over 100 ns. Top graph corresponds to implicit solvent and bottom graph to explicit solvent. Red, blue, yellow, green and white correspond to α helix, 3_10 helix, π helix, turn, and “other” secondary structural motifs, respectively.

Figure 2.11 Secondary structure calculated with STRIDE for the ff12SB umbrella sampling runs. The y-axis shows the residue and the x-axis shows the end-to-end distance used as restraint in the umbrella sampling runs. Each small window represents an average for the given residue over 100 ns. Top graph corresponds to implicit solvent and bottom graph to explicit solvent. Red, blue, yellow, green and white correspond to α helix, 3_10 helix, π helix, turn, and “other” secondary structural motifs, respectively.

Figure 2.12 Scatter plot of the RMSD values of the epitope ELDKWA with respect to the corresponding sequence motifs of the peptide in the 1TJG crystal structure (x-axis) and in the 1LCX NMR structure (y-axis). The coordinate data are from the umbrella simulation under explicit ff99SB (top) and explicit ff12SB (bottom) for an end-to-end distance of 28 Å.

Figure 3.1 (a) Snapshot of a sample conformation of gp41_659-671 with secondary structure assigned by DSSP (left), STRIDE (middle), and KAKSI (right). (b) Snapshot of a sample conformation of N_18 with secondary structure assigned by DSSP (left), STRIDE (middle), and KAKSI (right). (c) Snapshot of two sample conformations of dimer N_8-N_8 with secondary structure assigned by DSSP (left) and STRIDE (right). Colors purple, blue, green, yellow and white represent α helix (H), 3_10 helix (G), turn (T), β sheet (E) and coil (C) for DSSP and STRIDE assignments; for KAKSI, red represents helix and white represents coil.

Figure 3.2 Length distribution of α helix, 3_10 helix, and π helix of gp41 as assigned by DSSP (blue) and STRIDE (green).

Figure 3.3 Length distribution of α helix, 3_10 helix, and π helix of N_18 as assigned by DSSP (blue) and STRIDE (green).

Figure 3.4 Length distribution of “Helix” for gp41_659-671 (top) and N_18 (bottom).
Figure 3.5  The population distribution of $\text{gp41}_{659-671}$ (left) and $N_{18}$ (right) on the plane determined by the alignment angle $\theta$ and the hydrogen bond length. The alignment angle $\theta$ is defined as the N-(H)-O angle, and the hydrogen bond length is defined as the distance between atoms N and O in each hydrogen-bond pair. The hydrogen bonds are assigned by DSSP (top) and STRIDE (middle). The two bottom pictures show the corresponding population distributions versus the alignment angle, with hydrogen bonds assigned by DSSP in blue and by STRIDE in green.  

Figure 3.6  Odd rows show Ramachandran plots (y-axis is $\psi$ angle in degrees, x-axis is $\phi$ angle in degrees) and even rows show translation length distribution (y-axis) as a function of helical rise (x-axis, in Å) for $3_{10}$ helix residues in $\text{gp41}_{659-671}$ (top) and $N_{18}$ (bottom). The DpSp (left column), DpSn (middle column) and DnSp (right column) are residues assigned by DSSP and STRIDE as pp, pn and np, where p indicates that the residue is assigned as $3_{10}$ helix by the respective algorithm, and n indicates that the residue is not assigned to that state by the algorithm.

Figure 3.7  Odd rows show Ramachandran plots (y-axis is $\psi$ angle in degrees, x-axis is $\phi$ angle in degrees) and even rows show translation length distribution (y-axis) as a function of helical rise (x-axis, in Å) for $\alpha$ helix residues in $\text{gp41}_{659-671}$ (top) and $N_{18}$ (bottom). The DpSp (left column), DpSn (middle column) and DnSp (right column) are residues assigned by DSSP and STRIDE as pp, pn and np, where p indicates that the residue is assigned as $\alpha$ helix by the respective algorithm, and n indicates that the residue is not assigned to that state by the algorithm.

Figure 3.8  The dihedral angle sum $\psi_i + \phi_{i+1}$ distribution of $3_{10}$ helix residues (dashed line) and $\alpha$ helix residues (solid line) as assigned by DSSP (blue) and STRIDE (red). The standard dihedral angles sum values are marked by the black lines, located at $-75^\circ$ (dashed line) for the $3_{10}$ helix and at $-104^\circ$ (solid line) for the $\alpha$ helix. Top: $\text{gp41}_{659-671}$; bottom: $N_{18}$.

Figure 3.9  The dihedral angle sum $\psi_i + \phi_{i+1}$ distribution of $3_{10}$ helix residues (top) and $\alpha$ helix residues (bottom) as assigned by DSSP and STRIDE: DpSn (DSSP positive, STRIDE negative) is in blue, DnSp (DSSP negative, STRIDE positive) is in red.

Figure 3.10  Pie chart of secondary structure distribution assigned by DSSP for all the residues identified as “turn” by STRIDE (i.e., the STRIDE graphs in either case are a full green “turn” circle). Turn: green; bend: yellow; coil: blue; other: red. Left: $\text{gp41}_{659-671}$; right: $N_{18}$. 

xi
Figure 4.1  Eight classes of steric zippers, following Sawaya et al. The sheet closer to the viewer is shown in blue, and the one behind in orange. The solid colors show the “front side” of each sheet with the side chains identified by even numbers, and the faded colors show the “back side” of the sheets with odd-number side chains.

Figure 4.2  For not too short homopeptides, some of the classes described in 4.1 converge. (a) For parallel sheets class 1 and 4 become the same (with sheets stacked in antiparallel fashion), and class 2 and 3 become the same (with sheets stacked in parallel fashion). (b) Schematic of the quarter-stagger displacement of one sheet with respect to the other. For antiparallel sheets undergoing a quarter-stagger, class 5-8 converge into class A (right). (c) Side-chain interdigitation of steric zippers.

Figure 4.3  This figure contrasts the relative stability of polyQ versus polyN hairpins. The top panel shows sample initial and final configurations after a 100 ns solvated simulation of Q23 and N23 β-sheets with an initial β-hairpin (top panel). After 100 ns the polyQ hairpin partially retains its structure, while polyN is disordered. A more quantitative expression is given by the middle panel which gives the percentage of β-sheet content of the hairpins as function of time, while the bottom panel shows the β-sheet content of each residue averaged over the last 50 ns of the simulation. Here, black and red lines denote results for Q23 and N23, respectively, and averages are performed over 10 runs.

Figure 4.4  For each 4 × 2 model labeled in the figure and described in Table 4.1, sample initial configurations after equilibration are shown on the left and sample final configurations at 400 ns are shown on the right.

Figure 4.5  For each 4 × 2 model labeled in the figure and described in Table 4.1, sample initial configurations after equilibration are shown on the left and sample final configurations at 400 ns are shown on the right. Configurations on the left column correspond to polyN aggregates and on the right column, to polyQ aggregates.

Figure 4.6  For each 2 × 2 model labeled in the figure and described in Table 4.1, sample initial configurations after equilibration are shown on the left and sample final configurations at 100 ns are shown on the right. Top row corresponds to polyN aggregates and bottom row corresponds to polyQ aggregates.
Figure 4.7 RMSD values of the main chain atoms N, O, Cα and C' in various models with respect to the corresponding initial structure (after minimization). The values are averaged every 2 ns for 4×2 models, and every 0.2 ns for 2×2 models, and then averaged over 10 runs.

Figure 4.8 β-sheet content of the 4×2 models 4T1 (black), 4T2 (red), 4T3 (blue), 4T4 (orange), 4O9 (green) in top left chart; 4×2 models 4O1 (black), 4O2 (red), 4O3 (blue), 4O4 (orange) in top right chart; 4×2 models 4O5 (black), 4O6 (red), 4O7 (blue), 4O8 (orange), 4O10 (green) in bottom left chart; 2×2 models 2T1 (black), 2T2 (red), 2T3 (blue), 2O2 (orange), 2O4 (green), 2O9 (cyan) and 2O10 (magenta) in bottom right chart. Results are averaged over 10 independent runs.

Figure 4.9 Time dependence of the rmsd and β-sheet content (taken with respect to the initial structure) of model 4O10 (panels a,c) and 4T3 (panels b,d) subject to an infrared laser pulse with modulus of the electric field amplitude corresponding to $E_0=1.8$ (black), $2.0$ (red), $2.2$ (green) and $2.4$ (blue) V/nm, respectively.

Figure 4.10 Side-chain packing and hydrogen-bond patterns in 4T1 (a), 4T3 (b) and 4O10 (c) models. Four inner chains labeled A (tan), B (green), C (cyan) and D (purple) are shown. The orange dash lines indicate hydrogen bonds. The N, ND2 (in N-residues), NE2 (in Q-residue) atoms are shown in blue balls. The O, OD1 (in N-residues), OE2 (in Q-residue) atoms are shown in red balls. The CB atoms are shown in yellow balls. The label $X^A_4$ indicates the X-residue (X = Q or N) of the chain A at fourth position (counting the first one from N-terminal for each chain).

Figure 4.11 Dynamic evolution of the β-sheet content of all possible mutations of the 2 × 2 2O10 class A structure (QQQQQQQQ) based on mutating one or two Qs into Gs. The time evolution of the β-content is plotted in orange. The mutated residues Gly (G) and Tyr (Y) on the different sequences are marked in green and purple, respectively. The blue lines separate the different models.

Figure 4.S1 The frequency dependence of (a) maximum of the C=O bond length fluctuations; (b) kinetic energy difference; (c) total energy difference, and (d) potential energy difference. Energy differences are taken with respect to the corresponding energy average value as obtained from an equilibrium nanosecond simulation.

Figure 4.S2 Energy differences (with respect to the equilibrium energy value) of laser pulse simulations as a function of different electric field strengths $E_0$: (a) 1.8; (b) 2.0; (c) 2.2; and (d) 2.4 V/nm, respectively.
Figure 4.S3 Intersheet distance for each model. The values are averaged every 2 ns for 4×2 models, and every 0.2 ns for 2×2 models, and then averaged over 10 runs. ........................................... 69

Figure 4.S4 Interstrand distance for each model. The values are averaged every 2 ns for 4×2 models, and every 0.2 ns for 2×2 models, and then averaged over 10 runs. ........................................... 70

Figure 4.S5 RMSD (top), β-content (2nd), intersheet distance (3rd) and interstrand distance (bottom) of final step simulations with their error bars on hairpin systems (last 2 ns simulation, blue), 4×2 models T (last 2 ns, orange) and models O (last 2 ns, green), as well as 2×2 models (last 0.2 ns, magenta). ........................................... 71

Figure 5.1 Structures of duplex configurations. Gs are in blue and Cs in red. 79

Figure 5.2 PCA distributions of G-G (1st column) and C-C (4th column) mismatched duplexes over the 500 ns - 1 µs interval projected onto the first three principal components for DNA models (a) and RNA models (b). And the corresponding visualization of the atomic motions involved in the 1st (blue) and 2nd (red) principal components. . 81

Figure 5.3 Backbone sugar torsion δ and the glycosyl torsion χ angle distributions for DNA duplexes (left) and RNA duplexes (right) corresponding to G mismatches (top three rows) and C mismatches (bottom three rows). ........................................... 84

Figure 5.4 Open angle distribution of G mismatching models (left two columns) and C mismatching models (right two columns) in DNA (odds columns) and RNA (even columns) forms. Black and red lines show the distribution of G and C bases, respectively, that participate in Watson-Crick bonding. The green line shows the distribution for the internal mismatched bases (G or C). ........................................... 85

Figure 5.5 Simulations results for DNA SC-3. (a) This figure shows the DNA SC-3 evolution in the simulation set A from 0 ns (after equilibration) to 1 µs (final simulation time). First, one strand shifts by one residue in the 3'-5' direction, then the free ending bases bond to the Watson-Crick paired bases, causing the breaking of the hydrogen bonds. (b) In the simulation set B, one strand shifts by one residue in the 3'-5' direction, transforming the duplex into a partial DC-1 structure. The flipping out of a C basis allows the retention of the SC-3 configuration in the second half of the duplex. . . . . . 86

Figure 5.6 (a) Open angle of DNA DC-1 as a function of time (left) and corresponding distribution (right). (b) The e-motif configuration. (c) Open angle of DNA DC-2 as a function of time (left) and corresponding distribution (right). (d) Evolution of the C6-C19 mismatch (sideviews) and top view of the C6-C19 evolution. . . . . . 88
Figure 5.7 (a) Torsion angle $\alpha$ (black) and $\gamma$ (red) of C7 and C18 as a function of time for DC-1 in DNA form. Sum of torsion angles ($\alpha + \gamma$) for C7 (blue) and C18 (orange) as a function of time for DC-1 in DNA form. (b) Torsion angle $\epsilon$ (black) and $\zeta$ (red) of C18 and C19 as a function of time for DC-2 in DNA form. Difference of torsion angles ($\epsilon - \zeta$) for C18 (blue) and C19 (orange) as a function of time for DC-2 in DNA form.

Figure 5.8 Open angle of RNA DC-1 (a) and DC-2 (b) as a function of time (left) and corresponding distribution (right). (c) Configurations A and B corresponding to the average structures of DC-1 at 400-500 ns and 650-700 ns, and typical hydrogen bonds associated with them.

Figure 5.9 (a) Open angle of RNA SC-3 as a function of time (left) and corresponding distribution (right). (b) Typical structures of C10-C16 pair base.

Figure 5.10 Average over last 100 ns of twist angles of DNA duplexes (top) and RNA duplexes (bottom) as a function of step numbers for G-rich sequences.

Figure 5.11 Typical hydrogen bond of G bases (a) and C bases (b) in mismatched bases. Due to the syn-anti configuration, a G-G pair forms two hydrogen bonds in DNA form: N7-N2 and O6-N1; and forms three hydrogen bonds in RNA form: N7-N2, O6-N1 and OP2-N2. The C-C pair forms either N3-N4 or O2-N4 hydrogen bonds. (c) Presence of hydrogen bonds in only N3-N4 (red), in only O2-N4 (blue), or formed simultaneously (green), as a function of time for DC-1 (top 6) and DC-2 (bottom 6) in DNA (left) and RNA (right).

Figure 5.12 Atomistic details of direct binding sites of Na$^+$ for different mismatches. (a) G-G mismatch in DG models; (b) a G-C base pair adjacent to a G-G mismatch (c) C-C mismatch where the bases stay inside the helix and form N3-N4 H-bonds; (d) C-C mismatches where the bases form O2-N4 H-bonds; (e) a G-C base pair adjacent to a C-C mismatch can also contribute to trap the ion. In addition, the figure shows two examples of ions binding to the helix backbone. (f) ion bridging four OP1 atoms in DNA DC-1; (g) ion linking the ends of a DG-2 RNA helix.

Figure 5.S1 Definition of open angle for a duplex. The local axis U is the average vector of U1 and U2, defined as the vectors of C1'(i-1)-C1'(i+1) (or C1'(i-2)-C1'(i+2) for mismatches) along each strand. The open angle is defined by the vectors C1'-N9(N1) and C1'-C1' projected on the plane perpendicular to the local axis U.
Figure 5.S2 Displacement of the first principal component as a function of time, as well as snapshots corresponding to the bimodal distributions. The red structure is an average structure over 600 ns to 650 ns, and the blue structure is an average structure over 885 ns to 935 ns. 115

Figure 5.S3 Left: PCA distributions of R-SC-3 and SC-3+C DNA duplexes over the 500 ns - 1 \(\mu\)s interval projected onto the first three principal components. Right: Visualization of the atomic motions involved in the 1st (blue) and 2nd (red) principal components. 116

Figure 5.S4 Displacement of the first principal component as a function of time, as well as snapshots corresponding to the bimodal distributions. The red structure is an average structure over 575 ns to 620 ns, and the blue structure is an average structure over 850 ns to 900 ns. 117

Figure 5.S5 Displacement of the first principal component as a function of time, as well as snapshots corresponding to the bimodal distributions. The red structure is an average structure over 600 ns to 650 ns, and the blue structure is an average structure over 875 ns to 925 ns. 118

Figure 5.S6 Displacement of the first principal component as a function of time, as well as snapshots corresponding to the bimodal distributions. The red structure is an average structure over 500 ns to 550 ns, and the blue structure is an average structure over 950 ns to 1000 ns. 119

Figure 5.S7 Displacement of the first principal component as a function of time, as well as snapshots corresponding to the bimodal distributions. The red structure is an average structure over 650 ns to 700 ns, and the blue structure is an average structure over 925 ns to 975 ns. 120

Figure 5.S8 Displacement of the first principal component as a function of time, as well as snapshots corresponding to the bimodal distributions. The red structure is an average structure over 725 ns to 735 ns, and the blue structure is an average structure over 740 ns to 750 ns. 121

Figure 5.S9 \(\chi\)-angle of DC-1 (odd rows) and DC-2 (even rows) in DNA (top two) form and RNA (bottom two) form as a function of time (left) and the distribution (right). 122

Figure 5.S10(a) Open angle (blue line) of C18 superposed to the radius of gyration \((R_g)\) (red line) of the RNA duplex as a function of time. (b) Two conformations: A is bent towards the major groove when the mismatching C bases turn to the minor groove. B is an unbent duplex where the mismatching bases are in the major groove. 123

Figure 5.S11 Occupation (%) of ions trapped in the major (red) and minor (blue) grooves as a function of base pair for G mismatching duplexes (left) and C mismatching duplexes (right) in DNA form. 124
Figure 5.S12 Occupation (%) of ions trapped in major (red) and minor (blue) grooves as a function of base pair for G mismatching duplexes (left) and C mismatching duplexes (right) in RNA form. 125

Figure 5.S13 Radial distribution function of Na\(^+\) at C(O2) (left) and C(N4) (right) for C mismatching helices in DNA form. 126

Figure 5.S14 Radial distribution function of Na\(^+\) at C(O2) (left) and C(N4) (right) for C mismatching helices in RNA form. 127

Figure 5.S15 Secondary structure prediction with least Gibbs energy computed with Mfold of (a) DNA G mismatching models (b) DNA C mismatching models; (c) RNA G mismatching models; and (d) RNA C mismatching models, resulting in configurations (a) SG-3, (b) DC-1, (c) DG-1 and (d) DC-2. 128
Introduction

Proteins can take on four general classes of conformations: highly ordered (folded), molten globular, pre-molten globular and unfolded structures [1]. Unlike the ordered globular proteins, which are characterized by rigid 3-D structures, many proteins fail to form well-defined 3-D structures, and are best characterized by an ensemble of dynamic conformation in which the atomic positions and backbone Ramachandran angles vary significantly over time. Such proteins and regions are known as intrinsically disordered proteins (IDPs) or proteins with intrinsically disordered regions (IDRs) [2]. The structure of IDPs and IDRs can be either collapsed as a molten globule, or extended into a coil or pre-molten globular structure. Eukaryotic proteomes are rich in such “low-complexity” sequences, which are distinguished by the prevalence of certain amino acids. Lacking well-defined 3-D structures also means that these IDPs/IDRs are under-represented in the RCSB PDB [3]. In general, these sequences are quite abundant (14% of the proteins of Homo Sapiens include at least one simple sequence [4]) and either involve single amino-acid repeats, or just a few different amino acids. These disordered proteins have been shown to be functionally important to the cell. Their considerable plasticity makes it easier for IDPs/IDRs to interact efficiently with several different targets, or play crucial roles in cellular processes, such as transcription, translation, signaling, and in cell cycle control [2,5]. The impairment of their physiological functions results in a wide variety of disease: cancers (associate with the disordered proteins such as p53 [6], AFP [7, 8] and BRCA1 [9]), neurodegenerative diseases (including Parkinsons’ Disease (PD), Alzheimer’s disease (AD), Huntington disease, and other Synucleinopathies), prion diseases, diabetes, and cardiovascular disease (CVD) [10].
This thesis presents a study of the structural characterization of (1) Selected disordered proteins, specifically these including the challenging small IDPs $\text{gp41}_{659-671}$, $\text{N}_{18}$, and $\text{N}_{68}-\text{N}_{8}$; (2) ordered amyloids proteins, which result from the misfolding of polyQ and polyN that are IDPs in their monomeric forms; (3) DNA and RNA double helices with mismatched base pairs as obtained from the GGGGCC and CCCCCG hexanucleotide repeats.

1.1 Computational approaches for the characterization of challenging intrinsically disordered peptides

Because the secondary structure of IDPs is typically transient, weakly populated, and confined to short stretches of the protein sequence, the structural characterization of these proteins is often difficult both experimentally and theoretically [11]. Crystals of disordered molecules are hard to obtain. Even in the rare cases when they form, such crystals represent only one structure out of thousands of possible conformations in solution. Therefore, it is not feasible to adequately perform X-ray crystallography for the secondary structure characterization of such proteins. However, it has been shown that small-angle X-ray scattering (SAXS), NMR spectroscopy and single-molecule fluorescence combined with computational methods represent a powerful paradigm for the detailed structural characterization of IDPs [12]. In Chapter 2 and 3, we report on two characterization studies on a set of structurally challenging peptides: specifically, $\text{gp41}_{659-671}$ (ELLELDKWASL), the homo-peptide polyasparagine ($\text{N}_{18}$), and polyasparagine dimers ($\text{N}_{68}-\text{N}_{8}$). This part of the thesis illustrates how a judicious combination of computational tools (classical molecular dynamics (MD), as well as secondary structure assignment algorithms) can be used to characterize challenging IDPs and to unify seemingly contradictory experimental result.

In the HIV-1 virus, $\text{gp41}$ and $\text{gp120}$ are two transmembrane glycoproteins that mediate the fusion of the virus with the host cell. The structure of $\text{gp41}$ consists of a hydrophobic N-terminus fusion domain; two core, highly conserved helical regions that have been solved both by X-ray crystallography [13–15] and by Nuclear Magnetic Resonance (NMR) [16]; and the C-terminus. Understanding the structure of these domains and the different
mechanisms of protein fusion—which eventually result in a six-helix bundle across the host cell membrane with an open channel for the entry of the HIV-1 nucleocapsid—may eventually lead to the development of an anti-HIV vaccine. Unfortunately, neutralizing antibodies so far cannot access the helical core structure [17]. The terminal regions, on the other hand, have proved to be a more susceptible target. For instance, it is known that mutations or deletions in the tryptophan-rich C-terminal region (residues 660–683) can abolish viral fusion [18]. This region overlaps part of the peptide T-20 (gp41_{638–673}), which was the first gp41 fusion inhibitor approved by the US Food and Drug Administration in 2003.

There have been several experimental studies to determine the structure of gp41_{659–671} in aqueous solution. However the original experimental data of the solvated monomeric peptide was highly contradictory. An early investigation based on NMR at pH 7.7 indicated that the conformation of the monomeric peptide in water is an amphiphilic 3_{10} helix with minor random coil representation [18]. A second NMR study found no major population of 3_{10}-helical conformers, but a mixture of various conformers [19]. Later, a UV Resonance Raman Spectroscopy (UVRRS) investigation [20] concluded that there is a rough energy landscape with a greater variety of conformations than found in the NMR studies and previous Circular Dichroism (CD) studies [18]. The authors suggested that gp41_{659–671} exhibits a broad distribution of conformations that includes a significant population of β-turns, as well as 3_{10} helix and π helix motifs but little α helix. Recently, a far-UV CD spectroscopy study [21] also revealed the conformational plasticity of gp41_{659–671}, with no stable α-helical, 3_{10}-helical, or turn motifs. On the other hand, the crystal structure of the peptide bound to the 2F5 antibody shows an extended conformation [22,23].

MD simulation studies also gave a variety of different results [21,24–26]. Some of the discrepancies between these results may be traced back to the use of different force fields and sampling issues. A study based on the GROMOS96 43a1 force field [27,28] found no stable structures in aqueous solution, but an equilibrium of interchanging globular structures, as well as non-negligible populations of bends, turns, and α helices [24]. This force field, however, has been shown to exhibit conformational biases [29]. Another recent MD study used the CHARMM22 [30], AMBER ff03 [31,32] and AMBER ff99SB [33] force fields with the TIP3P water model [25]. In all cases, the authors found far more α-helical content than experimentally observed, with CHARMM22 giving the poorest agreement and ff03 being somewhat closer to experiments. The same group [21] carried out a second
study combining simulations based on CHARMM22 and experimental data. Unfortunately, the MD results agree poorly with the presented CD experimental data. All three force fields were later shown to have significant secondary structure biases. In particular, both CHARMM22 and ff03 over-stabilize helices [34], while ff99SB is believed to under-stabilize helices [34] (although the ff99SB results in Ref. [25] also seem to overestimate the α-helical content with respect to the CD data). Out of the eight force fields employed in the validation study, the best performing fields are CHARMM22∗ and ff99SB∗-ILDN which, respecting the original CHARMM22 and ff99SB force fields, contain modified parameters for backbone and certain side chains. Although force fields have come a long way since their initial development, the fact that they can still show disagreement with experimental data for certain biomolecules is an issue of great concern. The same gp41_{659−671} peptide also highlights some of the challenges of validating theoretical simulation results against experimental data, given that the conclusions of the experimental studies themselves may be in contradiction due to their sources of error, uncontrolled variables, and limitation.

To address some of these issues, in Chapter 2, we test a relatively new AMBER force field, ff12SB [35], and use this to investigate the conformations of gp41_{659−671}. We compare the results against the more extensively tested ff99SB force field, whose performance has been deemed satisfactory by a group of other gp41_{659−671} studies [33,36]. In addition, we also characterize the equilibrium conformations of this peptide as it is pulled away from its equilibrium length in both an implicit and explicit solvent environment.

The use of different methods to characterize the secondary structure motifs can further complicate comparison and interpretation of the data. Secondary structure assignment codes were built to explore the regularities associated with the periodic motifs of proteins, such as in the dihedral angles or in the hydrogen bonds between backbone atoms. The programs map the atomic coordinates of every atom into intuitive visual diagrams [37], and facilitates protein structural comparison and analysis [38–41]. Pauling and Corey predicted the first and most significant structural motifs, namely the α helix [42] and β sheet [43] in 1951. Since then, other periodic motifs have been identified, including 3_{10} helix, π helix, β turn, γ turn, etc. Secondary structures are crucial for methods aiming at the 3-D structures of proteins [44–48]. However, a precise and evidence-based secondary structure assignment is quite challenging because real-life secondary structures are susceptible to bending, twist, fraying and other deformations that can distance them from their ideal geometrical prototypes [47,49–52].
Automatic secondary structure assignment programs were introduced more than three decades ago. These programs attempt to assign secondary structure in a consistent manner, and they are built to explore the regularities associated with periodic structures, such as the dihedral backbone angles ($\phi, \psi$) used for Ramachandran plots, the $\alpha$-Carbon distances, the pattern of hydrogen bonds between backbone atoms, the 3-D geometry of local fragments, the backbone curvature, etc. The first implementation of these methods was introduced by Levitt and Greer in 1977 [53] and was based on distances and dihedral angle profiles of $\alpha$-Carbon atoms over a four-residue sliding window. Other methods include DSSP [54], DEFINE [55], P-CURVE [56], STRIDE [57], P-SEA [58], XTLSSTR [59], SECSTR [60], VoTAP [61], KAKSI [52], PROSIGN [62] and, more recently, a Bayesian method of minimum message length inference, where the protein is described as a collection of segments, with each segment belonging to one of eight potential models (the secondary structure motifs) [63]. Of these, DSSP and STRIDE use quite close criteria for the assignment of secondary structure and are the most widely used methods.

The DSSP (Dictionary of Secondary Structure of Proteins) method, developed by Kabsch and Sander, uses hydrogen bond patterns based on an electrostatic model to assign the secondary structures. The STRIDE (STRuctural IDEntification) method, developed by Frishman and Argos, makes use of both hydrogen-bond patterns and backbone dihedral angles to assign the secondary structures. The more recent KAKSI method [52] uses backbone dihedral angles and $\alpha$-Carbon distances. It was introduced to improve the treatment of irregularities, such as kinks in $\alpha$ helices, in which case STRIDE tends to assign a longer, kinked helix, while KAKSI assigns several short helices.

Although results from codes such as DSSP and STRIDE converge for well-ordered structures [52,58,61,64,65], the agreement between the secondary structure assignments is known to deteriorate as the conformations become more distorted. This is particularly important for unfolded proteins and intrinsically disordered proteins, which are known to exhibit residual and/or transient secondary structure whose characterization is challenging.

In Chapter 3, we carry out MD simulations of (relatively) disordered peptides. Specifically, we consider $\text{gp}41_{659-671}$ (ELLELDKWASLWN), the homopeptide polyasparagine ($N_{18}$), and polyasparagine dimers. We analyze the resulting conformations with DSSP and STRIDE, based on hydrogen-bond patterns (and dihedral angles for STRIDE), and KAKSI, based on $\alpha$-Carbon distances; and carefully characterize the differences in the structural assignments. The full-sequence Segment Overlap (SOV) scores, which quan-
tify the agreement between two secondary structure assignments, vary from 70% for \( \text{gp41}_{659-671} \) (using STRIDE as the reference) to 49% for \( N_{18} \) (using DSSP as the reference). Significant differences are observed in turns, in the distinction between \( \alpha \) and \( 3_{10} \) helices, and in short parallel-sheet segments.

### 1.2 Amyloid protofibril structures of polyQ and polyN

Sequences rich in glutamine (Q) and asparagine (N) are intrinsically disordered in monomeric form [66–68], but can aggregate into highly ordered amyloids [69,70]. Amyloids are fibrillar protein aggregates rich in \( \beta \)-sheet structures that can self-propagate through protein-conformational chain reactions. Pure polyglutamine (polyQ) stretches in humans are associated with at least nine late-onset progressive neurodegenerative diseases, which are caused by polyQ expansions greater than a given threshold in proteins with little or no similarity except for the polyQ region. For instance, in Huntington’s disease, the normal polyQ length in Huntingtin (Htt) is 10-34 repeats, and pathological lengths are 36-120 repeats. Although each disease has a different pathology, they all share some common feature such as the formation of polyQ aggregates [71], and eventual neuronal death. Many structural models have been proposed for polyQ aggregates, such as \( \alpha \)-helical coiled coils [72], \( \beta \) helices [73], \( \alpha \) sheets [74,75] and \( \beta \) sheets [76,77]. Current evidence based on polyQ fibers and crystallites supports the presence of cross-\( \beta \) structures [76–81], common to other amyloid fibrils. PolyQ expansions occur at the DNA level as a consequence of the unstable expansion of CAG codons.

By contrast, prion diseases can have a genetic, infectious, or sporadic origin that involve modification of the prion protein (PrP). This also results in fatal neurodegenerative diseases such as Creutzfeldt-Jakob (CJD) and Straussler-Scheinker diseases of humans, bovine spongiform encephalopathy (BSE), and scrapie in sheep. The cellular prion protein (PrP\(^C\)) is encoded by a gene, and a post-translational modification converts it into the infectious abnormal isoform, (PrP\(^Sc\)), which is rich in \( \beta \)-sheet content [82]. Independently of whether the presence of PrP\(^Sc\) in the cell is the product of a post-translation mishap or whether it is an external infectious agent, PrP\(^Sc\) acts as a template upon which PrP\(^C\) further aggregates thus enabling the non-Mendelian inheritance characteristic of prions. Prions also cause non-Mendelian inheritance in yeast, which has become an imperative model system for the study of prions. Aggregation-prone, Q/N rich domains are present in the baker’s yeast
Saccharomyces cerevisiae, in proteins such as Sup35p and Ure2p, whose amyloid prions are \( \text{[PSI}^+ \text{]} \) and \( \text{[URE}^3 \text{]} \), respectively. Although prions can cause disease in yeast, cells with certain prion forms have been shown to survive better under adverse conditions, which supports the view that at least some prions developed under positive evolutionary selection [83]. In this sense, fungal prions belong to the category of functional amyloids, such as those found for instance in biofilms in bacteria: HET-s fibrils of Podospora anserina, required by heterokaryon incompatibility [84]; the chaplin peptides from the bacterium Streptomyces coelicolor, associated with fimbriae formation [85]; as well as curli fibrils of Escherichia coli and Salmonella enterica with adhesive function [86]. One important research direction in the characterization of prions involves the investigation of the relative roles of the polar, uncharged residues Q and N in Q/N-rich yeast prions [67,87–89]. As an example, a recent study found opposing effects of Q and N in the formation of the Sup35 prion-forming domain (PrD) [89]. The experiments found that if all the Ns in the wild-type Sup35 PrD were switched to Qs, then amyloid formation decreased but toxicity to the cell increased. Conversely, changing all Qs to Ns enhanced benign, self-templating amyloid formation and decreased cell toxicity [89]. The correlation between the lack of amyloid formation and increased toxicity for polyQ sequences is supported by a large body of literature in neurodegenerative polyQ diseases, where the formation of aggregates and toxicity in neuronal cells are not believed to be linked as cause-effect as was once thought [90]. Instead, it is believed that the formation of large aggregates represents a protective cellular response against further toxicity [90–92], and that it is the soluble monomeric or oligomeric intermediates that are toxic to the cell [90–95].

A complete understanding of the mechanisms of polyQ toxicity still eludes us. However, one can address structural, thermodynamic and kinetic matters related to the amyloids themselves. For instance, why would N-rich sequences form amyloid-like prions while Q-rich sequences would not? The results obtained for the Sup35 PrD [89] were unexpected because it was previously thought Qs and Ns are equivalent in terms of prion formation [96,97], and because algorithms for recognizing amyloidogenic sequences neither provide a distinction of the effects of Q and N switching nor predict amyloid formation for the Q/N-rich PrD [89]. Since these experiments do not have an atomic resolution, Monte Carlo simulations were employed in search of an atomistic explanation of these results. The results of the simulations of Q_{30} and N_{30} homopolymers suggested that the N-rich sequences increase the formation of \( \beta \)-hairpin turns and thus of \( \beta \) sheets, which in turn would lead to the
formation of the N-rich amyloids. A more recent study employs several biophysical tools to characterize and compare the aggregation kinetics of N$_{24}$ and Q$_{24}$ peptides [98]. The main finding is that the kinetics of aggregation of N$_{24}$ is much faster than that of the associated Q$_{24}$ counterpart.

In Chapter 4, we study and compare the fibril structural characteristics of Asn and Gln amino acids. We carry out MD simulations that complement and integrate previous experimental and simulation results on N/Q-rich sequences. In particular, we explore all possible steric zipper interfaces for N-rich and Q-rich oligomers to determine which crystallographic class [99] provides the most stability to the N-rich or Q-rich fibril precursor. Our results can be summarized as follows. First, different crystal classes maximize stability in N-rich or Q-rich aggregates. By comparing the most stable (in their optimum class) fibril precursor aggregates, we show that polyQ aggregates are more stable than polyasparagine (polyN) aggregates. Second, the presence of $\beta$-hairpin turns in polyN leading to $\beta$ sheets does not necessarily explain the observed trend of polyN faster aggregation. Since N-rich sequences tend to favor parallel $\beta$ sheets, as opposed to antiparallel ones, and $\beta$-hairpin turns are more stable in polyQ than in polyN. In particular, tight hairpin turns in polyQ have been recently observed experimentally [79,80]. In the previous work in our group, we have measured a non-negligible comparable population of loose turns both in disordered polyQ [68] and in disordered polyN [100], and other simulations have also found polyQ $\beta$ hairpins [101]. Third, enthalpic contributions to the free energy favor the formation of polyQ protofibrils, while entropic contributions favor the formation of polyN protofibrils. Fourth, when other amino acids are present, such as in the Sup35 PrD, their shorter side chains favor steric-zipper formation for N but not Q, as they preclude the in-register association of the long Q side chains.

1.3 DNA/RNA duplex obtained from the GGGGCC and CCCCGG hexanucleotide repeats

Frontotemporal dementia (FTD) and amyotrophic lateral sclerosis (ALS) are two neurodegenerative diseases with similar genetic and neurological pathways. FTD is the most common cause of early-onset dementia due to degeneration of the frontal and anterior temporal lobes, while ALS is characterized by progressive muscle weakness and paralysis due to loss of motor neurons in the brain and the spinal cord. These diseases can occur
simultaneously, and are believed to be part of the same spectrum [102]. A (GGGGCC) hexanucleotide repeat (HR) expansion in the first intron of the C9ORF72 gene has been shown to be the primary cause behind both FTD (here specifically indicated as C9FTD) and ALS [103, 104]. While the impervious population carries fewer than 20 repeats (generally no more than a couple), large expansions greater than 70 repeats and usually encompassing 250-1600 repeats have been found in C9FTD and ALS patients.

Nucleotide repeat disorders can cause toxicity through different, non-exclusive mechanisms. First, the expansions may originate in the DNA itself, and these expansions can alter the local chromatin structure, affecting RNA transcription and protein translation in the gene. Second, transcribed RNA can cause gain and/or loss of function. The transcribed introns containing these large expansions seem to contribute to neuropathology both through loss of function, as mRNA levels of C9ORF72 are decreased in C9FTD/ALS patients [103, 104]; and through gain of function, as RNA transcripts containing the (GGGGCC) HRs are accumulated in nuclear foci in the frontal cortex and spinal cord, leading to the sequestration of RNA-binding proteins [103]. There is evidence that antisense transcripts of the expansion, i.e., (CCCCGG) expanded repeats resulting from the bi-directional transcription of the DNA HR expansions, also form nuclear RNA foci [105, 106]. Third, translated repeats can also cause toxicity in the corresponding protein and its interaction partners. Even though the hexanucleotide expansions reside in a noncoding region of the C9ORF72 gene, it has been shown that these expansions can trigger protein translation in the absence of the start ATG codon, giving rise to the unconventional repeat-associated non-ATG (RAN) translation [106–109]. RAN translation of the (GGGGCC) expansion can lead to Gly-Ala, Gly-Pro, and Gly-Arg poly-dipeptide expansions, while RAN translation of the antisense (CCCCGG) expansion can result in Pro-Gly, Pro-Ala, and Pro-Arg poly-dipeptide expansions. Generically, these are known as “C9RAN” proteins and have been detected in C9FTD/ALS patients.

For simple-sequence nucleotide repeat diseases, a significant breakthrough has been the recognition that stable non-B DNA secondary structure in the expanded repeats is “a common and causative factor for expansion in human disease” [110]. Chemical and enzymatic probing of an r(GGGG GCC) repeat expansion points to a general scenario where the repeat expansion adopts a hairpin structure with G-G mismatches in equilibrium with a quadruplex structure [111]. Recently, CD, optical melting and 1D $^1$H NMR spectroscopy, combined with chemical and enzymatic analysis reveal that r(GGGG GCC) expansions
fold into a G-quadruplex in equilibrium with hairpin structures [112]. The equilibrium is temperature dependent, with $T = 37^\circ$ favoring hairpins and higher annealing temperatures favoring quadruplexes. The equilibrium is also controlled by the type of ion (with K$^+$ ions favoring G-quadruplexes and Na$^+$ ions favoring hairpins), and ionic strength [112].

In Chapter 5, we present an extensive in silico investigation of the structure and dynamics of all possible DNA and RNA duplexes that can be formed from the GGGGCC sense and CCCCCG antisense HRs. First, we carefully generate the possible duplexes that can be obtained by shifting the reading frame in the HRs. This results therefore in three duplexes for each G-rich or C-rich DNA or RNA helical duplex, for a total of twelve duplexes. Then we carry out 1 $\mu$s classical MD simulations to investigate the global conformational space of dodecamers and associated dynamics; characterize the local conformation of the mismatches and the ion distribution and binding; and when possible, compare the relative stability of the double helices.

References


The $\text{gp}41_{659-671}$ HIV-1 Antibody Epitope: A Structurally Challenging Small Peptide


Copyright © 2013 American Chemical Society
The gp41_{659−671} HIV-1 Antibody Epitope: A Structurally Challenging Small Peptide

Yuan Zhang †,‡ and Celeste Sagui*,†,‡

†Department of Physics, North Carolina State University, Raleigh, North Carolina 27695, United States
‡Center for High Performance Simulations (CHiPS), North Carolina State University, Raleigh, North Carolina 27695, United States

ABSTRACT: We report on extensive molecular dynamics (MD) simulations of the tridecapeptide corresponding to residues 659−671 of the envelope glycoprotein gp41 of HIV-1, which spans the 2F5 monoclonal antibody epitope ELDKWA. Previously, X-ray crystallography, nuclear magnetic resonance, and circular dichroism experiments have yielded conflicting conformational information, but there is a growing consensus that the monomeric peptide in aqueous solution is disordered. Here, we use the latest, state-of-the-art AMBER force fields to describe the complex conformational landscape of gp41_{659−671}. We have analyzed the conformational ensembles of the peptide in solution both without applied restraints and under successive tensile restraints. In contrast to previous MD simulations, our results are consistent with the bulk of the experimental findings. The amount of helical population is important in aqueous solution, but this structure forms part of a flexible conformational ensemble with a rugged free energy landscape with shallow minima. Under uniaxial tension, the disordered peptide first becomes fully helical before melting into turns, loops, and 3_{10}-helices. The conformational ensemble includes epitope conformations close to an NMR solution structure (PDB ID 1LCX) as well as epitope conformations close to a very different, extended crystal structure (PDB ID 1TJH).

INTRODUCTION

The recent interest in the structural properties of the peptide gp41_{659−671} (ELLELDKWASLWN) is due to two related issues: (i) as possible target for vaccines as it resides in the membrane proximal region of the HIV-1 fusogenic subunit gp41 and (ii) as a short peptide with an elusive structure that challenges both experiments and simulations.

In the HIV-1 virus, gp41 and gp120 are two transmembrane glycoproteins that mediate the fusion of the virus with the host cell. The structure of gp41 consists of a hydrophobic N-terminus fusion domain; two core, highly conserved helical regions that have been solved both by X-ray crystallography and by nuclear magnetic resonance (NMR), and the C-terminus. Understanding the structure of these domains and the different mechanisms of protein fusion—which eventually result in a six-helix bundle across the host cell membrane with an open channel for the entry of the HIV-1 nucleocapsid—can eventually lead to vaccines. Unfortunately, neutralizing antibodies so far cannot access the helical core structure. The terminal regions, on the other hand, have proved to be a more susceptible target. For instance, it is known that mutations or deletions in the tryptophan-rich C-terminal region (residues 660–683) can abolish viral fusion. In addition, this region overlaps part of the peptide T-20 (gp41_{658−673}), which was the first gp41 fusion inhibitor approved by the US Food and Drug Administration in 2003.

A vaccine can also be based on immunogens containing conserved epitopes for monoclonal antibodies in a way that mimes their structure on the native gp41 and gp120 glycoproteins. In particular, the gp41_{659−671} sequence (which is included in T-20) contains the sequence ELDKWA which is the core epitope for the HIV-1 neutralizing monoclonal antibody 2F5. The 2F5 antibody reacts strongly with peptides that contain the ELDKWA sequence; however, when this sequence is incorporated in various immunogens, these immunogens fail to induce antibodies that react with the native form of the HIV-1 envelope glycoprotein complex. Experimental findings show that all residues in gp41_{659−671} are important to induce antibodies that act like 2F5. Experiments also show that 2F5 only binds to the prefusogenic form of gp41 before gp120 binds to the T-cell membrane) and suggest that the gp41_{659−671} epitope is solvent exposed in this situation. Thus, it is desirable to know the structure of the gp41_{659−671} epitope in its solvated prefusogenic state.

There have been several experimental and simulation studies to determine the structure of gp41_{659−671} in aqueous solution. An early study based on NMR at pH 7.7 indicated that the conformation of the monomeric peptide in water is an amphiphilic 3_{10}-helix with minor random coil representation. A second NMR study found no major population of 3_{10}-helical conformers, but a mixture of various conformers. Later, a UV resonance Raman spectroscopy (UVRRS) investigation investigation.
concluded that there is a rough energy landscape with a wider variety of conformations than found in the NMR studies and previous circular dichroism (CD) studies. The authors suggested that gp41\,659-671 exhibits a broad distribution of conformations that includes significant population of \( \beta \)-turns as well as \( 3_{10} \)-helix and \( \alpha \)-helix motifs but little \( \alpha \)-helix. Recently, a far-UV CD spectroscopy study\(^{17} \) also revealed the conformational plasticity of gp41\,659-671 with no stable \( \alpha \)-helical, \( 3_{10} \)-helical, or turn motifs. On the other hand, the crystal structure of the peptide bound to the 2F5 antibody shows an extended conformation.\(^{17,14} \) Figure 1 shows sample structures obtained experimentally: PDB ID 1TJH shows the crystal structure of an \( 11 \)-mer peptide from the 1LCX NMR solution structure.\(^{15} \) The flanking residues E662 and A667 of the epitope ELDKWA are shown. The color scheme for the secondary structure (generated by VMD\(^{16} \)) is as follows: \( 3_{10} \)-helix: blue; turn: cyan; coil: white.

Figure 1. Conformations of the gp41 \( 11 \)-mer peptide (gp41\,659-671 minus the terminal residues, i.e., LLELDKWASLW) from the 1TJH crystal structure\(^{13} \) and the gp41\,659-671, \( 13 \)-mer peptide from the 1LCX NMR solution structure.\(^{15} \) The flanking residues E662 and A667 of the epitope ELDKWA are shown. The color scheme for the secondary structure (generated by VMD\(^{16} \)) is as follows: \( 3_{10} \)-helix: blue; turn: cyan; coil: white.

Although force fields have come a long way since their initial development, the fact that they can still show disagreement with experimental data for certain biomolecules is an issue of great concern. The previous work showed how a small peptide such as gp41\,659-671 gives rise to rather different results depending on the choice of force field and that these results in general do not agree with experiments. Yet, the same gp41\,659-671 peptide also highlights some of the challenges of validating against experimental data, given that the conclusions of the experimental studies themselves can be quite contradictory, too, since of course all experiments have their own sources of errors and uncontrolled variables.

In this work, we test how a brand new AMBER force field, mainly ff12SB,\(^{25} \) performs in reproducing the conformations of gp41\,659-671. We compare it against the more extensively tested ff99SB, whose performance has been deemed satisfactory by a number of studies\(^{22,24} \) and for which there are already MD results for gp41\,659-671.\(^{23} \) However, it has also been recognized that ff99SB understabilize helical conformations of transiently folded peptides.\(^{24} \) In order to improve the accuracy of secondary structure propensities, the backbone angle \( \phi \) has been modified in ff12SB. Since residues such as isoleucine, leucine, aspartate, asparagine, and others were shown to sample conformations differently from experiments,\(^{24,25} \) the \( \gamma \)-helical and \( \delta \)-turn angles of the side chains in ff12SB have been refined by fitting energy profiles to match ab initio quantum data. Similar side-chain dihedral corrections have been extended to lysine, arginine, glutamate, glutamine, methionine, serine, threonine, valine, tryptophan, cysteine, phenylalanine, tyrosine, and histidine (i.e., all the side chains but Gly, Ala, and Pro), and thus ff12SB is expected to enhance the reproduction of experimentally indicated geometries over ff99SB, although more extensive testing is still needed. In addition to sampling equilibrium configurations, we compare the conformational landscape as the peptide is pulled away from its equilibrium length, and we compare results for both implicit and explicit solvent. Given the dispersion of results in the NMR and CD studies, we hope to motivate single-molecule experiments such as Förster resonance energy transfer (FRET) and single-molecule pulling experiments, which may throw additional information about the conformation of this peptide.

### COMPUTATIONAL METHODS

To investigate the conformational phase space of gp41\,659-671, we have carried out several simulations with the AMBER ff99SB and ff12SB versions of the Cornell et al. force field,\(^{22} \) using both implicit and explicit solvent for each force field. These simulations explore the conformational landscape of the peptide at equilibrium using regular MD. We also use equilibrium umbrella sampling to probe the conformations as the peptide is restrained to different end-to-end distances. Each simulation includes one gp41\,659-671 molecule with sequence Ace-ELLELDKWASLWN-NH\(_2\). The N-acetylated and C-amidated forms are used to avoid excessive interactions between the charged N- and C-terminus in the implicit solvent. These capping groups are generally assumed to make the peptide more similar to the corresponding segment in the native protein. Experimental comparisons of gp41\,659-671 in its zwitterionic and capped forms show essentially the same diffusion constant and the same conformations with a small increase of the helical population in the capped form.\(^{6} \) Simulations were carried out at room temperature (300 K) via Langevin dynamics with a coupling parameter \( \gamma = 1.0 \text{ ps}^{-1} \).
The peptide has two net negative charges that were neutralized by the addition of two Na\textsuperscript{+} ions. Coordinates were sampled every 2 ps for regular MD simulation and 10 ps for the umbrella sampling runs.

**Implicit Solvent Simulations.** Topology and parameter files along with the coordinates corresponding to the unfolded peptide were generated via the LEAP program of the AMBER v.11\textsuperscript{27} and v.12\textsuperscript{28} simulation packages. The implicit water model is based on the generalized Born approximation (GB)\textsuperscript{29} including the surface area contributions using the LCPO model\textsuperscript{30} (GB/SA) with the surface tension set to 0.005 kcal/ (mol Å\textsuperscript{2}). For the GB model, we used the GB	extsuperscript{99}II model,\textsuperscript{31} with a cutoff of 25 Å. With this model, the effective Born radii are rescaled to account for interstitial spaces between atom spheres missed by the GB\textsuperscript{HCT} approximation.

After a short minimization, an initial 10 ns MD simulation was used to generate equilibrium conformations of the peptide. Five different conformations from this run were chosen as initial conformations for five MD runs (with different random number generator seeds), each 40 ns long. One conformation from these sets was then selected as initial conformation for another 100 ns simulation. The 100 ns set and the second half of five 40 ns sets were used for data analysis of equilibrium conformations. Thus, the simulation was run for a cumulative time of 300 ns, and data are presented for 200 ns. Figure 2 shows the five initial conformations for the 40 ns MD simulations for ff99SB. The same procedure was repeated for ff99SB.

![Figure 2](image.png)

**Figure 2.** Snapshot of chosen initial conformations for the unrestrained MD simulations under the ff99SB force field. The top of the peptide corresponds to E659 and the bottom to W671. These conformations all evolved from a fully extended initial structure.

We then used equilibrium umbrella sampling to "pull" the peptide through different end-to-end distances by applying a harmonic restraint to the collective variable. The end-to-end distance collective variable is defined as the distance between the peptide-bond nitrogen atoms in E659 and in N671. The end-to-end distance \(d\) was restrained with a harmonic potential \(V_{\text{col}} = (1/2)[(d - d_0)^2]\) in the range 10–40 Å through 2 Å windows (i.e., \(d_0 = 10, 12, 14, \ldots, 36\) Å). Then, for each of the configurations centered at 10 Å, 12 Å, etc., we “equilibrated” the system by using a harmonic restraint of 10 kcal/(mol Å\textsuperscript{2}) for 1 ns and then 1 kcal/(mol Å\textsuperscript{2}) for 1 ns. The conformations at this point were taken as initial conformations for 100 ns runs with a harmonic restraint of 1 kcal/(mol Å\textsuperscript{2}), which were finally used for data analysis.

**Explicit Solvent Simulations.** The TIP3P water model\textsuperscript{32} was used for the explicit solvent simulations under periodic boundary conditions. Five initial conformations from the implicit solvent simulations for each force field were solvated in a rectangular box, with an average number of waters of approximately 5670. Electrostatics were handled by the PME method,\textsuperscript{33,34} with a direct space cutoff of 9 Å and an average mesh size of approximately 1 Å for the lattice calculations. The equilibration process took place in four steps. First, we applied steepest descent followed by conjugate gradient minimization keeping the peptide atoms fixed at their initial positions. Then we carried out unrestrained steepest descent followed by conjugate gradient minimizations. This was followed by short MD runs under constant volume while the system was heated from 0 to 300 K with weak restraints on the peptide atoms. Finally, the system was kept at 300 K via Langevin dynamics with a collision frequency \(\gamma = 1.0\) ps\textsuperscript{−1} and at constant pressure (1 atm) via the Berendsen barostat\textsuperscript{35} with the isothermal compressibility \(\beta = 44.6 \times 10^{-5}\) bar\textsuperscript{−1} and the pressure relaxation time \(\tau_p = 1.0\) ps. These NPT Langevin dynamics simulations were carried out for 2 ns, and the density of the system was found to be stable around 1.0 g/cm\textsuperscript{3}. The two sets of five equilibrated conformations for each force field provided the initial conformations for NPT simulations with Langevin dynamics, each simulation 40 ns long. One configuration for each force field was selected for a further 100 ns run. The second half of each of the 40 ns runs and the 100 ns runs were used for data analysis.

For the umbrella sampling simulations, we also used the NPT ensemble as before, and the collective variable was again selected as the end-to-end distance, with conformations restrained around 14 windows separated by 2 Å, from 10 to 36 Å. The system was pulled through the different end-to-end distances and equilibrated in the same way as in the implicit solvent case. The bond force constant was set at 1.0 kcal/(mol Å\textsuperscript{2}), and the equilibrium bond length was set at 10, 12, 14, ..., 36 Å.

**RESULTS**

Figure 3 shows the free energy profiles as a function of the end-to-end distance for the two force fields in implicit and explicit solvents. To first order, the two force fields provide similar results, with the equilibrium length at 17.92, 17.75, 16, and 17.18 Å under implicit ff99SB, explicit ff99SB, implicit ff12SB, and explicit ff12SB. The equilibrium end-to-end distances are in good agreement with each other except for explicit ff12SB, which seems to yield a slightly more compact structure. A general feature for these free energies is that for distances greater than the equilibrium length, the explicit solvent lowers the free energy. This can be easily explained, as in the absence of waters, the peptide tends to form hydrogen bonds with itself, making elongated conformations more costly than in the presence of explicit waters, which can replace the internal hydrogen bonds.

Figure 4 shows the Ramachandran plots for each residue under the two force fields and implicit and explicit solvents. For most residues, in addition to \(\phi, \psi\), there is an important presence...
of β and PPII populations. There is better agreement between the implicit and explicit solvent representations for f12SB than for f99SB. In the latter, residues Glu-659 to Leu-663 and Leu-669 and Trp-670 show slightly larger β and PPII populations in explicit f99SB than in implicit f99SB (indicating less α-helical content in the explicit solvent). On the other hand, for both force fields Trp-666 shows a considerably large PPII population in implicit solvents that is absent in the explicit solvents. Although certain residues in the implicit solvent representation have populations in the αi region of the Ramachandran plot that are absent in the explicit plots, the αi region seems to be consistently present in Leu-660 and, especially, Asn-671 (which is systematically characterized as “loop” in Figures 5 and 6).

The secondary structure motifs were characterized by using the program STRIDE, since it is the code used in a couple of other simulations of gp41α−−−γ71, and therefore comparisons are more straightforward. Figures 5 and 6 show the secondary structure population for each residue (y-axis) as a function of “sampling window”: strictly speaking, it is not completely as a function of time because the first 100 ns are obtained as aggregates of the last 20 ns in each of the five independent 40 ns runs; the period between 100 and 200 ns corresponds to the time evolution of another independent run (each simulation was therefore “equilibrated” for over 20 ns, and this data is not shown). We see that the second half of the plot constructed in this way is qualitatively similar to the first half, suggesting that the system has undergone ergodic sampling. The average secondary structure for each residue can be appreciated at a glance in Figure 7. The major differences between f99SB and f12SB occur in the N-terminal ELLEL sequence. This is due to both corrections in residues L and E, which was found to be very helical in f99SB. Thus, the α-helical content of ELLEL is 51% in implicit f99SB while only 11% in implicit f12SB, and it is 31% in explicit f99SB while only 19% in f12SB. For the individual E’s in ELLEL, the differences in α-helical content of E659 and E662 amounts to 43% for implicit f99SB while only 8% for implicit f12SB and 24% for explicit f99SB while 18% for explicit f12SB.

For each force field, the maps between the implicit and the explicit solvent are in qualitative agreement. If one considers the implicit and explicit solvent populations for a given type of secondary structure for each residue, and then sums the absolute value of their differences, then it is possible to obtain some sort of measure of how much the implicit and explicit solvent representations in Figures 5 and 6 differ (notice that there is emphasis on absolute differences of individual residues; in the averages presented in Table 1 positive differences at a given residue can cancel out negative differences at another residue). Overall, the differences between implicit and explicit f99SB are larger than the differences between implicit and explicit f12SB. In other words, there seems to be a better agreement between both solvents in f12SB than in f99SB, a result that is more pronounced if DSSP57,58 is used for the secondary structure analysis. The major differences between implicit and explicit f99SB occur in the population of α-helices, while the major differences between implicit and explicit f12SB occur in the population of turns. Although α structures have been reported before,11 we find that the population of α-helix is negligible, in agreement with more recent measurements.12,23

The amount of 3_10-helices and turn helices is also important. In particular, 3_10-helices, which have been reported in a previous NMR study,6 vary between 5% for implicit f99SB and 13% for explicit f12SB (the analysis was repeated with DSSP, which yields 13% and 14% for 3_10-helices in implicit and explicit f99SB and 19% and 21% in implicit and explicit f12SB). In other words, for each force field 3_10-helices tend to show up more in explicit than implicit solvents, and explicit f12SB exhibits more of them than explicit f99SB. Turns comprise 28% of the total populations of both implicit and explicit f99SB and 41% and 31% of the total populations of implicit and explicit f12SB. The third most common structural "motif" is "loops", i.e., disordered structures which comprise 24% and 30% of the total populations of implicit and explicit f99SB and 20% and 22% of the total populations of implicit and explicit f12SB. The average populations for the whole peptide and for the epitope are shown in Table 1.

Figure 8 shows the distribution of conformations of the ELDKWA epitope based on the root-mean-square deviation (RMSD) values of the backbone atoms N, O, C’, and Cα with respect to a solution NMR structure (ILCX)6 of gp41α−−−γ71 and with respect to the crystal structure of a heptamer ELDKWAS in complex with 2F5 (1TJG).13 The 1LCX solution NMR structure is the one reported by Biron et al.,6 which is a 3_10-helix from E662 to W670 according to DSSP and from D664 to L669 according to STRIDE (as shown in Figure 1). The crystal structure of the heptamer shows the same structure as 1TJH for the 11-mer (Figure 1) and 1TJ1 for the 17-mer, and all correspond to the conformations found by Ofek et al.13 When the structure of gp41α−−−γ71 is analyzed as part of the complex in 1TJH, then DSSP and STRIDE predict a β-strand on residues E662 and L663, followed by a turn, followed by disorder. If we try the isolated gp41α−−−γ71 then the secondary structure is PPII for E662 and L663 and a type 1 β-turn for residues D664—A667. The simulations results in Figure 8 for both explicit f99SB and explicit f12SB show an important
population presence in the bottom quadrants, which partly correspond to the solution NMR structure, with RMSDs between 1 Å (even less for ff99SB) and 2 Å. There is another important contribution in the upper-right quadrant, for conformations that are similar neither to the crystal structure nor to the NMR solution structure and that are mainly disordered. In addition, both force fields show a contribution to the lower-left quadrant, closer to the crystal structure. The implicit solvent representations (not shown) tend to increase the population in the upper-right quadrant and to decrease the rare, relatively extended conformations in the lower-left quadrant that have smaller RMSDs with respect to 1TJG.

Another comparison to experiment can be carried out through NMR chemical shifts. Figure 9 gives the distribution function of conformations versus the RMSD of experimental ($\delta_{\text{exp}}$) and predicted ($\delta_{\text{pred}}$) chemical shifts computed via SHIFTX2 for HN and HA, according to

$$\text{RMSD} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (\delta_{\text{exp}}^i - \delta_{\text{pred}}^i)^2}$$

where $N$ is the number of residues of each conformation. The experimental data are provided in ref 6 and correspond to the structure 1LCX. We notice that the RMSD for the algorithm itself as implemented in SHIFTX2 is 0.2351 ppm for HN and 0.1081 ppm for HA. This means that in addition to discrepancies between simulations and experiment, there are extra sources of errors due to the accuracy of the chemical shift predicting algorithm. Thus, if we run the 1LCX structure itself through SHIFTX2, the computed RMSDs are 0.3425 ppm for HN and 0.2073 ppm for HA (these are indicated as a red bar in the graphic). Again, although both explicit force fields are similar, ff12SB seems to perform slightly better than ff99SB.

Figure 4. Ramachandran plots for each residue in ELLELDKWASLN for the unrestrained MD simulations. Results for the AMBER ff99SB are presented in blue and for ff12SB in red/yellow. For each force field, the top two rows show the plots for implicit solvent, while the bottom two rows show the plots for the explicit solvent.
Figures 10 and 11 show the average secondary structure population for each residue (y-axis) as a function of the end-to-end distance (in Å) used in the restraint for the umbrella sampling runs. Each window represents an average over 100 ns. For ff99SB, the implicit solvent structures show more α-helical content than the corresponding explicit solvent ones, a trend that was already present in the unrestrained simulations. There is more consistent agreement between implicit and explicit solvents under the ff12SB force field. If one chooses a middle residue in the implicit graphs, and follows it along the row as the peptide is stretched, one can see a relatively smooth behavior, where the α-helical content (red) increases from its initial value at 10 Å until it reaches a maximum at approximately 20 Å and then it decreases accompanied by an increase of 310-helices, turns, and disordered extended structures. This behavior is considerably less "regular" in the explicit solvents (especially in ff99SB, where no clear trend is discernible). In particular, the number of α-helical structures decreases dramatically after 22 Å, giving rise to many more 3_10-helices, turns, and disordered extended structures, in agreement with the lowering in free energies in Figure 3 for the explicit solvents with respect to the implicit solvents.

Finally, Figure 12 shows a similar plot to Figure 8, i.e., the distribution of conformations of the ELDKWA epitope based on STRIDE.
on the root-mean-square deviations (RMSD) values of the backbone atoms N, O, C', and Cα with respect to the solution NMR structure (1LCX) of gp41659−671 and with respect to the crystal structure 1TJG, except that the simulation coordinates belong to the umbrella sampling runs with end-to-end distance of 28 Å. The net effect of stretching the polymer in explicit ff99SB (top panel) results in approximately the same spread in RMSDs with respect to 1TJG but augments the population with conformations closer to 1LCX. On the other hand, the effect of stretching the polymer in explicit ff12SB is slightly opposite: it gives approximately the same spread in RMSDs with respect to 1LCX but increases the spread of RMSDs with respect to 1TJG that extend from 0.5 Å to more than 3.5 Å. Remarkably, there is an important population contribution to the left quadrants, i.e., close to the extended crystal structure in 1TJG, including a small population of RMSDs of approximately 0.5 Å. Implicit solvent representations (not shown) suppress the extended conformations and are closer to 1LCX.

### DISCUSSION

#### Comparison to Previously Published Experimental Data.

The NMR study by Biron et al.6 of the peptide in aqueous solution at pH 7.7 found a monomeric, amphiphilic 310-helix with minor random coil representation6 (Figure 1), a finding that seemed to be confirmed by CD measurements carried out by the same group. In particular, the exposed surface of the helix was found to consist of conserved, hydrophobic HIV-1 residues. The reconstructed, energy-minimized structure found a 310-helix conformation with 3.2 residues per turn. This analysis was carried out for the zwitterionic form of the peptide, but the authors observed additional helical stabilization with acetylation and amidation of the peptide. However, a second NMR study and a CD spectrum analysis both carried out by Barbato et al.10 found no major population of 310-helical conformers, but a mixture of various conformers. These authors showed that all the 13 residues in gp41659−671 were important for binding the human monoclonal antibody 2F5, as the flanking residues induce local structure (rather than providing additional contacts to 2F5). The authors also proposed that gp41 exhibits a "conformational plasticity" where the final configuration depends on the microenvironment. Thus, the peptide could undergo a conformational transition between disordered and helical structures as it approaches the lipid membrane, a fact that was later further explored and confirmed by Gregor et al.23 Close to the membrane, in membrane-mimetic environments, or in conformations that involve dimers or trimers, the X-ray

<table>
<thead>
<tr>
<th>Peptide</th>
<th>Simulation Model</th>
<th>α-helix</th>
<th>3_10-helix</th>
<th>Other</th>
<th>Turn</th>
</tr>
</thead>
<tbody>
<tr>
<td>gp41659−671</td>
<td>ff99SB-impc</td>
<td>43</td>
<td>5</td>
<td>28</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td>ff99SB-exp</td>
<td>33</td>
<td>9</td>
<td>28</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>ff12SB-impc</td>
<td>31</td>
<td>8</td>
<td>41</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>ff12SB-exp</td>
<td>33</td>
<td>13</td>
<td>31</td>
<td>22</td>
</tr>
</tbody>
</table>

**Table 1. Secondary Structure Population Based on STRIDE for gp41659−671 and the Epitope ELDKWA**
and NMR structures of gp41659−671 are found to be mainly helical (see for instance PDB IDs 1ENV,2 2PV6,40 1JAV,41 3G9R,42 3H01,43 and 2X7R44).

A UVRRS investigation11 revealed a rough energy landscape with a wide variety of conformations. In particular, an examination of the conformation-sensitive AmIII3 region suggested that gp41 659−671 exhibits a broad distribution of conformations that includes significant population of β-turns as well as 3 10-helix and π-helix motifs but little α-helix. The authors of this work also found a lack of temperature-induced change (between 1 and 30 °C) in the AmIII3 band intensities, a fact that they interpreted as indicative of an energy landscape where folded and unfolded conformations occupy broad, degenerate free energy minima. Approximately half of the peptides were found in folded conformations (predominantly 310- and π-helices) and the other half in unfolded (β-turn/PPII) conformations. However, the assignation of secondary structure to this type of spectra is not straightforward and not unequivocal.11,12 Recently, a far-UV CD spectroscopy study carried out by Tulip et al.12 also revealed the conformational plasticity of gp41659−671, which exhibited different competing folding propensities and no stable α-helical, 310-helical, or turn motifs. The spectrum of gp41659−671 at 5 °C in various aqueous conditions showed a strong negative band just before 200 nm and a negative shoulder at approximately 225 nm, which are indicative of disordered polypeptides with residues in the αR and β regions of the Ramachandran plot.

Our simulations are consistent with the bulk of these experimental findings. First, the amount of helical population is important in aqueous solution, but this structure forms part of a flexible conformational ensemble with a rugged free energy, with shallow minima, as reported by Barbato et al.10 and Tulip et al.12 This is strongly supported by the Ramachandran plots in Figure 4. The analysis of secondary structure, which depends partially on the criteria used, shows that the 310-helix is present but in small amounts: for explicit ff12SB its population reaches 13% under STRIDE or 21% under DSSP. For all cases, we find negligible π-conformation, which is in agreement with the experimental findings.
knowledge that $\pi$-helices are seldom found to be stable experimentally. Mainly, the simulations results exhibit a large presence of $\alpha$-helical populations, turns, and loops, as clearly shown in Figures 5 and 6. Still, the explicit solvent conformations show important populations whose epitope is relatively close to the NMR structure LCLX of Biron et al. as well as an ensemble of conformations comprising other structures (Figure 8). The relatively extended conformation of Olek et al. seems to be under-represented in the regular MD simulations in Figure 8 (left quadrants). However, in the crystal structure gp41$_{659-671}$ is bound to 2F5, and this clearly imposes restraints in the conformational space. For instance, in a validation study of different force fields against experimental data, CHARMM22 fails to maintain the structure of protein GB3 (which unfolds during the simulation), and both CHARMM22 and ff99SB in general overstabilize helices. On the other hand, ff99SB is believed to underestimate helices.

The simulation results by Gregor et al. disagree considerably with the experimental data, especially in the case of ff99SB in explicit water, which greatly overestimates the $\alpha$-helical content with respect to the CD data. As we are particularly interested on how well the AMBER force fields perform on disordered peptides and intrinsically disordered proteins, we carried out our simulations specifically comparing the two most recent force fields with the TIP3P water model. In particular, we found that the ff99SB force field with the TIP3P water model produces different results than those reported by Gregor et al. This can be appreciated by comparing the bottom panel in Figure 5 and the bottom panel of Figure 4 in ref 23 (the secondary structures in both figures have been computed with STRIDE). Except for the terminal residues, the results in ref 23 show a much higher $\alpha$-helical content than our results. One
obvious difference in the simulations is that the previous authors employed the zwitterionic form of the peptide, while we are using capped ends. This, however, favors the opposite trend: the capped peptide favors the $\alpha$-helical conformation more than in its zwitterionic version. Given that we are using the same force field, the same water model, the same treatment of electrostatics, and the same program for the analysis of the secondary structure, we speculate that the main sources of disagreement can be attributed to the size of the simulation box and/or to the amount of sampling. With respect to the system size, we notice that Gregor et al. used only 33% of the numbers of waters that we used. Given that the peptide is disordered and in principle it can adopt a fully unfolded conformation, it is important to pad the box with a large number of waters to avoid periodic artifacts which could indeed stabilize a helical conformation (especially in a zwitterionic form): we chose the minimum number of waters that seemed to be "safe". A definite answer to what the minimum number of waters for this system would probably require a careful finite effect (i.e., periodic artifact) study. With respect to the amount of sampling, we notice that the simulations in ref 23 only extend to 10ns. Although we show only 200 ns of data, we actually carried out each simulation for 300 ns (see computational details). Thus, the extensive $\alpha$-helical conformations in ref 23 could simply be due to lack of sampling. For example, the first 20 ns window for explicit ff99SB in our Figure 5 (which corresponds to the last 20 ns in the run) contain a fit mostly to $\alpha$-helical conformations and qualitatively resemble Figure 4 in ref 23. However, later there are also extended periods where "turns" are more stable (these were found unstable in ref 23). All in all, our simulation data seem to be more consistent with the experimental results and with the concept of a rugged free energy with multiple conformations.

Comparison between the Force Fields. We are interested in how well the AMBER force fields can describe disordered peptides and intrinsically disordered proteins, whose satisfactory description is more challenging. Thus, we compared a brand new AMBER force field, mainly ff12SB\(^{56}\) against the more extensively tested ff99SB, whose performance has been deemed satisfactory by a number of studies\^{22,23,46,47} but which has also been found to suffer a number of inaccuracies in backbone and side-chain dihedral angles.\^{3,42} Basically, ff12SB builds upon ff99SB with modified parameters for the backbone $\phi$ torsion angle and for all side-chain torsions, except Gly, Ala, and Pro.

To first order, the force fields are in good agreement. The implicit and explicit simulations for each force field yield essentially the same free energy for conformations near equilibrium and for the compact structures. However, as the end-to-end distance grows, the polymer in explicit solvent tends to make more hydrogen bonds with waters, and therefore it is easier to stretch than the polymer in implicit solvent, which has more intramolecular hydrogen bonds. Other general trends for gp41\(_{559-671}\) valid independently of whether one computes the secondary structure with STRIDE or with DSSP, are the following. The implicit ff99SB gives the largest number of $\alpha$-helical conformations, while explicit ff99SB and both implicit and explicit ff12SB are comparable in this respect. The percentage of $3_\beta$-helices is larger for explicit ff12SB, and the percentage of loop structures is larger for explicit ff99SB. The major differences between the two force fields are in N-terminal ELLEL sequence, where refinements in the side-chain torsions of Glu and Leu decrease dramatically the $\alpha$-helical content of ELLEL in ff12SB.

Force fields are parametrized near equilibrium conformations, and comparisons far from equilibrium present a more stringent test of their performance. Here, we stretched the polymer away from its equilibrium end-to-end distance. The differences between the two force fields as reflected in Figures 10 and 11 are in general larger than the differences near equilibrium as shown in Figure 7. Implicit ff99SB always shows more helical content than the explicit version, while ff12SB shows better agreement between implicit and explicit solvents. Unfortunately, in spite of the measurable differences that we have described, we cannot unequivocally discern from the simulations which force field is better. The stretch simulations perhaps favor ff12SB as the better field, as there is least disagreement between the implicit and explicit solvent representations and as ff12SB encompasses in it conformational ensemble a larger population of turns, including those structures—when the polymer is stretched—that approach the crystallographic structure.

Conformational Equilibrium in Stretched Conformations. To study the mechanical properties of polypeptides, it is customary to carry out single-molecule stretching experiments using, for instance, atomic force microscopy or optical tweezers. The results are generally shown as a plot of the tensile force as a function of the end-to-end distance of the polymer. Computationally, this can be easily achieved, e.g., by using steered molecular dynamics (SMD). In this work, we have used umbrella sampling, so instead of pulling the peptide continuously, we take pictures of the equilibrium conformations under different (rather weak) end-to-end distance constraints.

First, the equilibrium, unbiased free energy is similar to that obtained in other simulation approaches, such as in the case of alanine-rich peptides explored with adaptively biasing techniques.\^{48} The stretched conformations in the implicit simulations (top panels in Figures 10 and 11) reveal a smooth behavior with stretching: for each residue, the $\alpha$-helix population along a row (as the conformations is stretched) increases from its initial value until it reaches a maximum in the window centered at 20 Å. After that it decreases with a concomitant increase in $3_\beta$-helices, turns, and disordered extended structures, which is qualitatively similar to the results in ref 48. An analogous behavior can also be observed in explicit ff12SB, although the $\alpha$-helical content decreases abruptly after 22 Å, giving rise to many more $3_\beta$-helices, turns, and disordered extended structures than the implicit solvents, in agreement with the free energy maps in Figure 3.

However, one main difference between our work and most work on stretching helical peptides is precisely given by the fact that at the free energy minimum gp41\(_{559-671}\) is not a helical but a disordered peptide (with some helical content), and its $\alpha$-helical content increases to its maximum on stretching, before it starts decreasing again. This agrees with the theoretical prediction that a partially helical chain should increase its helicity upon tensile stretching by an applied force due to the loss of the conformational entropy of the disordered state under uniaxial tension. Also, this situation however has not been verified experimentally at the single molecular level due to the difficulties associated with determining the secondary structure of a single polypeptide chain,\^{3,54} although it has been confirmed at the macroscopic level.\^{3,54}
of helical content provides an indirect confirmation of the theoretical argument proposed by Chakrabarti and Levine\(^5,6\) that the tensile "melting" of a partially helical chain will be determined by the stretching mode of the locally denatured parts of the polymer. This local denaturation can take place spontaneously due to local energy fluctuations. In the explicit solvent the carbonyl of the backbones can form hydrogen bonds with the waters,\(^7\) providing an additional source of fluctuations over the implicit solvent and the consequent easier melting of the helical chain in explicit waters.

**CONCLUSION**

In this work, we have studied the conformational ensembles of gp41\(^{659-671}\), under the most recent AMBER force fields\(^8\) and \(^9\) in both implicit and explicit waters via MD for a cumulative time longer than 7.2 \(\mu\)s. Intrinsic disordered peptides are hard to properly characterize, but our results are consistent with the bulk of the experimental findings. The amount of helical population is important in aqueous solution, but this structure forms part of a flexible conformational ensemble with a rugged free energy landscape with shallow minima. The newest force field \(^11\) seems to perform slightly better than \(^9\), especially under restraints that take the peptide away from its equilibrium end-to-end distance. When the originally disordered peptide at the equilibrium end-to-end distance corresponding to the free energy minimum is subject to uniaxial stretching, it first becomes almost fully helical due to the loss of the conformational entropy of the disordered state under uniaxial tension. Upon further stretching, the helix melts into turns, loops, and \(3\beta\)-helices. The explicit waters provide additional hydrogen bonds that facilitate the "melting" of the helical structure and the lowering of the free energy when compared with the implicit solvent. Important populations in the conformational ensemble show epitope conformations that are close to the NMR aqueous solution structure (1LCX) of Biron et al.\(^6\) as well as epitope conformations close to the very different crystal structure of the peptide bound to 2FS (1TJH) of Ofek et al.\(^13\). The latter agreement is obtained under a weak end-to-end distance restraint that mildly mimics one of the several effects of the binding to 2FS.

**AUTHOR INFORMATION**

**Corresponding Author**

*E-mail sagui@ncsu.edu, Ph (919) 515-3111 (C.S.).

**Notes**

The authors declare no competing financial interest.

**ACKNOWLEDGMENTS**

We gratefully acknowledge the support of the National Science Foundation (Grants 1021883 and 1148144). We also thank the NC State HPC Center for extensive computational support.

**REFERENCES**


(42) Liu, J.; Dey, Y.; Li, Q.; Dey, A. K.; Moore, J. P.; Lu, M. Role of a Putative gp41 Dimerization Domain in Human Immunodeficiency Virus Type 1 Membrane Fusion. J. Virol. 2010, 84, 201−209.


Secondary structure assignment for conformationally irregular peptides: Comparison between DSSP, STRIDE and KAKSI


Copyright © 2014 Elsevier Inc. All rights reserved.
Secondary structure assignment for conformationally irregular peptides: Comparison between DSSP, STRIDE and KAKSI

Yuan Zhang a,b, Celeste Sagui a,b,∗

a Department of Physics, North Carolina State University, Raleigh, NC 27695, United States
b Center for High Performance Simulations (CHiPS), North Carolina State University, Raleigh, NC 27695, United States

1. Introduction

Secondary structure allows for a simple description of the complex three-dimensional structure of proteins. It maps the atomic coordinates for every atom into intuitive visual diagrams [1], and facilitates protein structural comparison and analysis [2–5]. The first and most important structural motifs, the α-helix [6] and β sheet [7] were predicted by Pauling and Corey in 1951. Since then, other periodic motifs have been identified, including 3_{10} helix, π helix, β turn, γ turn, etc. Secondary structures are extremely important for methods aiming at predicting protein three-dimensional structures [8–12]. However, a precise and evidence-based secondary structure assignment is quite challenging due to the fact that real-life secondary structures are susceptible to bending, twist, fraying and other deformations that can distance them from their ideal geometrical prototypes [13–16,11].

Automatic secondary structure assignment programs were introduced starting more than three decades ago. These programs try to assign secondary structure in a consistent manner, and they are built to explore the regularities associated with periodic structures, such as the dihedral backbone angles (ϕ, ψ) used for Ramachandran plots, the α-Carbon distances, the pattern of hydrogen bonds between backbone atoms, the three-dimensional geometry of local fragments, the backbone curvature, etc. The first implementation of these methods was introduced by Levitt and Greer in 1977 [17] and was based on distances and dihedral angle profiles of α-Carbon atoms over a four-residue sliding window. Other methods include DSSP [18], DEFINE [19], P-CURVE [20], STRIDE [21], P-SEA [22], XTLSSTR [23], SECESTR [24], VoTAP [25], KAKSI [16], PROSIGN [26] and, more recently, a Bayesian method of minimum message length inference, where the protein is described as a collection of segments, with each segment belonging to one of eight potential models (the secondary structure motifs) [27]. Of these, DSSP and STRIDE use quite close criteria for the assignment...
of secondary structure and are the most widely used methods. The DSSP (Dictionary of Secondary Structure of Proteins) method, developed by Kabsch and Sander, uses hydrogen bond patterns based on an electrostatic model to assign the secondary structures. DSSP is the method of choice to annotate depositions to PDB [28]. The visualization software RASMOL [29] uses a fast DSSP-like algorithm, and both AMBER [30] and GROMACS [31] analysis tools use the DSSP software. The STRIDE (STRUCTural IDENTification) method, developed by Frishman and Argos, makes use of both hydrogen-bond patterns and backbone dihedral angles to assign the secondary structures. STRIDE is used by the popular visualization tool VMD [32]. Both DSSP and STRIDE have been used in structure prediction methods, for instance in Hidden Markov Models where the secondary structure assignments are used as a basis for generalized “alphabets” of backbone geometry [15]. The more recent KAKSI method [16] uses backbone dihedral angles and α-Carbon distances. It was introduced to improve the treatment of irregularities, such as kinks in α helices, in which case STRIDE tends to assign a longer, kinked helix, while KAKSI assigns several short helices.

Recent comparisons between DSSP and STRIDE show, in general, good agreement for well-defined secondary structures [22,33,25,34,16]. Thus, for instance, the results of both assignments were compared for 29 reference crystal structures (both in their reference state and in slightly distorted states) in order to relate them with the results of Circular Dichroism (CD) spectroscopy [33]. The results showed good agreement for α helices and β strands, while displaying more discrepancies in the “turns”. The overall performance of DSSP assignments was considered slightly better in the context of this CD spectra analysis. A more recent study [16] focused on how several of the assignment methods handled irregularities of the structures, such as the edges of the motifs, and various distortions [16]. The study was based on 3 X-ray datasets with, respectively, high, medium and low resolution and an NMR dataset. The authors found that the results from DSSP and STRIDE were very close for all the four datasets: a direct comparison between DSSP and STRIDE gave between 93.4% and 95.4% agreement using the C3 scores; while a comparison of each of these methods with respect to KAKSI based on SOV [35,36] scores yielded 89% and 92% agreement. The KAKSI method seemed to perform better than STRIDE when there were kinks in the motifs (DSSP was not used in this comparison). There seemed to be less agreement with other methods.

Clearly, there is not an exact and unique way to assign secondary structural motifs, since this depends on the definition for the secondary structure underlying the method [37,11]. The hope is that these automatic computational methods converge (e.g., to a secondary structure assignment provided by an experienced crystallographer). Most disagreements between the various methods occur in the terminal regions of the assigned structural motifs [14,38,37,34,16,39]. Conflicts also arise in determining whether the deviation from a given motif is simply a distortion of the motif or a break in the structure [14,16]. The disagreements between the methods is therefore expected to increase when the peptides exhibit irregular conformations.

In this work, we compare the predictions of these methods when applied to conformationally irregular peptides. In particular, we choose codes that give the closest results in comparisons of ordered structures: DSSP and STRIDE, based on hydrogen bonds; and KAKSI, based on α-Carbon distances. Clearly, these codes were developed to be applied mainly to periodic structures. However, the performance of these codes on more “disordered” structures is of interest for the following reasons. First, as mentioned above, these assignment programs tend to disagree in the terminal regions at the beginning and end of repetitive structures [14,38,37,34,16,39], resulting on different lengths for these structures, and further difficulties for the analysis of connecting loops [34]. Second, it is now understood that disordered proteins and peptides do not quite behave as random coils: both unfolded protein states and intrinsically Disordered Proteins [40] (IDPs) show considerable residual secondary structure [41,42]. Indeed, the use in NMR of chemical shifts, scalar couplings and residual dipolar couplings have found residual and/or transient secondary structure in both unfolded proteins and IDPs [43–59,61–64]. These are particularly present when IDPs become structured when binding a partner [65], or when the IDPs are prone to aggregation [58,66,67]. Both population analysis and free energy landscapes [68–71] reveal that the conformational preferences are much less entropic than those of a random coil. Third, in clear need of a tool to analyze residual secondary structures, researchers fall back into the secondary structure assignment codes, which unfortunately in these irregular or disordered structures do not necessarily achieve consensus.

We recently came across one of such cases [72] when we studied the structural preferences of the tridecapeptide corresponding to residues 659–671 of the envelope glycoprotein gp41 of HIV-1, which spans the 2PS monoclonal antibody epitope ELDKWA. There have been many studies to determine the structure of gp41 [659–671] (ELLELDKWASLWN) in aqueous solution, giving (at least initially) conflicting results. These include NMR and CD studies [73–75], a UV Resonance Raman Spectroscopy investigation [76], and crystal studies [77,78], Molecular Dynamics (MD) simulations also showed varying results [79,75,80,72], that can be attributed to the use of different force fields and sometimes inadequate sampling [72]. In addition, the use of different methods to characterize the secondary structure can further complicate comparisons and interpretation of the data. In this work, we compare the secondary structure assignments by DSSP, STRIDE and KAKSI when applied to gp41 [659–671]. In addition, we also compare these methods as applied to the disordered homopeptide polyasparagine. Asparagine-rich peptides are disordered at the monopeptide level, but are also found aggregated in highly ordered amylods [81–83]. Since the β content of both peptides is low, further comparisons are carried on polyasparagine dimers, expressly built as β-sheet dimers.

## 2. Methods

In this section we briefly review the basics of the DSSP, STRIDE and KAKSI secondary structure assignment methods, as well as the Segment Overlay method (SOV) for the evaluation of secondary structure prediction methods.

### 2.1. DSSP

The DSSP method [18] by Kabsch and Sander carries out the helical and sheet assignments entirely based on the backbone hydrogen bonds, as defined by an electrostatic model [84]. The electrostatic interaction energy is given by:

\[
E = f \left[ q_1 \frac{1}{r_{\text{ON}}} + q_2 \frac{1}{r_{\text{CH}}} - \frac{1}{r_{\text{OH}}} - \frac{1}{r_{\text{CN}}} \right],
\]

where \( q_1 = 0.42e, q_2 = 0.20e, f = 332 \text{ Ä kcal/(e}^2\text{ mol)} \), and \( r_{\text{AB}} \) is the distance between atoms A and B. A hydrogen bond exists between the C=O group of residue i and the N–H group of residue j if the electrostatic interaction energy satisfies \( E < -0.5 \text{ kcal/mol} \).

The basic turn pattern is defined if a hydrogen bond exists between residue i and residue \( i + n \) \( \rightarrow (i + n) \). If the residues in a basic turn do not belong to a helix, the DSSP output labels them as Turn ("T") (the smallest helix has to have two consecutive hydrogen bonds). Continuous "n helices" are defined when two adjacent hydrogen bonds are at \( i - 1 \rightarrow i + n - 1 \) and at \( i + 1 \rightarrow i + n \), and the residues \( i, i + 1, \ldots, i + n - 1 \) are labeled as "n helix". The notations for the "n helix" are "G" (n=3) for 3↓ helix, "H" (n=4) for...
alpha helix, and "I" (n=5) for pi helix. This helix definition does not provide assignments to the edge residues with the initial and final hydrogen bonds in the helix.

Another basic pattern, bridge, has been defined as follows: two non-overlapping stretches of three residues each, i−1, i, i+1 and j−1, j, j+1 form a parallel bridge when there are two hydrogen bonds i−1→j and j+1→i, or two hydrogen bonds j−1→i and i+1→j; they form an anti-parallel bridge if there are two hydrogen bonds i→j and j→i, or two hydrogen bonds i−1→j+1 and j−1→i+1. The residues i and j are labeled as bridge by using lower case for parallel bridge and upper case for anti-parallel bridge. The parallel and anti-parallel information are included in the sheet "structure" section, but not in the "summary" section. One or more consecutive bridges of identical type form a ladder; and one or more connected ladders form a sheet. In the "summary" section, the single bridge (ladder of length 1) is denoted as "B", and the longer ladder is denoted as "E", without distinguishing parallel and antiparallel information. A beta bulge connects two perfect parallels or bridges of the same type, and the connected ladders are treated as a single ladder, with all the residues, including the bulges, being labeled as "E". There are at most one bulge residue on one strand and 4 bulge residues on the other strand.

DSSP also defines Bend, Chirality, SS bonds and Chain Breaks. Bend is defined when the angle C_i−2→C_i→C_i+2 is smaller than 110°, and is labeled as "S". The dihedral angle ω(i) = (C_i−2, C_i, C_i+1, C_i+2) is used to define the Chirality, and only the sign of ω(i) is reported as "+"; if 0° < ω < 180°; or as "−", if −180° < ω < 0°. SS bonds represent the disulfide bonds joining Cys residues. The chain breaks indicate covalent breaks in the peptide sequence, and are labeled as "I". The information about chirality, SS bonds and chain breaks is included in the DSSP output files but not in the summary section which only includes "H", "G", "T", "B", "E" (parallel and anti-parallel are labeled together here), "T", "S" (any bend that is not assigned to any of the six previous categories) and "C"; the last one used to indicate coil (as the default for a conformation that does not belong to any of the other categories).

2.2. STRIDE

The STRIDE method [21] by Frishman and Argos makes use of hydrogen bonds in a similar manner as DSSP, except that the backbone hydrogen bonds are defined differently. STRIDE uses an empirical function to assign the hydrogen bond energy E_HB. In addition, the backbone torsion angles evaluated according to the regions of the Ramachandran plot in which the residues fall, are also involved in the definition of alpha helix and beta strands. The hydrogen bond energy E_HB is ignored if the backbone torsions are unfavorable.

For the alpha helix assignment, STRIDE uses a similar definition as DSSP, so that when two adjacent hydrogen bonds are located at i−1→i+3, then the residues i, i+1, . . . , i+3 are assigned as an alpha helix ("H"). STRIDE extends the criterion so that if the probabilities of the torsional angles of edge residues i−1 and i+4 satisfy some additional constraints, the edge residues are also included in the helix. The 3j helix and pi helix definitions and notation agree with those of DSSP (except for the definition of the hydrogen bond itself).

STRIDE has several definitions of bridge, which include three bridges with a similar definition to that in DSSP, and two definitions not found in DSSP. Consecutive bridges form a beta sheet, with no more than one bulge between bridges and no more than four bulges on the other strand, and the participating residues are denoted as "E", except the bridges flanking the given beta sheet where only internal residues are labeled as "E". An isolated bridge is labeled as "B", except for the type III bridges [21] that on one side neither of the residues in the bridge is internal, which are assigned as "b". This "b" is not to be confused with the one in the sheet structure section in DSSP (which is used to label a parallel bridge).

The turn assigned by STRIDE is based on Richardson's definition [1], which classifies six distinct turn types (I, L, II, III, VIa and VIb) based on the dihedral angles (ϕ, ψ); as well as the extended definition (type VIII) proposed by Wilmot and Thornton [85].

2.3. KAKSI

The KAKSI method [16] was introduced to optimize the fit to the secondary structure assignments obtained from the PDB files. The assignment is based on the Cα distances and (ϕ, ψ) dihedral angles, and it is done by sliding one window along the sequence for alpha helix, and two windows for beta sheet, to ensure that the assigned beta strand is involved in a beta sheet. As the windows slide along the sequence, the distance criterion (C1) for helix is checked first, this is followed by the angle criterion (C2) for helix; all the eligible residues are assigned as helix temporarily. The kink criteria (K1) and (K2) are applied to detect kinks along these helix residues, and all the kinks are assigned as coil. Helices with length shorter than 5 residues are discarded. The beta sheet distance (C1) and angle (C2) criteria are then used to detect beta strands after helices have been assigned. Finally, the contiguous criteria (C3) is used to introduce a coil between adjacent helices and beta strands, a process that shortens the intervening helix by one residue. The output includes only three kinds of secondary structure: helix, beta strand and coil, which are labeled as "H", "b" and "C".

2.4. SOV

The Segment Overlap measure [35,36] (SOV) by Rost et al. is usually employed to evaluate the protein secondary structure prediction assessment. SOV was developed to take into account that the correct characterization of the type and location of the secondary structure elements is frequently far more important than the assignment at the residue level. By contrast, the C3 scores define the overall fraction of residues predicted in a given pattern, which sometimes can be quite misleading as to the accuracy of the prediction (as the SOV authors noted, assigning the entire myoglobin chain as a single helix gives a C3 score of approximately 80% [36]). In addition, SOV is constructed so as to minimize the ambiguity in the definition of the segment ends due to the differences of criteria of the different secondary structural assignment methods.

The SOV measure is defined as follows. Let (si, sj) denote a pair of overlapping segments, S(i) denote the set of all the overlapping pairs of segments (si, sj) in state i, and S(i) be the set of all the segments si and sj in state i, with no overlap between si and sj. The segment overlap measure for state i is defined as [36]:

\[
\text{SOV}(i) = 100 \times \frac{1}{N(i)} \sum_{s(i)} \left\{ \frac{\min(s(i), s(j)) + \delta(s(i), s(j)) \times \text{len}(s(i))}{\max(s(i), s(j)) \times \text{len}(s(i))} \right\}
\]

\[
N(i) = \sum_{s(i)} \text{len}(s(i))
\]

\[
\delta(s(i), s(j)) = \min \left\{ \left( \frac{\text{max}(s(i), s(j)) - \min(s(i), s(j))}{\text{len}(s(i))} \right) ; \min(s(i), s(j)); \text{int} \left( \frac{\text{len}(s(i))}{2} \right) \right\}
\]

In this expression, len(s(i)) is the number of residues in segment si; \min(s(i), s(j)) is the length of the overlapping area between si and sj when they are in state i; \max(s(i), s(j)) is the total extent for which either of the segments si or sj has a residue in state i; \sum s(i) is taken over all the segment pairs in state i which overlap by at least one residue; and \sum s(i) is taken over the remaining segments in state i (that do not overlap). The SOV measurement for multi-state
secondary structure assignment is defined as

\[
SOV = 100 \sum_{i=1}^{N} \left[ \min(x_i, s_i) - \min(x_i, s_i) \right] \times \frac{\text{len}(x_i)}{N(i)},
\]

where \(N\) is the sum of \(N(i)\) over all secondary structure assignments:

\[
N(i) = \sum_{i=1}^{N} N(i)
\]

Note that since one of the two assignments is chosen as observed assignment (reference), the roles of \(x_i\) and \(s_i\) are not symmetric.

3. Simulation details

As example of conformationally irregular peptides we chose the gp41<sub>659–671</sub> peptide, polyasparagine with 18 residues, here denoted as N<sub>18</sub>. In addition, we also considered polyasparagine dimers (N<sub>8</sub>–N<sub>8</sub>).

3.1. gp41<sub>659–671</sub> Peptide

This simulation [72] includes one gp41<sub>659–671</sub> molecule with sequence Ace-ELLELDNASLWN-NH<sub>2</sub>. The capping groups are assumed to make the peptide more similar to the corresponding segment in the native protein. The simulation was carried out with the AMBER f12SB version of the Cornell et al. force field [86] under explicit solvent.

Topology and parameter files along with the coordinates corresponding to the unfolded peptide were generated via the LEAP program of the AMBER v.12 [30] simulation packages. We started the simulations with an implicit solvent. The implicit water model is based on the Generalized Born approximation (GB) [87] including the surface area contributions using the LCPO model [88] (GB/SA) with the surface tension set to 0.005 kcal/mol/Å<sup>2</sup>. For the GB model, we used the Gromacs II model [89], with a cutoff of 25 Å. After a short minimization, an initial 10 ns MD simulation was used to generate equilibrium conformations of the peptide. Five different conformations from this run were chosen as initial conformations for the explicit simulations.

The TIP3P water model [90] was used for the explicit solvent simulations under periodic boundary conditions. The five initial conformations from the implicit solvent simulations were solvated in a rectangular box, with an average number of waters of approximately 5670. The peptide has two net negative charges that were neutralized by the addition of two Na<sup>+</sup> ions. Electrostatics were handled by the PME method [91,92], with a direct space cutoff of 9 Å and an average mesh size of approximately 1 Å for the lattice calculations. The equilibration process took place in four steps. First, we applied steepest descent followed by conjugate gradient minimization keeping the peptide atoms fixed at their initial positions. Then we carried out unrestrained steepest descent followed by conjugate gradient minimizations. This was followed by short MD runs under constant volume while the system was heated from 0K to 300K with weak restraints on the peptide atoms. Finally, the system was kept at 300K via Langevin dynamics with a collision frequency \(\gamma = 1.0 \text{ ps}^{-1}\), and at constant pressure (1 atm) via the Berendsen barostat [93] with the isothermal compressibility \(\beta_T = 44.6 \times 10^{-6} \text{ bar}^{-1}\) and the pressure relaxation time \(\tau_p = 1.0 \text{ ps}\). These NPT Langevin dynamics simulations were carried out for 2 ns, and the density of the system was found to be stable around 1.0 g/cm<sup>3</sup>. The five equilibrated conformations provided the initial conformations for NPT simulations with Langevin dynamics, each simulation 40 ns long. One configuration was selected for a further 100 ns run. The second half of each of the 40 ns runs and the 100 ns runs were used for data analysis (for a total of 300 ns). Coordinates were sampled every 2 ps.

3.2. Polyasparagine

Polyasparagine of sequence Ace-Asn<sub>18</sub>-NH<sub>2</sub> (N<sub>18</sub>) was simulated in an implicit water model. In order to enhance the conformational sampling of this peptide we used temperature replica exchange molecular dynamics (T-REMD) [94]. The unfolded peptide was generated via the LEAP program of AMBER v.12, and the implicit water model was the same one as described for gp41<sub>659–671</sub>. Eighteen replicas were used for the T-REMD scheme at temperatures 300, 322, 346, 371, 398, 427, 459, 493, 529, 567, 609, 654, 702, 753, 809, 868, 932, 1000 K. After a short minimization, an initial 500 ps MD simulation for each replica was used to heat the system to the respective temperature via Langevin dynamics with a collision frequency \(\gamma = 1.0 \text{ ps}^{-1}\), and at constant pressure (1 atm) via the Berendsen barostat [93]. Then another 500 ps MD simulation for each replica was run to generate equilibrium conformations of the peptide. The T-REMD scheme was used on the equilibrium conformations from this run. We ran 200 ns T-REMD simulations and the coordinates of the T = 300K replica were sampled every picosecond. Only the second half of the sampled conformations was used for secondary structure analysis.

3.3. Polyasparagine dimers

In order to enhance the conformational sampling of \(\beta\) sheet, polyasparagine dimers (N<sub>8</sub>–N<sub>8</sub>), consisting of two Ace-Asn<sub>8</sub>–NMe peptides, were simulated in an explicit water model with parallel and anti-parallel orientations.

The unfolded peptide was generated via the LEAP program of AMBER v.12. The dimers were assembled in parallel and antiparallel conformations with the proper symmetry and then solvated with TIP3P waters in a truncated octahedron box with 2772 waters and periodic boundary conditions. Electrostatics were handled by the PME method. The dimers tend to quickly unfold under regular MD. Since our aim is to compare how the codes assign structure to \(\beta\) sheet conformations that are not ideal, this system provides a good benchmark as long as the \(\beta\) sheet is somehow enforced. Thus, we carry out two sets of simulations: in one set, we kept the dimers restrained (allowing, of course, some range of motion); in the other set, we took out the restraints and allowed the dimers to evolve, not at room temperature but at 200 K. Initial equilibration started by minimizing the energy of the system through steepest descent followed by conjugate gradient minimization keeping the peptide atoms fixed at their initial positions. Then we carried out unrestrained steepest descent followed by conjugate gradient minimizations. This was followed by short MD runs under constant volume while the system was heated from 0K to 300K with weak restraints on the peptide atoms. Finally, the system was kept at 300K via Langevin dynamics with a collision frequency \(\gamma = 1.0 \text{ ps}^{-1}\), and at constant pressure (1 atm) via the Berendsen barostat [93] with the isothermal compressibility \(\beta_T = 44.6 \times 10^{-6} \text{ bar}^{-1}\) and the pressure relaxation time \(\tau_p = 1.0 \text{ ps}\). These NPT Langevin dynamics simulations were carried out for 2 ns, and the density of the system was found to be stable around 1.0 g/cm<sup>3</sup>. The five equilibrated conformations provided the initial conformations for NPT simulations with Langevin dynamics, each simulation 40 ns long. One configuration was selected for a further 100 ns run. The second half of each of the 40 ns runs and the 100 ns
conformations for each gp41_{659-671} and N_{18} system, as well as to 10,000 conformations of each of the four N_{9}–N_{18} dimer systems. We discuss each result and explain the differences in terms of the criteria that these methods have used to assign the secondary structure.

### 4.1. Secondary structure assignment and SOV scores

Table 1 shows the average secondary structure content (over all the residues), including α helix (H), 3_10 helix (G), π helix (I), turn (T), isolated β bridge (B), extended strand (E), and sheet (B+E) as assigned by DSSP, STRIDE and KAKSI (when it provides output). First we consider the results corresponding to the mainly disordered peptides gp41_{659-671} and N_{18}. Notice that even though the strand and sheet contents represent small percentages, the absolute numbers of residues in these conformations are not negligible, as they are calculated over 100,000 conformations for each peptide. The results from DSSP and STRIDE are comparable (DSSP gives higher helical content and lower strand content) except for turn, where DSSP gives considerably lower values. KAKSI gives similar helical content, lower sheet content, and no output for turns. Fig. 1a and b show sample conformations of gp41_{659-671} and N_{18} displaying secondary structure assignments by the three different methods.

The β sheet dimers show good agreement for the parallel conformations (the restrained one in better agreement than the unrestrained) and less agreement on the anti-parallel conformations. In particular, the unrestrained, antiparallel dimers exhibit the smallest β sheet population, and are in rather poor agreement. Sample conformations of this situation are given in Fig. 1c.

Next, we compare SOV and C3 scores for the different assignments. The simplest residue by residue comparison would involve counting the number of residues that agree for a given category versus the total number of residues (for a full sequence comparison). There are several possibilities for evaluating this. The most commonly used is the overall C3 score, which is defined as:

\[
C_3 = \frac{N_H + N_E + N_C}{N}
\]

where \(N_H\), \(N_E\) and \(N_C\) represent the number of identical residues as assigned by the secondary structure algorithms in the categories helix ("H"), sheet ("B") and coil ("C"). Here "H" includes "H", "G", "I", "T", "B", "E", "T", and "C". In principle, one could define a "C7" score with the same definition as the C3 score, except that it includes 7 categories ("H", "G", "I", "T", "B", "E", "T", and "C"). Naturally, C7 scores would generally be lower than C3 scores. As an example, given two sequences SA = GGGGGHHHH and SB = HHHHHHHH, the C3 score is 100% while the C7 score is 0%. One could also define a C3 score for a particular category, by counting the number of residues that coincide versus the total number of residues in that category. This would generally not be symmetric and would depend on which category is taken as reference. For instance, given the sequences SA = CHHHHHHHHHH and SB = CCCCCCCCC, then C3_S = 50% when SA is taken as reference, and C3_B = 100% when SB is taken as reference. In general, this type of comparison is not employed and only the symmetric overall C3 score is used (C3 = 58% in this case).

The SOV measure was introduced to take into account both the type and the location of the secondary structure elements, and in general (but not always) gives more realistic scores. Its results depend on which structure is taken as reference. For instance, consider the two twelve-residue segments SA and SB with sequences SA = CHHHHHHHHHH and SB = CCCCCCCCC, for which the overall C3 score is C3 = 58%. If SA is taken as reference, the helical overlap SOV_H is 70%, and the overall overlap SOV_{all} is 63%. If SB is taken as reference, then SOV_H = 70% while SOV_{all} = 46%. As a more extreme example, consider the 10-residue sequences SA = CECECECECE and SB = EEEEEEEEE. Taking either sequence as reference gives SOV_H = 10%, with an overall value SOV_{all} = 5% for SA as reference and SOV_{all} = 10% for SB as reference. The overall C3 scores instead are 50%.

Table 2 gives the average value of the SOV measures over all the conformations between DSSP and STRIDE, the latter being chosen as reference. First we look at the disordered peptides gp41_{659-671} and N_{18}. The SOV_{all} value is larger for gp41_{659-671} than N_{18}, which reflects the fact that gp41_{659-671} has more helical population than N_{18}. The SOV measures between DSSP and STRIDE on α helix are in good agreement, with values greater than 95%. But the SOV measures on other secondary structure assignments are considerably less good, especially for the turn assignment, whose SOV measures are less than 45%. The overall SOV measures for the entire peptides are also lower (as low as 54% for N_{18}), indicating a relative low overall overlap between the assignments of secondary structure by DSSP and STRIDE. Since the SOV measure is not symmetric, we also show the values obtained when DSSP is taken as reference in Table 3. In this case, there is a further lowering of the overall overlap. Table 4 gives the average value of the SOV

![Fig. 1. Snapshots of sample conformations. (a) gp41_{659-671} and (b) N_{18}, with assignments by DSSP (left), STRIDE (middle), and KAKSI (right). (c) Dimer N_{9}-N_{18} analyzed with DSSP (left) and STRIDE (right). Colors purple, blue, green, yellow and white represent α helix (H), 3_10 helix (G), turn (T), β sheet (E) and coil (C) for DSSP and STRIDE assignments; for KAKSI, red represents helix and white represents coil.](image-url)
Table 1
Secondary structure content (%) of gp41_{41-61}, N8, and dimers N8-N8, in parallel (P) and anti-parallel (AP) orientation, with a restrained (R) or unrestrained (U) backbone.

<table>
<thead>
<tr>
<th>Peptide</th>
<th>Codes</th>
<th>ω</th>
<th>3ω</th>
<th>τ</th>
<th>Helix</th>
<th>β bridge</th>
<th>β strand</th>
<th>Sheet</th>
<th>Turn</th>
</tr>
</thead>
<tbody>
<tr>
<td>gp41_{41-61}</td>
<td>STRIDE</td>
<td>33.36</td>
<td>16.56</td>
<td>1.12</td>
<td>51.04</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>19.54</td>
</tr>
<tr>
<td></td>
<td>KAKSI</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>44.56</td>
<td>–</td>
<td>–</td>
<td>0</td>
<td>–</td>
</tr>
<tr>
<td>N8</td>
<td>STRIDE</td>
<td>11.91</td>
<td>4.87</td>
<td>0.12</td>
<td>39.56</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>13.11</td>
</tr>
<tr>
<td></td>
<td>KAKSI</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>20.13</td>
<td>–</td>
<td>–</td>
<td>0.04</td>
<td>–</td>
</tr>
<tr>
<td>N8-N8, U-P</td>
<td>DSSP</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>STRIDE</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>N8-N8, R-P</td>
<td>DSSP</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>STRIDE</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>N8-N8, U-A</td>
<td>DSSP</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>STRIDE</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Table 2
Average SOV measures (%) between DSSP and STRIDE, with STRIDE taken as reference. Values are provided for helix (H), 3ω helix (G), τ helix (I), turn (T), isoalted bridge (B), β strand (E), coil (C), as well as the overlap for the complete peptide, SOV_{total}. Data is shown for gp41_{41-61}, N8, and dimers N8-N8 in parallel (P) and anti-parallel (AP) orientation, with a restrained (R) or unrestrained (U) backbone.

<table>
<thead>
<tr>
<th>Peptide</th>
<th>SOV_{H}</th>
<th>SOV_{G}</th>
<th>SOV_{I}</th>
<th>SOV_{T}</th>
<th>SOV_{E}</th>
<th>SOV_{C}</th>
<th>SOV_{total}</th>
</tr>
</thead>
<tbody>
<tr>
<td>gp41_{41-61}</td>
<td>98.06</td>
<td>78.11</td>
<td>79.31</td>
<td>43.53</td>
<td>–</td>
<td>–</td>
<td>70.90</td>
</tr>
<tr>
<td>N8</td>
<td>95.18</td>
<td>61.82</td>
<td>52.74</td>
<td>41.01</td>
<td>86.64</td>
<td>45.25</td>
<td>72.68</td>
</tr>
<tr>
<td>N8-N8, R-P</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>0.00</td>
<td>98.54</td>
<td>98.54</td>
</tr>
<tr>
<td>N8-N8, R-AP</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>0.00</td>
<td>56.36</td>
<td>59.89</td>
<td>49.07</td>
</tr>
<tr>
<td>N8-N8, U-P</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>0.00</td>
<td>89.73</td>
<td>89.73</td>
<td>49.07</td>
</tr>
<tr>
<td>N8-N8, U-AP</td>
<td>–</td>
<td>–</td>
<td>22.49</td>
<td>6.94</td>
<td>95.79</td>
<td>67.67</td>
<td>48.85</td>
</tr>
</tbody>
</table>

Table 3
Same as Table 2, except that DSSP is taken as reference.

<table>
<thead>
<tr>
<th>Peptide</th>
<th>SOV_{H}</th>
<th>SOV_{G}</th>
<th>SOV_{I}</th>
<th>SOV_{T}</th>
<th>SOV_{E}</th>
<th>SOV_{C}</th>
<th>SOV_{total}</th>
</tr>
</thead>
<tbody>
<tr>
<td>gp41_{41-61}</td>
<td>83.98</td>
<td>62.72</td>
<td>61.69</td>
<td>22.77</td>
<td>–</td>
<td>N/A</td>
<td>46.31</td>
</tr>
<tr>
<td>N8</td>
<td>65.13</td>
<td>46.01</td>
<td>30.86</td>
<td>44.51</td>
<td>87.11</td>
<td>89.11</td>
<td>44.92</td>
</tr>
<tr>
<td>N8-N8, R-P</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>98.54</td>
<td>99.95</td>
</tr>
<tr>
<td>N8-N8, R-AP</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>94.03</td>
<td>92.53</td>
</tr>
<tr>
<td>N8-N8, U-P</td>
<td>–</td>
<td>–</td>
<td>40.43</td>
<td>15.79</td>
<td>27.02</td>
<td>55.34</td>
<td>51.89</td>
</tr>
<tr>
<td>N8-N8, U-AP</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

Table 4
Average SOV measures (%) between KAKSI and DSSP/STRIDE based on helix (including \( H, \) \( C, \) \( I \) from DSSP and STRIDE), \( \beta \) sheet (including \( B, \) \( E, \) and \( \beta \) from DSSP and STRIDE) and coil (including \( T, \) and \( I \) from DSSP and STRIDE), KAKSI is chosen as reference.

<table>
<thead>
<tr>
<th>Peptide</th>
<th>Codes</th>
<th>SOV_{H}</th>
<th>SOV_{G}</th>
<th>SOV_{I}</th>
<th>SOV_{T}</th>
<th>SOV_{E}</th>
<th>SOV_{C}</th>
<th>SOV_{total}</th>
</tr>
</thead>
<tbody>
<tr>
<td>gp41_{41-61}</td>
<td>DSSP</td>
<td>88.97</td>
<td>N/A</td>
<td>53.45</td>
<td>65.48</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>STRIDE</td>
<td>85.74</td>
<td>N/A</td>
<td>60.26</td>
<td>69.80</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N8</td>
<td>DSSP</td>
<td>73.12</td>
<td>33.79</td>
<td>71.46</td>
<td>70.61</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>STRIDE</td>
<td>62.56</td>
<td>37.58</td>
<td>76.95</td>
<td>74.37</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N8-N8, R-P</td>
<td>DSSP</td>
<td>66.67</td>
<td>53.46</td>
<td>58.45</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>STRIDE</td>
<td>65.79</td>
<td>51.34</td>
<td>57.83</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N8-N8, R-AP</td>
<td>DSSP</td>
<td>64.33</td>
<td>72.63</td>
<td>66.31</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>STRIDE</td>
<td>47.09</td>
<td>72.63</td>
<td>59.27</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N8-N8, U-P</td>
<td>DSSP</td>
<td>76.11</td>
<td>49.13</td>
<td>53.86</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>STRIDE</td>
<td>69.72</td>
<td>46.27</td>
<td>50.23</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N8-N8, U-AP</td>
<td>DSSP</td>
<td>–</td>
<td>75.63</td>
<td>75.63</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>STRIDE</td>
<td>–</td>
<td>92.94</td>
<td>92.94</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
content. Evaluating the SOV scores of DSSP and STRIDE against KAKSI notably decreases the overall agreement for the parallel dimers, and increases the agreement for the antiparallel dimers. This is particularly the case for the results of the unrestricted parallel dimer, where both STRIDE and KAKSI predict a large amount of turns and coils.

Table 5 compares the average overall C3 scores over all the conformations. Comparisons with KAKSI give lower values, reflecting the different approach of KAKSI. C7 scores between DSSP and STRIDE are 66.82% for gp41659–671 and 58.35% for N18.

4.2. Segment length distribution

Figs. 2 and 3 show the population distribution of conformations versus the segment length, e.g., the number of adjacent residues with the same assignment. The 310 helix and π helix for DSSP and STRIDE are in good agreement. The preferred lengths for the 310 helix are 3 and 6, which means that the secondary structure consists of a single 310 helix loop or two 310 helix contiguous loops. For the π helix, the favorite length is 5, i.e., the length of a single π helix loop. STRIDE tends to give longer α helices and turns. The favorite length of the α helix from DSSP is 4, which is the length of a single α helix loop with 2 hydrogen bonds between residues (i−1, i+1) as well as (i, i+4); while the favorite length from STRIDE is 5, which is one typical α helix loop with one of the end residues. This is obviously due to the different criteria used by these two algorithms. In the case of turn, DSSP gives a short turn with a length around 2 residues, but STRIDE favors a longer turn with a length around 4 residues, and even longer. Fig. 4 shows how the population distribution changes when the helical secondary structures G, H and I are grouped together into the single “Helix” category. By design, the minimum helical segment length for KAKSI is 5.

4.3. Hydrogen bonds

Both DSSP and STRIDE base their secondary structure algorithms on the definition of hydrogen bonds, and many of the differences between their results are due to the different criteria employed for the definition of these bonds. Fig. 5 shows the population distribution on the plane determined by the alignment angle and the hydrogen-bond length. The alignment angle θ is defined as the angle N-H...O, and the hydrogen bond length as the distance between atoms O and N in each hydrogen bond pair. We observe good agreement between both predictions except that there is a 3.5 Å cutoff on the hydrogen-bond length in the STRIDE prediction. However, as we show in the next figures, the hydrogen bonds predicted by DSSP are more consistent with the helical definitions than those by STRIDE.

Table 5

<table>
<thead>
<tr>
<th></th>
<th>KAKSI</th>
<th>DSSP</th>
<th>STRIDE</th>
</tr>
</thead>
<tbody>
<tr>
<td>gp41659–671</td>
<td></td>
<td>85.12</td>
<td>70.35</td>
</tr>
<tr>
<td></td>
<td>STRIDE</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>DSSP</td>
<td>89.56</td>
<td>80.71</td>
</tr>
<tr>
<td></td>
<td>STRIDE</td>
<td></td>
<td>83.72</td>
</tr>
</tbody>
</table>

Fig. 4. Length distribution of “Helix” for gp41659–671 (top) and N18 (bottom).
4.4. Helical properties

We find that there are some differences between the helical predictions by DSSP and STRIDE, especially for the 310 helix. Here we show results for the dihedral angles \( \phi, \psi \) and the translations of the helices as assigned by these two algorithms.

Two important characteristics of a helix are the hydrogen bond geometry and the rotation angle \( \Omega \) per residue. The 310 helix is defined by the repeated hydrogen bond \((i+3 \rightarrow i)\) between the \( \mathrm{N-H} \) group of an amino acid and the \( \mathrm{C=O} \) group three residues earlier. The rotation angle of a 310 helix is about 120°, resulting in 3 residues per turn. The \( \alpha \) helix is defined by the \((i+4 \rightarrow i)\) hydrogen bond, with a rotation angle of about 100°, which includes 3.6 residues per turn. A general formula for the rotation angle \( \Omega \) per residue of the helix structure is given by:

\[
3 \cos \Omega = 1 - 4 \cos^2 \left( \frac{\psi_i + \phi_{i+1}}{2} \right)
\]

where \( \psi_i \) is the backbone dihedral angle \( \psi \) in residue \( i \), and \( \phi_{i+1} \) is the backbone dihedral angle \( \phi \) in the next residue. Since the 310 helix and \( \alpha \) helix adopt rotation angles at approximately 120° and 100°, we can identify the helix by checking the sum of the backbone dihedral angle \( \psi_i + \phi_{i+1} \). For the 310 helix, it is roughly -75° and for the \( \alpha \) helix it is approximately -104°. The \( \alpha \) helix has 1.5 Å translation length along the axis and its backbone dihedral angles are distributed around \((-60°, -45°)\). The 310 helix translational length is about 2.0 Å. The generally accepted values of the dihedral angles are \((-49°, -26°)\) [95,96]. The ideal 310 helix has angles \((-74°, -4°)\) [97,98]. Sometimes values of \((-49°, -18°)\) (with rise per residue of 1.7 Å) and \((-71°, -18°)\) (with 3.2 residues per turn) are also cited [99,101].

Figs. 6 and 7 show the Ramachandran plot and the population distribution as function of translation length for the 310 and \( \alpha \) helices. The residues are assigned by DSSP and STRIDE as \( \psi \) and \( \alpha \), where \( \psi \) (positive) indicates that the residue is assigned as 310 or \( \alpha \) helix by the respective algorithm (based on the \( \mathrm{H-H} \) bond pattern), and \( \alpha \) (negative) indicates that the residue is not assigned to that state by the algorithm. Thus, for instance, “DpSn” denotes the set of residues found in a given state by DSSP (DSSP positive) but not by STRIDE (STRIDE negative). The three possible combinations, DpSp, DpSn and DnSp are shown for the 310 helix in Fig. 6 and for the \( \alpha \) helix in Fig. 7. For gp41 [659-671] the conformation ensemble Dpsp assigned as 310 helix both by DSSP and STRIDE has a translation length distribution located around 1.8 Å and the Ramachandran plot shows the backbone dihedral distributed around \((-64°, -23°)\); both metrics are similar to the \( \alpha \) helix properties. The DpSn ensemble assigned as 310 helix by DSSP but not STRIDE has a better translation located exactly at 2 Å but the DnSp ensemble has a 1.5 Å translational length with dihedral angles located at \((-65°, -34°)\). But the DnSp ensemble has a 1.8 Å helical rise, and dihedral angles \((-65°, -18°)\), which is very close to the DnSp ensemble in the 310 helix case. These results show that STRIDE tends to give tighter 310 helices and losser \( \alpha \) helices, which suggests that STRIDE cannot clearly distinguish between 310 helices and \( \alpha \) helices. Similar results are observed for N18.
Fig. 6. Odd rows show Ramachandran plots (y-axis is $\psi$ angle in degrees, x-axis is $\phi$ angle in degrees) and even rows show translation length distribution (y-axis) as a function of helical rise (x-axis, in Å) for $\beta_3$ helix residues in gp41659–671 (top) and N18 (bottom). The DpSp (left column), DpSn (middle column) and DnSp (right column) are residues assigned by DSSP and STRIDE as pp, pn and np, where p indicates that the residue is assigned as $3_{10}$ helix by the respective algorithm, and n indicates that the residue is not assigned to that state by the algorithm.

Fig. 7. Odd rows show Ramachandran plots (y-axis is $\psi$ angle in degrees, x-axis is $\phi$ angle in degrees) and even rows show translation length distribution (y-axis) as a function of helical rise (x-axis, in Å) for $\alpha$ helix residues in gp41659–671 (top) and N18 (bottom). The DpSp (left column), DpSn (middle column) and DnSp (right column) are residues assigned by DSSP and STRIDE as pp, pn and np, where p indicates that the residue is assigned as $\alpha$ helix by the respective algorithm, and n indicates that the residue is not assigned to that state by the algorithm.
Fig. 8 shows the dihedral angle sum $\psi_i + \phi_i$ distribution of $3_{10}$ helix and $\alpha$ helix as assigned by DSSP and STRIDE. The standard $3_{10}$ helix is expected to show values of the sum at around $-75^\circ$ (marked by the vertical dashed line), however both DSSP and STRIDE assignments are shifted $20^\circ$ towards $-95^\circ$, which is closer to the $\alpha$ helix. The $\alpha$ helix assignments by these two algorithms have a better agreement with the theoretical dihedral angle sum, $-104^\circ$. This suggests that the $3_{10}$ and $\alpha$ helices identified by DSSP or STRIDE for these peptides tend to be quite close to each other in their helical properties. Fig. 9 shows the $\psi_i + \phi_i$ distribution of $3_{10}$ helix and $\alpha$ helix residues when the assignments by DSSP and STRIDE do not agree (i.e., the DpSn and DnSp sets). The positive DSSP assignments (blue lines) have better agreement with the standard value of the dihedral angle sum for both the $\alpha$ helix ($-104^\circ$) and the $3_{10}$ helix ($-75^\circ$). This, together with the better helical rise prediction shown in Figs. 6 and 7, points out to a better helix recognition by DSSP.

4.5. Turn and bend

DSSP and STRIDE use different definitions for turn. The $n$-turn pattern is defined as a single H bond from CO($i$) to NH($i+n$) by DSSP. If there is no helix structure detected at the $n$-turn, the DSSP output identifies these residues as turn, i.e., “T”. STRIDE predicts turn patterns by using Richardson’s definition [1], plus the extended turn structure definition by Wilmot and Thornton [85]. The $\beta$ turn “T” from STRIDE is given by the following criterion: the atoms C$^\beta$ – C$^\gamma$, distance is <7 Å and the central residues are not helix. The extended turn type such as $\beta$, $\beta$II, etc. are given by using the backbone dihedral informations. Even though these two definitions are equivalent in many cases, there may still be considerable disagreement between the predictions.

Fig. 10 shows the secondary structure distribution predicted by DSSP on all the residues which are predicted as turn type by STRIDE. Only 36.7% and 40.5% of turn type residues from STRIDE are predicted as “T” by DSSP; 38.7% and 48.1% residues are predicted as either bend (“S”) or coil (“C”) by DSSP (the summary file reports the sum of bend and coil as coil). Finally, we dump the remaining 24.5% and 11.3% turn-type residues identified by STRIDE into an “other” category, mainly composed of helical residues. Checking the H-bonds detected by DSSP on these helices reveals that 46% (gp41469-671) and 49% (N18) of these H-bonds are longer than 3.5 Å, which is forbidden by STRIDE. Conversely, for the residues identified as turn by DSSP, STRIDE gives 58.4% (gp41469-671) and 86.2% (N18) turn; 22.0% (gp41469-671) and 5.5% (N18) coil; and 19.6% (gp41469-671) and 83.3% (N18) “other” (which is mainly helix).

Fig. 10. Pie chart of secondary structure distribution assigned by DSSP for all the residues identified as “turn” by STRIDE (i.e., the STRIDE graphs in each case are a full green “turn” circle). Turn: green; bend: yellow; coil: blue; other: red. Left: gp41469-671; right: N18. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)
5. Discussion and conclusions

Secondary structure analysis codes were originally developed to be applied mainly to periodic structures. Even though there is not an exact and unique way to assign secondary structural motifs [37,11], these automatic computational methods generally converge when applied to regular structures. However, as several comparison studies in the past have shown, not all codes converge equally. DSSP and STRIDE, based on hydrogen bonds and also on dihedral angles for STRIDE, are the most popular codes in current use, and they give the closest results in comparisons of ordered structures [22,25,34,23,16]. KAKSI, using a different criterion based on $\alpha$-Carbon distances and backbone dihedral angles, also gives relatively close results [16].

As the protein conformations become more distorted, the agreement between the secondary structure assignment codes is expected to deteriorate. Indeed, it has been observed that most disagreements between the various methods occur in the terminal regions of the assigned structural motifs [14,38,37,34,16,39], which results on different lengths for these structures, and further difficulties for the analysis of connecting loops [34]. Conflicts also arise in determining whether the deviation from a given motif is simply a distortion of the motif or a break in the structure [14,16]. Conformationally irregular peptides offer therefore a great test bench to explore the limits and differences between these codes. Molecular dynamics simulations are optimal to generate large number of conformations (there are 100,000 conformations for each peptide presented in this work), especially when one considers the dearth of experimental structural data for disordered peptides.

Secondary structure analysis codes, or perhaps suitable extensions of them, are also needed to help characterize residual or transient secondary structure in unfolded proteins and IDPs [43–59,41,60–64]. Indeed, it is recognized that disordered proteins and peptides do not quite behave as random coils at the residue level [100,41,42]; yes, many of these chains (especially if they are long) satisfy global properties and scaling laws typical of random coils [101,102], such as radius of gyration $R_g \propto N^{0.6}$, where $N$ is the number of residues, and a Gaussian distribution for end-to-end distances. However, they show conformational preferences at the residue level, and thus do not satisfy Flory’s model for a polymer chain, where each monomer is randomly oriented with respect to its neighboring monomers. This of course is by no means contradictory. For instance, Fiztkee and Rose [103] modified proteins of known structures by varying backbone torsion angles for ~8% of the residues, while the remaining 92% stayed fixed in their native conformations. By generating ensembles of these “disordered” structures, they were able to recover the typical random-coil statistics. This is simply because the residual structure can be grouped into Kuhn segments, and a chain of Kuhn segments can be treated as a random coil.

In experiments, the use in NMR of chemical shifts, scalar couplings and residual dipolar couplings has been particularly useful to find residual and/or transient secondary structure in both unfolded proteins and IDPs. Residual structures in the unfolded states introduce conformational biases that greatly enhance the folding of the protein [104,105]. In addition, many unfolded states seem to prefer PPII conformations [68], (PPI is not a state recognized by these codes, where PPII conformations are simply characterized as “coils”). Transient and in some cases even residual secondary structures are particularly present when IDPs become structured upon binding a partner [65], or when the IDPs are prone to aggregation [58,66,67]. Both population analysis and free energy landscapes [68–71] reveal that the conformational preferences are much less entropic than those of a random coil. For instance, a polyglutamine peptide in aqueous solution is intrinsically disordered, but it still exhibits relatively large percentages of secondary structure motifs [71]. Moreover, Q$_{80}$, a polyglutamine of 40 residues just above the threshold of $\sim$36 for Huntington’s disease, has been found to exhibit long-range (over 20 residues) structural correlations [71], that are not present in shorter, non-pathological polymer lengths. These correlations may underlie the aggregation phenomena related to polyglutamine diseases, and are destroyed when a C-terminal hexaproline is added to Q$_{40}$.

Given the difficulties to characterize the structure of IDPs experimentally, MD has become a valuable tool for this purpose. For instance, our recent work [72] on the gp41$_{659–671}$ peptide was spurred by its role as HIV antibody. However the original experimental data of the monomeric peptide in aqueous solution was highly contradictory. An early study based on NMR indicated that the conformation of the monomeric peptide in water was an amphiphilic $\bar{3}_1$ helix with minor random coil representation [73]. A second NMR study found no major population of $\bar{3}_1$ helical conformers, but a mixture of various conformers [74]. Later, a UV Resonance Raman Spectroscopy (UVRRS) investigation [76] concluded that there is a rough energy landscape with a wider variety of conformations than found in the NMR studies and previous Circular Dichroism (CD) studies [72]. The authors suggested that gp41$_{659–671}$ exhibits a broad distribution of conformations that includes significant population of $\beta$ turns, as well as $\bar{3}_1$ helix and $\pi$ helix motifs but little $\alpha$ helix. Recently, a far UV CD spectroscopy study [75] also revealed the conformational plasticity of gp41$_{659–671}$ with no stable $\alpha$ helical, $\bar{3}_1$ helical, or turn motifs. On the other hand, the crystal structure of the peptide bound to the ZF antibody shows an extended conformation [77,78]. MD simulations also showed varying results [79,75,80,72], due to different force fields and sometimes inadequate sampling [72]. In addition, the use of different methods to characterize the secondary structure motifs can further complicate comparisons and interpretation of the data.

In this work, we have compared DSSP, STRIDE and KAKSI as applied to mainly disordered gp41$_{659–671}$, N$_{14}$ and N$_{14–16}$ peptides. As expected, the agreement between the codes is considerably poorer than that for regular structures. The SOV scores between STRIDE and DSSP for the full peptide comparison, SOV$^{\beta}$ are 70.4% (gp41$_{659–671}$) and 53.8% (N$_{14}$) when STRIDE is taken as reference; and 62.2% (gp41$_{659–671}$) and 49.3% (N$_{14}$) when DSSP is taken as reference. For the dimers that show some measure of (non-ideal) $\beta$ sheet, SOV$^{\beta}$ scores are good for parallel dimers, in the range 90–98% when STRIDE is taken as reference (93–95% with DSSP as reference). However, for antiparallel dimers, for which STRIDE assigns considerably more turns and fewer strands, the SOV$^{\beta}$ scores are in the range 44–49% when STRIDE is taken as reference (46–52% with DSSP as reference). Segment length distributions are the same for $\bar{3}_1$ and $\pi$ helices, but are different for $\alpha$ helices and turns, due to the difference in definitions as discussed in the Results section. STRIDE and DSSP define more secondary structure elements than KAKSI, and thus can provide more detail when this information is needed. When it comes to distinguishing between $\bar{3}_1$ and $\alpha$ helices, DSSP performs better than STRIDE, as shown in Figs. 6–9. In particular, the DpSn ensembles (the subset of residues found in a given state by DSSP but not by STRIDE) correctly produce a helical rise distribution centered at 2 Å for $\bar{3}_1$ helices and at 1.5 Å for $\alpha$ helices; while the DnSp ensembles exhibit distributions centered at shorter helical rises for $\bar{3}_1$ helices and longer helical rises for $\alpha$ helices. Distributions for the dihedral angle sum $\psi + \phi$ of $\bar{3}_1$ also confirm a better assignment of $\bar{3}_1$ and $\alpha$ helices by DSSP. The largest source of differences is due to the definition of turns: in DSSP their definition is based on the hydrogen bonds, and include $\alpha$ turns, $\beta$ turns, $\gamma$ turns; in STRIDE only $\beta$ turns and $\gamma$ turns are recognized based on the distance $C_{\alpha} - \phi$ and the backbone dihedral angles ($\phi$, $\psi$). This results in a great disparity of these conformations: of all the turns defined by STRIDE, only approximately 40%
are recognized as "turns" by DSSP, while the remaining residues are classified as bends, coils and other structures.

Acknowledgments

We gratefully acknowledge help from Dr. Viet Hoang Man, especially for the preparation of the Nα→Nβ dimers. We thank NSF (Grants No. CHE-0845482 and No. 0822183) for funding. We also thank the NC State HPC Center for extensive computational support.

References


47
Amyloid properties of asparagine and glutamine in prion-like proteins


Copyright © 2016 American Chemical Society
Amyloid Properties of Asparagine and Glutamine in Prion-like Proteins

Yuan Zhang, Viet Hoang Man, Christopher Roland, and Celeste Sagui*

Department of Physics, and Center for High Performance Simulations (CHiPS), North Carolina State University, Raleigh, North Carolina 27695, United States

Supporting Information

Current evidence based on polyQ fibers and crystallites supports the presence of cross-β structures,15−20 common to other amyloid fibrils. PolyQ expansions occur at the DNA level as a consequence of the unstable expansion of CAG codons. By contrast, prion diseases can have a genetic, infectious, or sporadic origin that involves modification of the prion protein (PrP). This also results on fatal neurodegenerative diseases such as Creutzfeldt−Jakob (CJD) and Strausler−Scheinker diseases of humans, bovine spongiform encephalopathy (BSE), and scrapie of sheep. The cellular prion protein (PrP) is encoded by a gene, and a post-translational modification converts it into the infectious abnormal isoform, (PrPSc), which is rich in β-sheet content.21 Independently of whether the presence of PrPSc in the cell is the product of a post-translation misfold or whether it is an external infectious agent, PrPSc acts as a template upon which PrPSc further aggregates thus enabling the non-Mendelian inheritance characteristic of prions. Prions also cause non-Mendelian inheritance in yeast, which has become a very important model system for the study of prions. Aggregation-prone, Q/N rich domains are present in the baker’s yeast Saccharomyces cerevisiae, in proteins such as Sup35p and Ure2p, whose amyloid prions are [PSF] and [URE3], respectively. Although prions can cause disease in yeast, cells with certain

Received: December 18, 2015
Accepted: February 25, 2016

ABSTRACT: Sequences rich in glutamine (Q) and asparagine (N) are intrinsically disordered in monomeric form, but can aggregate into highly ordered amyloids, as seen in Q/N-rich prion domains (PrDs). Amyloids are fibrillar protein aggregates rich in β-sheet structures that can self-propagate through protein-conformational chain reactions. Here, we present a comprehensive theoretical study of N/Q-rich peptides, including sequences found in the yeast Sup35 PrD, in parallel and antiparallel β-sheet aggregates, and probe via fully atomic molecular dynamics simulations all their possible steric-zipper interfaces in order to determine their protofibril structure and their relative stability. Our results show that polyglutamine aggregates are more stable than polyasparagine aggregates. Enthalpic contributions to the free energy favor the formation of polyQ protofibrils, while entropic contributions favor the formation of polyN protofibrils. The considerably larger phase space that disordered polyQ must sample on its way to aggregation probably is at the root of the associated slower kinetics observed experimentally. When other amino acids are present, such as in the Sup35 PrD, their shorter side chains favor steric-zipper formation for N but not Q, as they preclude the in-register association of the long Q side chains.

KEYWORDS: Prion, amyloid, protein aggregation, intrinsically disordered proteins, polyglutamine disease

1. INTRODUCTION

Eukaryotic proteomes are rich in “low-complexity” sequences, which are distinguished by the prevalence of certain amino acids.1−5 These sequences lack well-defined three-dimensional structures and thus are under-represented in the RCSB PDB.6 In general, these sequences are quite abundant (14% of the proteins of Homo sapiens include at least one simple sequence1) and involve either single amino-acid repeats or a few different amino acids. Variations in the length of the sequences are a source of quantitative genetic variation in evolution but can also cause severe diseases in humans. In particular, sequences rich in glutamine (Q) and asparagine (N) are intrinsically disordered in monomeric form,7−9 but can aggregate into highly ordered amyloids.10 Amyloids are fibrillar protein aggregates rich in β-sheet structures that can self-propagate through protein-conformational chain reactions. Pure polyglutamine (polyQ) stretches in humans are associated with at least nine late-onset progressive neurodegenerative diseases, which are caused by polyQ expansions greater than a given threshold in proteins with little or no similarity except for the polyQ region. For instance, in Huntington’s disease, the normal polyQ length in Huntingtin (Htt) is 10−34 repeats, and pathological lengths are 36−120 repeats. Although each disease has a different pathology, they all share some common feature such as the formation of polyQ aggregates,10 and eventual neuronal death. Many structural models have been proposed for polyQ aggregates, such as α-helical coiled coils,11 β-helices,12 β-sheets,13,14 and β-sheets.15 It is not completely clear whether these models are the same or different, and there is little understanding of how these models relate to each other. However, recent studies have shown that the aggregation of polyQ peptides is facilitated by the presence of their shorter polyN counterparts.16−19 In this work, we focus on the case of polyN and polyQ stretches, which are distinguished by the prevalence of certain amino acids.20 These sequences lack well-defined structures and thus are under-represented in the RCSB PDB.3 In particular, sequences rich in glutamine (Q) and asparagine (N) are intrinsically disordered in monomeric form,5 although each disease has a different pathology. Here, we present a comprehensive theoretical study of N/Q-rich peptides, including sequences found in the yeast Sup35 PrD, in parallel and antiparallel β-sheet aggregates, and probe via fully atomic molecular dynamics simulations all their possible steric-zipper interfaces in order to determine their protofibril structure and their relative stability. Our results show that polyglutamine aggregates are more stable than polyasparagine aggregates. Enthalpic contributions to the free energy favor the formation of polyQ protofibrils, while entropic contributions favor the formation of polyN protofibrils. The considerably larger phase space that disordered polyQ must sample on its way to aggregation probably is at the root of the associated slower kinetics observed experimentally. When other amino acids are present, such as in the Sup35 PrD, their shorter side chains favor steric-zipper formation for N but not Q, as they preclude the in-register association of the long Q side chains.
prion forms have been shown to survive better under adverse conditions, which supports the view that at least some prions developed under positive evolutionary selection. In this sense, fungal prions belong to the category of functional amyloids, such as those found for instance in biofilms in bacteria: HET-s fibrils of Podospora anserina, required by heterokaryon incompatibility; the chaperonin from the bacterium Streptomyces coelicolor, associated with fimbriae formation; as well as curli fibrils of Escherichia coli and Salmonella enterica with adhesive function.

One important research direction in the characterization of prions involves the investigation of the relative roles of the polar, uncharged residues Q and N in Q/N-rich yeast prions. As an example, a recent study found opposing effects of Q and N in the formation of the Sup35 prion-forming domain (PrD). The experiments found that if all the Ns in the wild-type Sup35 PrD were switched to Qs, then amyloid formation decreased but toxicity to the cell increased. Conversely, switching all Qs to Ns enhanced benign, self-templating amyloid formation and decreased cell toxicity. The correlation between the lack of amyloid formation and increased toxicity for polyQ sequences is supported by a large body of literature in neurodegenerative polyQ diseases, where the formation of aggregates and toxicity in neuronal cells are not believed to be linked as cause-effect as was once thought. Instead, it is believed that the formation of large aggregates represents a cellular protective response against further toxicity, and that it is the soluble monomeric or oligomeric intermediates that are toxic to the cell.

A complete understanding of the mechanisms of polyQ toxicity still eludes us. However, one can address structural, thermodynamic and kinetic issues related to the amyloids themselves; for instance, why would N-rich sequences form amyloid-like prions while Q-rich sequences would not? The results obtained for the Sup35 PrD were unexpected because it was previously thought Qs and Ns are equivalent in terms of prion formation, and because algorithms for recognizing amyloidogenic sequences neither provide a distinction for the effects of Q and N switching nor predict amyloid formation for the Q/N-rich PrD. Since these experiments did not have atomic resolution, Monte Carlo simulations were employed in search for an atomistic explanation of these results. The results of the simulations of Q4 and N4 homopolymers suggested that the N-rich sequences increase the formation of β-hairpin turns and thus of β-sheets, which in turn would lead to the formation of the N-rich amyloids. A more recent study employs several biophysical tools to characterize and compare the aggregation kinetics of N4 and Q4 peptides. The main finding is that the kinetics of aggregation of N4 is much faster than that associated with Q4.

In this work, we seek to study and compare the fibril structural characteristics of Asn and Gln amino acids. We carry out molecular dynamics (MD) simulations that complement and integrate previous experimental and simulation results on N/Q-rich sequences. Particularly, our work presents the first comprehensive theoretical study of all the possible interfaces, structures and relative stability of the important N/Q-rich peptides (including sequences found in the yeast Sup35 prion domain), as they form amyloid-like protofibrils. We explore all possible steric zipper interfaces for N-rich and Q-rich oligomers in order to determine which crystallographic class provides the most stability to the N-rich or Q-rich fibril precursor. Our observations can be summarized as follows. First, different crystal classes maximize stability in N-rich or Q-rich aggregates. By comparing the most stable (in their optimum class) fibril precursor aggregates, we show that polyQ aggregates are more stable than polyserylamine (polyN) aggregates. Second, the presence of β-hairpin turns in polyN leading to β-sheets not necessarily explains the observed trend of polyN faster aggregation, since N-rich sequences tend to favor parallel β-sheets, as opposed to antiparallel ones, and β-hairpin turns are more stable in polyQ than in polyN. In particular, tight hairpin turns in polyQ have been recently observed experimentally.

In previous work, we have measured a non-negligible comparable population of loose turns both in disordered polyQ and in disordered polyN, and other simulations have also found polyQ-β-hairpins. Third, enthalpic contributions to the free energy favor the formation of polyQ protofibrils, while entropic contributions favor the formation of polyN protofibrils. Fourth, when other amino acids are present, such as in the Sup35 PrD, their shorter side chains favor steric-zipper formation for N but not Q as they preclude the in-register association of the long Q side chains.

2. RESULTS AND DISCUSSION

2.1. Initial Structures. The purpose of our investigations is to contrast the differing roles played by Gln and Asn in the formation of intraneuron protein aggregates. We have therefore investigated the stability of various aggregate models that are compatible with the experimental data, specifically as it pertains to the yeast prion protein Sup35. Specifically, the steric interfaces characterizing the amyloids are based on parallel or antiparallel cross-linked β-sheets with various crystal symmetries and side-chain interdigititation. The structures considered therefore represent a generalization of the experimental structures (as originally proposed, for instance, in ref 15), accounting for all possible symmetries and the possible presence of different amino acids. We also note that we have previously investigated polyQ structures with larger number of residues (i.e., polyQs with arcs and turns) with results that are in good agreement with the experimental studies. To understand our results, it is important that we first discuss the different models considered, and the nomenclature that is used to describe them. The models investigated are listed in Table 1, and a description of their construction is to be found in the Supporting Information (SI). Model 2010 in Table 1 is then used to generate 29 different models (all in “class A”, see below) mainly with mutations of one or two Qs into Gs.

The models introduced here are based on a set of M × N oligomeric crystals, with integer M representing the number of strands in a sheet and integer N the number of sheets in a given oligomer. The simplest structures are considered to be M × 1 homopolymers Ace-Gln-N-Me and Ace-Asn-N-Me, corresponding to an antiparallel β-sheet connected by a “hairpin”.

The majority of the structures considered are based on 4 × 2 and 2 × 2 model aggregates, which fall into two categories each (4T, 4O and 2T, 2O series) based on their side-chain packing. The aggregate structure itself is based on the classification proposed by Sawaya et al., who proposed eight classes of rigid steric zippers according to the relative positions of the β strands and sheets. Aggregates are classified according to whether (i) sheets are made from parallel or antiparallel strands; (ii) the adjacent sheets are packed with the same (“face-to-face”) or reverse (“face-to-back”) surfaces; and (iii) the symmetry operation used to generate the second sheet from the first preserves the order of the strands, such that the upper and lower strands in one sheet are still in that order (“up-up”) in the second sheet, or inverts it (“up-down”). A schematic of the eight
possible classes of aggregate models is shown in Figure 1. In addition to these considerations, one can consider two more criteria based on the type of side-chain packing, as shown in Figure 2c. In the “1-by-1” packing the side chains of two adjacent sheets intercalate with each other “one-by-one” as in a true zipper, while in the “2-by-2” packing two side chains intercalate every second set of side chains. The “One-by-One” packing gives rise to the 4O and 2O sets of models, while the “two-by-two” packing is associated with the 4T and 2T sets.

The number of steric zipper class reduces considerably when (not too short) homopeptides are considered, as shown in Figure 2. In this case, the “face-to-face” and “face-to-back” structures are the same for the parallel sheets, and the first four classes of Sawaya et al.35 are reduced to two: aggregates with parallel-stranded sheets in which the sheets are either parallel (class 2 and class 3) or antiparallel (class 1 and class 4) with respect to each other. In class 1 = 4, the carboxamide dipoles in polyN or polyQ are aligned at the interface between the two sheets, while in class 2 = 3, the dipoles in one sheet point in the opposite direction to those in the other sheet. Concerning steric zippers in antiparallel sheets (classes 5 to 8), for (not too short) homopeptides classes 6 to 8 reduce to a single structure, where the strands facing each other in opposite sheets point in the same direction (i.e., the two sheets are stacked in a parallel way), while class 5 (which so far has not been observed experimentally) corresponds to two sheets stacked in an antiparallel way. Furthermore, antiparallel-stranded sheets in polyQ tend to shift with respect to each other such that the strands of different sheets lie on top of each other, but “in-between” each other in a zipperlike fashion. This is known as “quarter-stagger” in β-sheet crystals, and it has been observed both experimentally and in simulations.16,20 When this happens, all of Sawaya et al.35 steric zippers involving antiparallel-stranded sheets (classes 5 to 8) reduce to a single structure, which for convenience we here label as class A (for antiparallel sheets).

We now turn to the specific models we investigated, which are listed in Table 1. First, we constructed four models (labeled 4T1 to 4T4) based on the X-ray crystal structure of the GNNQQNY peptide (PDB ID 1YJP).35 The model 4T1 corresponds exactly to the 1YJP crystal structure with eight GNNQQNY peptides forming two parallel sheets with strands separated by 4.87 Å in each sheet. The sheets therefore reproduce the class 1 steric zipper with “2-by-2” side chain packing in a dry interface with an 8.5 Å separation between the sheets. To study the effects of Q versus N in the formation of Sup35 PrD, we then mutated all the Qs to Ns to obtain 4T2 which was constructed from GNNNNNY strands. Likewise, we also mutated the Ns to Qs to obtain the 4T4 structure, which for convenience we here label as class A (for antiparallel sheets).
structure 4T5, corresponding to the 2OL9 crystal structure with eight SNQNNF peptides forming two parallel sheets with strands separated by 4.79 Å in each sheet. The structure corresponds to class 2 of Sawaya et al.35

Figure 2. For not too short homopeptides, some of the classes described in Figure 1 converge. (a) For parallel sheets, class 1 and 4 become the same (with sheets stacked in antiparallel fashion), and class 2 and 3 become the same (with sheets stacked in parallel fashion). (b) Schematic of the quarter-stagger displacement of one sheet with respect to the other. For antiparallel sheets undergoing a quarter-stagger, class 5–8 converge into class A (right). (c) Side-chain interdigitation of steric zippers.

We also considered aggregates based on the “1-by-1” side-chain packing (the 4O series), obtained from pure polyQ and polyN strands consisting of seven residues so as to be commensurate with the 4T series. Although for longer peptides class 1 coincides with 3, and class 2 coincides with 4, we specifically packed the seven-residue peptides in the four separate symmetries in order to account for possible end effects due to the relative short length of the peptides. Finally, we constructed models 4O9 and 4O10, based on polyN and polyQ strands respectively, arranged in an antiparallel fashion with “1-by-1” side-chain packing.

In order to gain further insight into the stability of these structures, we constructed a number of 2×2 models as listed in Table 1. Essentially, these structures were taken from the core strand (which are more stable) in the final configuration of the corresponding 4T or 4O models after a 400 ns simulation in explicit waters. In addition, there are 29 sequences in class A (not listed in Table 1) obtained from mutations in the sequence of 2010.

2.2. Monomeric PolyQ β-Hairpins Are More Stable than Monomeric PolyN β-Hairpins. Figure 3 shows sample initial and final structures (at 100 ns) for the homopolymers Q23 and N23, respectively, and averages are performed over 10 runs. After 100 ns, Q23 still exhibits considerable β-sheet structure while N23 has become a completely disordered coil. Also shown in Figure 3 is the residue-averaged β-sheet content as a function of time (middle), and the time-averaged β-sheet content as a function of the residue number (bottom). The data thus suggests that the Q23 hairpin is more stable than the N23 hairpin. This is an important result, and is in contradiction to the results given in ref 28, which deemed a N30 β-hairpin to be more well-defined (leading to a much better antiparallel β-sheet) than the corresponding Q-based structure. The previous simulation results28 were attributed to the intrinsic difficulty of polyQ sequences to form tight hairpin turns.

Previous Experimental and Simulation Data. Our results are in agreement with previous data. In particular, tight hairpin turns in polyQ have been recently observed experimentally.15,16 Also, tight β hairpins were proposed and used16 to successfully model the X-ray diffraction patterns obtained from crystalline fibers and films of a D3Q4K1 polypeptide.12 In our own work, we have measured a non-negligible comparable population of loose turns both in polyQ and in polyN,36 and other simulations have also found polyQ β hairpins.37 Recent simulations of a Q30 monomer (which is above the aggregation threshold for polyQ) containing two tight hairpins and an arc turn (“lateral β-sheet stack”) turned out to be remarkably stable in time scales longer than 1 μs.29
We also note that although the di-structure, with face-to-back and up-down packing between the sheets, are more stable for polyQ than for polyN.

2.3. N-Rich Oligomers Are Most Stable in “Class 1” Steric Zippers with 2-by-2 Interdigitation. Here we discuss the stability of Q-rich and N-rich models with 2-by-2 side chain interdigitation as schematically shown in Figures 1 and 2. Model 4T1 is based on the experimental PDB 1YJP crystal structure with sequence GNNQQNY, and 4T3 on the experimental 2OL9 crystal structure with sequence SNQNNF. Models 4T1 through 4T4 correspond to class 1 structures: parallel-stranded sheets with face-to-face and up-up packing between the sheets (Figure 1), while model 4T5 corresponds to class 2: parallel-stranded sheet, with face-to-back and up-up packing between the sheets. We also note that although the different 4 × 2 models are constructed from dry crystal interfaces as described, the actual simulations of the structures take place in a fully solvated environment.

Sample configurations for the 4 × 2 aggregates are shown in Figures 4 and 5, and sample configurations for the 2 × 2 aggregates are shown in Figure 6. Figures 4 and 5 show initial (after equilibration) and final configurations at 400 ns for each of the 4 × 2 models. Figure 4 shows that N-rich sequences in class 1 structures with 2-by-2 side-chain packing are more stable than the corresponding Q-rich structure. Thus, models 4T1, 4T2, and 4T3 all preserve considerable β-sheet structure after 400 ns, while model 4T4 is completely disordered. Quantitative plots of the time evolution of the structural characteristics of the different models, mainly RMSDs and β-sheet content, are presented in Figures 7 and 8, while intersheet and interstrand distances are given in the SI. Model 4T5 was found to be completely unstable in solution where it lacks the close crystal contacts that stabilize it in crystal environment. Because of the rapid unfolding of this structure in solution, we believe that similar class 2 structures based on pure polyQ and polyN will also be unstable. We would like to notice that, although 4T5 consists of six (as opposed to seven) residues, a 4T3 model with 6 residues per strand (not shown) is still relatively stable. The instability of this model may be the reason why the PDB structure 2OL9 (which forms the basis for model 4T3), appears to be the only class 2 structure with this kind of interdigitation in the PDB databank. More polyN and polyQ class 2 models were investigated in the 4O series, as described below.

Next, we compare the pure polyN structures given by the parallel sheets in the 4T3 and the antiparallel sheets in 4O9 (with a sample configuration shown in Figure 5). Statistically, the RMSDs shown in Figure 7 and the β-sheet contents shown in Figure 8 indicate that the parallel-stranded 4T3 model is more stable than the antiparallel-stranded 4O9 model. This observation is also supported by the intersheet and intrastrand distances shown in the SI, Figures S3 and S4. In both cases, the corresponding, smaller 2 × 2 aggregates are unstable.

Previous Experimental and Simulation Data. Our results are in agreement with previous data. The stability of the heptapeptide GNNQQNY in model 4T1 (PDB ID 1YJP) has been confirmed by previous simulations. In all cases, the in-register parallel packing of GNNQQNY is consistent with the observed X-ray diffraction data. It has also been pointed out that GNNQQNY exhibits the same amyloid properties as the full-length Sup35, as seen in the cooperative kinetics of aggregation, the binding to the dye Congo red, fibril development, and in the typical cross-β diffraction pattern of the structure. The surfaces of the two sheets are complementary resulting in a dry steric zipper with tightly interdigitated side chains. The zipper is stabilized by van der Waals interactions, and hydrogen bonds between the polar groups. Through the π stacking of the aromatic Tyr-Tyr interactions in GNNQQNY plays a role in...
stabilizing the parallel arrangement of the strands, the pure polyN model 4T3 also favors the formation of this kind of zipper structure. Previously, two single strands of polyN or polyQ homopeptides were shown to assemble in dimers that favor parallel (polyN) or antiparallel (polyQ) β-sheets. In addition, solid-state NMR experiments show that the amyloid of the full prion domain of Sup35p (residues 1−123) has an in-register parallel β-sheet structure.

In summary, N-rich sequences in 4 × 2 aggregates featuring class 1 interfaces with 2-by-2 side-chain packing display stability in the times of the simulations (400 ns). The solution simulation for the sequence in class 2 is not stable for the 4 × 2 model, and therefore, no further mutations are tested. For class 1, the Q-rich model is not stable.

2.4. PolyQ Oligomers Are Most Stable in Antiparallel-Stranded β-Sheets with 1-by-1 Steric Zippers. Next, we turn to structures based on pure polyQ and polyN peptides exhibiting 1-by-1 side-chain packing. As before, we begin our discussion with the 4 × 2 models, as shown in Figure 5. For these models, there are three classes of steric zippers that need to be considered: class 1 = 4 and class 2 = 3 based on a parallel arrangement of strands within a sheet, and class A based on an antiparallel arrangement. However, we specifically packed the peptides in the four separate symmetries in parallel-stranded sheets in order to account for possible end effects due to the relative short length of the peptides.

The results are as follows. First, all the parallel 1-by-1 side-chain classes are unstable for polyN. For polyQ structures, models in class 1 and 4 (4O2 and 4O4) are more stable than the corresponding models in class 2 and 3 (4O6 and 4O8) indicating that an antiparallel packing of parallel-stranded β-sheets is preferred. The final configurations of 4O2 (class 1) and 4O4 (class 4) and their RMSDs in Figure 7 and β-content in Figure 8 are very similar (and so are their intersheet and interstrand...
distances, shown in the SI), suggesting that the end effects are not important and indeed class 1 and class 4 are the same. A similar conclusion is obtained for 4O6 and 4O8, indicating that the end effects in these peptides play no role and class 1 is the same as class 3. Finally, polyQ structures with 1-by-1 side-chain packing and class A zippers (model 4O10) appear to be very stable. PolyN structures with 1-by-1 side-chain packing and class A zippers (model 4O9) also show larger stability than the 1-by-1 side-chain packing polyN parallel arrangements (4O1, 4O3, 4O5, 4O7). However, the class A polyN model 4O9 is less stable than the 2-by-2 side-chain packing parallel arrangement in 4T3.

Previous Experimental and Simulation Data. Our polyQ results are in complete agreement with current evidence based on polyQ fibers and crystallites supporting the presence of cross-β structures with antiparallel β-sheets.15−20

2.5. The Most Stable PolyQ Structures Show Higher Stability than the Most Stable PolyN Structures. At this point, the results for the 4 × 2 aggregates (at 400 ns) include the following relatively stable structures: polyN 2-by-2 class 1 structures (4T1, 4T2, and 4T3) and polyQ 1-by-1 class A structure (4O9); and three stable 1-by-1 polyQ structures: class 1 (4O2), class 2 (4O4), and class A (4O10). Figures 7 and 8 suggest that the polyQ structures are more stable. Recognizing that cooperative effects in larger aggregates make them more stable than smaller aggregates,26 one can build smaller aggregates with the same symmetry and test them for stability. Indeed, in a situation where peptides are undergoing a transition from a disordered state to an amyloid precursor, the higher stability of a small nucleus would confer a higher chance to the formation of the protofibril. Thus, we used the most stable 4 × 2 structures to build the corresponding 2 × 2 models. These are listed in Table 1 and shown in Figure 6 for a sample configuration, while the time evolution is shown in Figures 7 and 8.

The reduction in size causes the 2 × 2 aggregates for class 1 models with 2-by-2 side-chain packing (i.e., GNNQQNY in 2T1, GNNNNNY in 2T2, and NNNNNNN in 2T3) to quickly become unstable. So does the polyN structure in the antiparallel, class A, 1-by-1 model 4O9. A previous simulation of GNNQQNY oligomers also found that these were unstable when in 2 × 2 aggregates.43 For polyQ, the 1-by-1 class 1 (2O2) and class 4 (2O4) also become unstable. On the other hand, the 1-by-1 class A (2010) model is very stable.

All the models probed here have "reactive", rigid, edge strands with exposed main-chain hydrogen bond donors and acceptors which would drive the elongation of fibrils after a size threshold.44,45 This same reactivity makes all the small 2 × 2 oligomers unstable, except for 2010. This feature is peculiar to polyQ, which in the 1-by-1 class A model exhibits tightly interdigitated side chains, stabilized not only by dipole–dipole interactions, and backbone–backbone, interside–chain, and backbone–side-chain hydrogen bonds, but also by strong van der Waals interactions between the closely fitting side chains.16,20

Previous Experimental and Simulation Data. We are not aware of studies comparing the stability of polyN and polyQ amyloid-like aggregates.

Melting by a Free-Electron Laser Pulse. Recently, free electron lasers have been used for melting biomolecular complexes.46,47 Kawasaki and co-workers47 have developed a mid-infrared free electron laser with highly oscillation characteristics having a high photon density, a picosecond pulse

Figure 8. β-Sheet content of the 4 × 2 models 4T1 (black), 4T2 (red), 4T3 (blue), 4T4 (orange), and 4O9 (green) in top left chart; 4 × 2 models 4O1 (black), 4O2 (red), 4O3 (blue), and 4O4 (orange) in top right chart; 4 × 2 models 4O5 (black), 4O6 (red), 4O7 (blue), 4O8 (orange), and 4O10 (green) in bottom left chart; 2 × 2 models 2T1 (black), 2T2 (red), 2T3 (blue), 2O2 (orange), 2O4 (green), 2O9 (cyan), and 2O10 (magenta) in bottom right chart. Results are averaged over 10 independent runs.
structure, and a range of tunable frequencies. To date, such a laser pulse with frequencies tuned to the amide I bands has been used in experiments to dissociate amyloid-like fibrils of lysozyme into its native forms,\textsuperscript{47} convert insulin fibrils into soluble monomers, and dissociate a short human thyroid hormone peptide. A simulated laser pulse has also been used in atomistic molecular dynamics simulations to dissociate amyloid fibrils,\textsuperscript{48} and to study the breakup of a peptide-based nanotube.\textsuperscript{49}

In order to corroborate the results obtained for the $2 \times 2$ aggregates, we decided to apply the laser melting techniques to the most stable $4 \times 2$ homopeptide aggregates. For polyN, this is class 1 model 4T3; and for polyQ this is class A model 4O10. Simulations were carried out as outlined in the SI using a laser frequency of 1671 cm$^{-1}$, which specifically targets the C=O main-chain bonds, thereby destabilizing the β-sheets. The application of this laser pulse results in an almost identical energy absorption for the 4T3 and 4O10 aggregates, as shown in SIFigure S2. However, the structures respond very differently, as shown in Figure 9, which displays two measures of structural integrity: the root-mean-square deviation (RMSD) with respect to the initial, intact structure and the β-sheet content. The behavior of these functions reflects the initial disruption of the hydrogen bonds (to a greater or lesser degree) once the laser pulse is applied, followed by a period of partial healing to the initial structure. For instance, considering the β-sheet content, one notes an initial strong dip in its numerical value as the laser pulse is applied followed by a slower rise to a constant value as the structure attempts to heal itself. For each electric field amplitude, the initial dip and subsequent constant value is systematically lower for 4T3, the most stable polyN structure, than for 4O10, the most stable polyQ structure. Similarly, the RMSDs in Figure 9 are systematically higher for 4T3 than for 4O10. All together, these results indicate that polyQ aggregates are more stable than polyN aggregates.

2.6. The Enthalpic Contributions to the Free Energy Favor PolyQ over PolyN. One can obtain a fairly good estimate of the enthalpic contributions to the free energy of the protofibrils by computing the energy difference between the aggregate and the individual unfolded monomers: $\Delta H = H_{\text{crystal}} - 8H_{\text{monomer}}$, where $H_{\text{crystal}}$ is the enthalpy of the $4 \times 2$ homopeptide protofibril after equilibration, and $H_{\text{monomer}}$ is the average enthalpy of a statistical ensemble of an unfolded monomer, and both terms are computed with the same force field and specifications as described before, but with the Generalized Born approximation for solvation. This gives $\Delta H_{\text{Q}} = -813$ kcal/mol for polyQ 4O10 [approximately $-14.5$ kcal/(mol residue)] and $\Delta H_{\text{N}} = -249$ kcal/mol for polyN 4T3 [approximately $-4.4$ kcal/(mol residue)].

Clearly, the stability of a β-sheet structure not only depends on the main-chain hydrogen bonds but also on the side-chain packing, associated with both side-chain–side-chain and side-chain–main-chain hydrogen bonds. In order to illustrate the contributions to stability from a structural point of view, we have sketched in Figure 10 the structural differences between the stable models 4T1 (a), 4T3 (b) and 4O10 (c). In the 4T1 model (a), the side chains of two pairs of residues Q4-N6 of two neighbor chains (A and B or C and D) in the same sheet form hydrogen bonds and create stable polygon structures which support the overall stability of the sheet layer. Additionally, the NE2 atoms of Q2 residues form hydrogen bonds with the nearest main chains (side-chain–main-chain hydrogen bonds, linking the colored polygons in (a) to the main chains), which contributes to binding...
The only difference between N and Q is the extra methylene group $\text{CH}_2$ in the side chain of Q. It is expected therefore that the change in free energy to bring together two strands and form a $2 \times 1 \beta$ sheet (parallel for N, antiparallel for Q) might be comparable, and indeed a simulation study for a Q2 pair and a N2 pair shows very similar free energy profiles in terms of the distance between the strands33 (with the difference in free energies slightly favoring the formation of the Q2 pair). In the $2 \times 1 \beta$ dimers the side chains remain perpendicular to the sheet, free to form occasional hydrogen bonds with the residues or with water. More important differences for the stability of the oligomers may therefore arise from the formation of the steric zipper. Ultimately, as this work shows, polyQ aggregates are more stable than polyN aggregates. Strong enthalpic terms as described above favor the formation of the polyQ aggregates. What about the entropic contribution?

The entropic free energy differences associated with the formation of steric zippers in the two types of peptide can be quickly estimated by considering the degrees of freedom of the side chains (assuming that the backbones have similar degrees of freedom). Asn has 2 side chain torsion angles $\chi_1$ and $\chi_2$, and Gln has 3 side chain torsion angles $\chi_1, \chi_2, \chi_3$ and $\chi_4$. The $\{\chi_1\}$ of N and the $\{\chi_1, \chi_2, \chi_3\}$ of Q are defined as rotameric $\chi_s$'s, which exhibit three narrow and symmetric peaks in their density probability distribution. The terminal side-chain torsion angle $\chi_4$ of N and $\chi_4$ of Q is defined as nonrotameric $\chi_s$ which exhibits a broad and asymmetric probability distribution. The 2002 Rotamer Library90 defines the degrees of freedom of rotameric $\chi_s$ as 3 and nonrotameric $\chi_s$ as 6 for N and Q. The total number of rotamers for the side chains of N and Q therefore becomes 18 and 54 for N and Q in their unfolded states. Since in the aggregate form the side chains are locked into one conformer, the total number of rotamers in the prototefibril structure for both peptides is 1. Thus, the entropic loss between a side chain locked in a fibril and a free side chain can be estimated as 

$$\Delta S_N = S_{\text{fibril}} - S_{\text{free}} = -R \ln \left( \frac{N_{\text{chains}}}{N_{\text{rotamers}}} \right) \quad \text{and} \quad \Delta S_Q = S_{\text{fibril}} - S_{\text{free}} = -R \ln \left( \frac{N_{\text{chains}}}{N_{\text{rotamers}}} \right)$$

At room temperature (300 K) this gives a difference in entropy favoring the disordered state of $\Delta G_N = -T \Delta S_N = 2.4 \text{ kcal/(mol residue)}$ and $\Delta G_Q = -T \Delta S_Q = 1.73 \text{ kcal/(mol residue)}$. (This estimate does not consider the degrees of freedom of the backbones, whose contribution to the free energies is probably similar for polyN and polyQ.)

Thus, entropic effects are less favorable to the polyQ steric zippers, as disordered polyQ chains have a more extensive phase space to sample (for $n_q$ residues, the ratio of phase spaces between polyQ and polyN grows as $3^n$) before settling in the preferred conformation. There are probably considerably larger entropic barriers associated with polyQ, which would explain the very long times for the formation of polyQ amyloids observed experimentally.$^{80,91}$

### 2.8. Other Interspersed Amino Acids Create More Disruption in PolyQ Zippers.

Finally, we briefly discuss the effects of other amino acid guests on the stability of the considered structures. Although a systematic study of the effect of every possible amino acid in the N-rich and Q-rich peptides is beyond the scope of this work, there are some observed trends that merit discussion. Aggregates 4T1 and 4T2 indicate that the seven-residue N-rich peptides are tolerant of the terminal G and Y residues, as well as of two inner Q residues. We believe that the 2-by-2 side-chain packing facilitates the accommodation of these residues in the N-rich strands. However, a similar analysis for the Q-rich strands is not possible, as the Q-rich four-residue N-rich peptide does not form a stable steric zipper.

The only difference between N and Q is the extra methylene group $\text{CH}_2$ in the side chain of Q. It is expected therefore that the change in free energy to bring together two strands and form a $2 \times 1 \beta$ sheet (parallel for N, antiparallel for Q) might be comparable, and indeed a simulation study for a Q2 pair and a N2 pair shows very similar free energy profiles in terms of the distance between the strands$^{33}$ (with the difference in free energies slightly favoring the formation of the Q2 pair). In the $2 \times 1 \beta$ dimers the side chains remain perpendicular to the sheet, free to form occasional hydrogen bonds with the residues or with water. More important differences for the stability of the oligomers may therefore arise from the formation of the steric zipper. Ultimately, as this work shows, polyQ aggregates are more stable than polyN aggregates. Strong enthalpic terms as described above favor the formation of the polyQ aggregates. What about the entropic contribution?

The entropic free energy differences associated with the formation of steric zippers in the two types of peptide can be quickly estimated by considering the degrees of freedom of the side chains (assuming that the backbones have similar degrees of freedom). Asn has 2 side chain torsion angles $\chi_1$ and $\chi_2$, and Gln has 3 side chain torsion angles $\chi_1, \chi_2, \chi_3$ and $\chi_4$. The $\{\chi_1\}$ of N and the $\{\chi_1, \chi_2, \chi_3\}$ of Q are defined as rotameric $\chi_s$'s, which exhibit three narrow and symmetric peaks in their density probability distribution. The terminal side-chain torsion angle $\chi_4$ of N and $\chi_4$ of Q is defined as nonrotameric $\chi_s$ which exhibits a broad and asymmetric probability distribution. The 2002 Rotamer Library$^{90}$ defines the degrees of freedom of rotameric $\chi_s$ as 3 and nonrotameric $\chi_s$ as 6 for N and Q. The total number of rotamers for the side chains of N and Q therefore becomes 18 and 54 for N and Q in their unfolded states. Since in the aggregate form the side chains are locked into one conformer, the total number of rotamers in the prototefibril structure for both peptides is 1. Thus, the entropic loss between a side chain locked in a fibril and a free side chain can be estimated as 

$$\Delta S_N = S_{\text{fibril}} - S_{\text{free}} = -R \ln \left( \frac{N_{\text{chains}}}{N_{\text{rotamers}}} \right) \quad \text{and} \quad \Delta S_Q = S_{\text{fibril}} - S_{\text{free}} = -R \ln \left( \frac{N_{\text{chains}}}{N_{\text{rotamers}}} \right)$$

At room temperature (300 K) this gives a difference in entropy favoring the disordered state of $\Delta G_N = -T \Delta S_N = 2.4 \text{ kcal/(mol residue)}$ and $\Delta G_Q = -T \Delta S_Q = 1.73 \text{ kcal/(mol residue)}$. (This estimate does not consider the degrees of freedom of the backbones, whose contribution to the free energies is probably similar for polyN and polyQ.)

Thus, entropic effects are less favorable to the polyQ steric zippers, as disordered polyQ chains have a more extensive phase space to sample (for $n_q$ residues, the ratio of phase spaces between polyQ and polyN grows as $3^n$) before settling in the preferred conformation. There are probably considerably larger entropic barriers associated with polyQ, which would explain the very long times for the formation of polyQ amyloids observed experimentally.$^{80,91}$

### 2.8. Other Interspersed Amino Acids Create More Disruption in PolyQ Zippers.

Finally, we briefly discuss the effects of other amino acid guests on the stability of the considered structures. Although a systematic study of the effect of every possible amino acid in the N-rich and Q-rich peptides is beyond the scope of this work, there are some observed trends that merit discussion. Aggregates 4T1 and 4T2 indicate that the seven-residue N-rich peptides are tolerant of the terminal G and Y residues, as well as of two inner Q residues. We believe that the 2-by-2 side-chain packing facilitates the accommodation of these residues in the N-rich strands. However, a similar analysis for the Q-rich strands is not possible, as the Q-rich four-residue N-rich peptide does not form a stable steric zipper.
other amino acids. However, Q-rich peptides prefer 1-by-1 side-chain packing. Thus, $4 \times 2$ parallel arrangements such as 4O2 and 4O4 for Q7 are stable, but the $4 \times 2$ parallel model 4T4 for GQ5Y is not. On the other hand, the presence of the terminal G and Y in GQ5Y only slightly diminishes the $\beta$-sheet content with respect to Q7 (Figure 11) when the peptides are in class A, which indicates that the structures may be more tolerant of terminal substitutions.

Since polyQ prefers a class A steric zipper, we investigated the effects of G substitutions in the 2010 polyQ aggregates, in order to probe the effect of a guest amino acid on the stability of these structures. Figure 11 shows that single G residues only moderately decrease the $\beta$-sheet content with respect to Q7, in a position-dependent manner. On the other hand, two G residues cause considerable disruption that strongly depends on the position of these residues.

PolyQ fibrils are characterized by the carefully balanced stacking of the side chains\(^{16,20}\) that maximizes the dipolar, hydrogen-bond, and van der Waals contacts. Thus, the presence of other amino acids will generally disrupt these contacts. The sensitivity of polyQ amyloid formation to the presence of other amino acids is displayed even in less intuitive cases. For instance, flanking sequences play a structural role in pure polyQ sequences, both in synthetic and natural peptides. A polyproline (polyP) region immediately adjacent to the C-terminal of a polyQ region has been shown to affect the conformation of the polyQ region; the resulting conformations depend on the lengths of both the polyQ and polyP sequences.\(^{17}\) Even for monomeric, disordered Q\(_{40}\) a C-terminal hexaproline suppresses both the population of “rare motifs” and the long-range correlation of the dihedral angles present in Q\(_{40}\) without the hexaproline.\(^{7}\)

3. CONCLUSIONS

In summary, we have carried out the first comprehensive theoretical study based on MD simulations of all the possible steric interfaces of important N/Q-rich peptides forming amyloids, especially sequences as found in the yeast Sup35 prion domain. We have compared the structural characteristics and thermodynamic stability of these protofibril aggregates in prion-like proteins. Specifically, we explored all possible steric zipper interfaces for N-rich and Q-rich oligomers in order to determine which crystallographic class provides the most stability to the N-rich or Q-rich fibril precursor. Our observations can be summarized as follows. First, different crystal classes maximize stability in N-rich or Q-rich aggregates: N-rich oligomers are most stable in parallel-stranded $\beta$-sheets with $2 \times 2$ side-chain interdigitation (model 4T3), while polyQ oligomers are most stable in antiparallel-stranded $\beta$-sheets with 1-by-1 steric zippers (model 4O10). Our results show that the polyQ 4O10 model is more stable than the polyN 4T3 model. Enthalpic contributions to the free energy favor the formation of polyQ protofibrils, while entropic contributions favor the formation of polyN protofibrils. The presence of $\beta$-hairpin turns in polyN leading to $\beta$-sheets not necessarily explains the observed trend of polyN faster aggregation, since N-rich sequences tend to favor parallel $\beta$-sheets, as opposed to antiparallel ones, and $\beta$-hairpin turns are more stable in polyQ than in polyN. The considerably larger phase space that disordered polyQ must sample on its way to aggregation probably is at the root of the associated slower kinetics observed experimentally. When other amino acids are present, such as in the Sup35 PrD, their shorter side chains favor steric-zipper formation for N but not Q, as they preclude the in-register association of the long Q side chains.
4. METHODS

4.1. Molecular Dynamics Simulations. The simulations were carried out with the AMBER 12 simulation package with the ff12SB force field. The TIP3P water model was used for the explicit solvent simulations with periodic boundary conditions in orthorhombic boxes (more than 7000 waters for the hairpins, 6000 for the 4 x 2 systems and 4500 for the 2 x 2 systems). The construction of the initial models and the equilibration procedures are described in the SI. Electrostatics were handled by the PME method, with a direct space cutoff of 9 Å, and with an average mesh size of approximately 1 Å for the lattice calculations. We used Langevin dynamics with a coupling parameter γ = 1.0 ps⁻¹. The NPT simulations were carried out mainly at 300 K and 1 atm via the Berendsen barostat with an isothermal compressibility of β = 44.6 × 10⁻¹² bar⁻¹ and pressure relaxation time τp = 1.0 ps. The length of the simulation was 100 ns for the hairpins and 2 x 2 systems and 400 ns for the 4 x 2 systems. Finally, for each of the systems we carried out 10 independent runs. We tracked the RMSDs with respect to the initial structures (after equilibration) and the β-sheet content of the aggregates as a function of time. We also tracked the interstrand and internsheet distances as a function of time. The definition of these two distances as well as the corresponding results are presented in the SI.

4.2. Laser Melting Simulations. While most of our conclusions are based on MD simulations in explicit waters, we have carried out a limited number of nonequilibrium MD simulations in which the structures were subject to a laser pulse with an electric field:

\[ E(t) = E_0 \exp \left[ \frac{(t - t_p)^2}{2\sigma^2} \right] \cos[2\pi(cot(t - t_p))] \]

where \( E_0 \) indicates the modules of the electric field amplitude, \( \sigma \) the pulse width, \( t \) the time after the pulse maximum \( t_p \), \( c \), the speed of light, and \( \omega \) the frequency. Such a classical laser pulse is characteristic of a free-electron based laser⁶⁷ with specific oscillation characteristics of a picosecond pulse structure, tunable wavelengths within the infrared regime, and a high photon density. Using this laser pulse, it has been possible to target specific bonds within a given structure and thereby study the unfolding of the structure. Specifically, this method was used to investigate the dissociation of amyloid fibrils into soluble monomers,⁵⁷ and the breakup of a peptide-based nanotube.⁵⁰ Here, we use such nonequilibrium MD simulations simply as an alternative way of exploring the relative stability of a given set of structures. Thus, by tuning the laser frequency to target key bonds that hold the structure together, the structure can be disrupted. According to the field intensity, the broken bonds may or may not reconnect. The application of such a laser pulse causes localized damage to the stable structure, given that all other conditions and parameters are equal. For our purposes, we use these laser pulse simulations simply as a tool to examine the relative stability of the polyn and polyQ most stable models (4T3 and 4O10), and do not focus on the microscopics of the laser melting. We emphasize here that these laser melting simulations are not needed per se, but are given because they lend additional support to stability results as obtained with the other MD simulations. Details of these laser pulse simulations and their analysis are relegated to the SI.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acschemneuro.5b00337.

Further details on simulations and analysis, frequency dependence of structures, infrared laser energy differences, interstrand and internsheet distances for models, RMSD for models, and β-content for different models (PDF)

AUTHOR INFORMATION

Corresponding Author

E-mail: sagui@ncsu.edu. Phone: 919-515-3111.

Author Contributions

Y.Z. and V.H.M. ran simulations; Y.Z. and V.H.M. analyzed data; all authors helped formulate problem and wrote paper.

Funding

This work has been supported by the National Science Foundation (NSF) via SI2-SSE-1148144 and 1534941.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We also thank the NC State HPC and especially the XSEDE TG-MCB150114 for computational support.

REFERENCES

exchange molecular dynamics simulations. J. Mol. Model. 19, 1627−
1639.

(38) Nelson, R., Sawaya, M. R., Balbirnie, M., Madsen, A. O., Riekel, C.,
Grothe, R., and Eisenberg, D. (2005) Structure of the cross-β spine of

side-chain interactions in the early steps of aggregation: Molecular
 dynamics simulations of an amyloid-forming peptide from the yeast

Stability of single sheet GNNQNY aggregates analyzed by replica
 exchange molecular dynamics: antiparallel versus parallel association.

(41) Balbirnie, M., Grothe, R., and Eisenberg, D. S. (2001) An amyloid-
forming peptide from the yeast prion Sup35 reveals a dehydrated β-

prion domain of Sup35 has an in register parallel β-sheet structure.

Insights into stability and toxicity of amyloid-like oligomers by replica

 proteins use negative design to avoid edge-to-edge aggregation. Proc.

change and assembly through edge strands in transthyretin and

(46) Yagi, H., Ozawa, D., Sakurai, K., Kawakami, T., Kuyama, H.,
Nishimura, O., Shimanouchi, T., Kubo, R., Naiki, H., and Goto, Y.
(2010) Laser-induced propagation and destruction of amyloid beta

Effect of mid-infrared free-electron laser irradiation on refolding of
amyloid-like fibrils of Lysozyme into native form. Protein J. 31, 710.

(48) Man, V. H., Derreuxaux, P., Li, M., Roland, C., Sagui, C., and
Nguyen, P. (2015) Picosecond dissociation of amyloid fibrils with
infrared laser: A nonequilibrium simulation study. J. Chem. Phys. 143,
155101.

(49) Man, V., Truong, P., Derreuxaux, P., Li, M., Roland, C., Sagui, C.,
and Nguyen, P. (2015) Picosecond Melting of Peptide Nanotube with
Infrared Laser: A Nonequilibrium Study. Phys. Chem. Chem. Phys. 17,
27275−27280.


(51) Darnell, G. D., Derryberry, J., Kurutz, J. W., and Meredith, S. C.
(2009) Mechanism of Cis-inhibition of polyQ fibrillation by polyP. PPII
oligomers and the hydrophobic effect. Biophys. J. 97, 2295−2305.

(52) Case, D. A. et al. AMBER 12; University of California: San
Francisco; 2012.

Phys. 103, 8577−8593.
Supporting information for: Amyloid properties of asparagine and glutamine in prion-like proteins

Yuan Zhang,†,‡ Viet Hoang Man,†,‡ Christopher Roland,†,‡ and Celeste Sagui∗,†,‡

†Department of Physics, North Carolina State University, Raleigh, NC 27695
‡Center for High Performance Simulations (CHiPS), North Carolina State University, Raleigh, NC 27695

E-mail: sagui@ncsu.edu

1 MD Simulations

The simulations were carried out with the AMBER 12 simulation package(1) with the ff12SB force field(2). The TIP3P water model(3) was used for the explicit solvent simulations with periodic boundary conditions in orthorhombic boxes. We used Langevin dynamics with a coupling parameter $\gamma = 1.0 \text{ ps}^{-1}$. Electrostatics were handled by the PME method (4), with a direct space cutoff of 9 Å, and with an average mesh size of approximately 1 Å for the lattice calculations. The constant pressure simulations were carried out at 1 atm via the Berendsen barostat(5) with an isothermal compressibility of $\beta = 44.6 \times 10^{-6} \text{ bar}^{-1}$ and pressure relaxation time $\tau_p = 1.0 \text{ ps}$.

We discuss the construction of the different models and give more details on the specific simulations associated with each of them. We begin with a consideration of the $Q_{23}$ and $N_{23}$ hairpin structures. First, the initial oligomers of one $Q_7$ was obtained from an 1 μs of a 4×4 polyQ steric zipper with antiparallel sheets(6). The $Q_7$ was then expanded to a $Q_{11}$ strand by adding N- or C-terminal residues. For a C-expansion, we considered a vector...
that connected the C atoms of the 6th and 8th residues, and used this vector to translate a
copy of the 7th residue and add it to the C-terminus of the Q7 strand. A similar definition
(using the 3rd and 1st residues to define the corresponding vector) was used to add a copy
of the 2nd residue to the N-terminus of the strand. We added one residue at the N-terminus
and two residues at the C-terminus. Then, a suitable Q-residue in a coil or turn structure
was picked out from a 100 ns molecular dynamics simulation of a monomeric Q40 system.
The two atoms to be connected were the C atom at the C-terminus of the first peptide(C1)
and the N atom at the N-terminus of the second peptide(N2), whose distance was 4.1-4.6
Å. Given that the peptide bond distance is in the range 1.32-1.33 Å, the distance between
the N and C atoms of the added residue should be larger than 2.0 Å. The residue was then
translated and rotated so that it fitted into the available space without any bond stretching,
with the side chain on a plane pointing outward away from the steric zipper. The hairpin
Ace-N23-Nme was obtained by mutating the Q side chains of Q23 to ones corresponding
to N23. To investigate the stability of these hairpin structures, they were placed in a box
containing more 7000 water molecules.

Turning to the amyloid-like fibrils, we note that the 4T models are based on the exper-
imental PDB 1YJP and 2OL9 structures as already described. The 4O models were built
from Q7 and N7 homopeptides using the LEAP program of the AMBER v.12(1) simulation
package. First, an unfolded Q7 peptide with all torsion angles equal to 180° was generated
with the ff12SB force field. Then, the backbond dihedral angles of consecutive monomers
were rotated using the PyMOL package(7) to values (-120°, 115°) for parallel and (-140°,
135°) for antiparallel conformations. Then the peptide was rotated so that its main chain
was oriented along the z-axis with side chains in the y-z plane. Two Q7 monomers were then
brought together such that the peptide axis lies along the z-axis and the hydrogen bonds
along the x-axis (i.e., the β sheet is then in the x-z plane) and the side chains are perpen-
dicular to the β-sheet plane along the y-axis. The corresponding polyN model was formed
by mutating Q to N. In each of these four models, the parallel sheets are separated by 8.34
Å. The side chain packing is “1-by-1”. These 4 × 2 (2 × 2) structures were then placed in orthorhombic boxes with 6000 (4500) waters, respectively.

The equilibration processes took place in four steps. First, we carried out a steepest descent simulation followed by conjugate gradient minimization keeping the peptide atoms fixed at their initial positions. Then we carried out unrestrained steepest descent followed by conjugate gradient minimizations. This was followed by short molecular dynamics runs under constant volume while the system was heated from 0K to 300K with weak restraints on the peptide atoms. After this, Langevin dynamics at constant temperature (300 K) and constant pressure (1 atm) were applied for 2 ns, after which the density of the system was found to be stable around 1.0g/cm³. Finally, choosing different conformations from the equilibration runs, we set up ten independent NPT Langevin simulation runs for each of the aggregate groups. The hairpin and the 2×2 systems were run for 100 ns, and the 4×2 systems were run for 400 ns. System coordinates were sampled every 2 ps. Figures were generated by VMD(8).

2 Laser Simulations

The laser simulations were carried out in a manner similar to previous work(9, 10). All simulations were carried out with the GROMACS 4.5.5 package(11) using the AMBER ff12SB force field(1) and the TIP3P water model. Long range electrostatic interactions were computed with the Particle-Mesh Ewald summation method (12). The non-bonded interaction pair list was updated every 10 fs with the cutoff of 1 nm. All covalent bonds were constrained by the LINCS algorithm (13) with a relative geometrical tolerance of 10⁻⁴. The V-rescale temperature coupling, which uses velocity rescaling with a stochastic term (14), has been used to couple the system to a heat bath. To ensure stability, a time step of 0.2 fs was used to integrate the equations of motion with the leapfrog algorithm (15).

The first step in carrying out the laser pulse simulations involves determining the correct
frequency to apply to the system so as to target the correct hydrogen bonds. To this end, we first carried out simulations involving a frequency scan. In the laser-induced simulations, a time-dependent electric field

\[ E(t) = E_0 \exp \left[ -\frac{(t - t_0)^2}{2\sigma^2} \right] \cos[2\pi c \omega (t - t_0)], \]  

(1)

was applied to mimic a laser pulse\( (16, 17) \). Here, \( E_0 \) represents the amplitude of the electric field, \( \sigma \) is the pulse width, \( t \) is the time after the pulse maximum \( t_0 \), \( c \) is the speed of light and \( \omega \) is the frequency. For these frequency scan simulations, we set the pulse width to \( \sigma = 2 \) ps, \( t_0 = 5 \) ps, the strength \( E_0 = 1 \) V/nm and allowed \( \omega \) to vary from 1630 to 1730 cm\(^{-1} \) in steps of 2 cm\(^{-1} \). For each \( \omega \) value, nine 16 ps laser-induced simulations (where the starting conformations were randomly selected from the first nanosecond of an equilibrium simulation) were performed. The data for each individual \( \omega \) value was then averaged over the 9 copies. Figure 1 shows data for these frequency scans for the 4T3 and 4O10 models.

From the frequency scan simulations, we selected 1671 cm\(^{-1} \) as the optimum frequency for targeting the C=O bonds. For the final runs, laser parameters were set to: \( \sigma = 2 \) ps, \( t_0 = 5 \) ps, \( \omega = 1671 \) cm\(^{-1} \); and the magnitude of the electric field strength was set to \( E_0 = 1.8, 2.0, 2.2, 2.4 \) V/nm. For each field strength, we performed 20 independent simulations 50 ps long, in which the starting structures were randomly selected from the first nanosecond of the equilibrium simulation. All the data was then averaged. Figure 2 shows the decay of the energy differences for the 4O10 and 4T3 models.

3 Analysis

Our structural analysis made use of the DSSP\( (18) \) package for the secondary structural assignments, noting that different secondary structure assignment codes can give different results for conformationally irregular peptides\( (19) \). Also we tracked the root-mean-square deviation (rmsd) with respect to the initial structures as a function of time. We also calcu-
lated the following: (i) the intersheet distance, which describes the distance between sheets in the steric zippers and is defined as the average distance of the center of mass (com) between two facing strands over all the strands and sheets; and (ii) the interstrand distance, which describes the distance between the strands on the same sheet and is defined to be the average distance of the center of mass between two residues in adjacent strands within the same sheet averaged over all the residues. Formulae for these are:

\[
\langle D_{\text{intersheet}} \rangle = \frac{\sum_{i} \sum_{j} (r_{i,j}^{\text{com}} - r_{i+1,j}^{\text{com}})}{N_{\text{strands}} \cdot (N_{\text{sheets}} - 1)}
\]

\[
\langle D_{\text{interstrand}} \rangle = \frac{\sum_{i} \sum_{j} \sum_{k} (r_{i,j,k}^{\text{com}} - r_{i,j+1,k}^{\text{com}})}{N_{\text{sheets}} \cdot (N_{\text{strands}} - 1) \cdot N_{\text{residues}}},
\]

with \(N_{\text{strands}}\), \(N_{\text{sheets}}\), and \(N_{\text{residues}}\) representing the number of strands, sheets and residues, respectively.
Figure S1: The frequency dependence of (a) maximum of the C=O bond length fluctuations; (b) kinetic energy difference; (c) total energy difference, and (d) potential energy difference. Energy differences are taken with respect to the corresponding energy average value as obtained from an equilibrium nanosecond simulation.
Figure S2: Energy differences (with respect to the equilibrium energy value) of laser pulse simulations as a function of different electric field strengths $E_0$: (a) 1.8; (b) 2.0; (c) 2.2; and (d) 2.4 V/nm, respectively.
Figure S3: Intersheet distance for each model. The values are averaged every 2 ns for $4 \times 2$ models, and every 0.2 ns for $2 \times 2$ models, and then averaged over 10 runs.
Figure S4: Interstrand distance for each model. The values are averaged every 2 ns for 4×2 models, and every 0.2 ns for 2×2 models, and then averaged over 10 runs.
Figure S5: RMSD (top), β-content (2nd), intersheet distance (3rd) and interstrand distance (bottom) of final step simulations with their error bars on hairpin systems (last 2 ns simulation, blue), 4×2 models T (last 2 ns, orange) and models O (last 2 ns, green), as well as 2×2 models (last 0.2 ns, magenta).
References


Chapter 5

Structure and dynamics of DNA and RNA double helices obtained from the GGGGCC and CCCCGG hexanucleotide repeats that are the hallmark of C9FTD/ALS diseases

Yuan Zhang, Christopher Roland, and Celeste Sagui. Structure and dynamics of DNA and RNA double helices obtained from the GGGGCC and CCCCGG hexanucleotide repeats that are the hallmark of C9FTD/ALS diseases. Under review by *ACS Chem. Neurosci.*, 2016.
ABSTRACT

A (GGGGCC) hexanucleotide repeat (HR) expansion in the C9ORF72 gene, and its associated antisense (CCCGGG) expansion, are considered the major cause behind frontotemporal dementia and amyotrophic lateral sclerosis. We have performed molecular dynamics simulations to characterize the conformation and dynamics of the twelve duplexes that result from the three different reading frames in sense and antisense HRs for both DNA and RNA. These duplexes display atypical structures relevant not only for a molecular level understanding of these diseases but also for enlarging the repertoire of nucleic-acid structural motifs. G-rich helices share common features. The inner G-G mismatches stay inside the helix in G\textsubscript{syn}-G\textsubscript{anti} conformations and form two hydrogen bonds (HBs) between the Watson-Crick edge of G\textsubscript{anti} and the Hoogsteen edge of G\textsubscript{syn}. In addition, G\textsubscript{syn} in RNA forms a base-phosphate HB. Inner G-G mismatches cause local unwinding of the helix. G-rich double helices are more stable than C-rich helices due to better stacking and HBs of G-G mismatches. C-rich helix conformations vary wildly. C mismatches flip out of the helix in DNA but not in RNA. Least (most) stable C-rich RNA and DNA helices have single (double) mismatches separated by two (four) Watson-Crick basepairs. The most stable DNA structure displays an “e-motif” where mismatched bases flip towards the minor groove and point in the 5’ direction. There are two RNA conformations, where the orientation and HB pattern of the mismatches is coupled to bending of the helix. Ion distributions and ion bridges that stabilize conformations are described.

5.1 Introduction

Frontotemporal dementia (FTD) and amyotrophic lateral sclerosis (ALS) are two neurodegenerative diseases with similar genetic and neurological pathways. FTD is the most common cause of early-onset dementia due to degeneration of the frontal and anterior temporal lobes, while ALS is characterized by progressive muscle weakness and paralysis due to loss of motor neurons in the brain and the spinal cord. These diseases can occur simultaneously, and are believed to be part of the same spectrum [1]. A (GGGGCC) hexanucleotide repeat (HR) expansion in the first intron of the C9ORF72 gene has been shown to be the major cause behind both FTD (here specifically indicated as C9FTD) and ALS [2,3]. While the unaffected population carries fewer than 20 repeats (generally no
more than a couple), large expansions greater than 70 repeats and usually encompassing 250-1600 repeats have been found in C9FTD and ALS patients.

Nucleotide repeat disorders can cause toxicity through different but non-exclusive mechanisms. First, the expansions originate in the DNA itself, and these expansions can alter the local chromatin structure, affecting RNA transcription and protein translation in the gene. Second, transcribed RNA can cause gain and/or loss of function. The transcribed introns containing these large expansions seem to contribute to neuropathology both through loss of function, as mRNA levels of C9ORF72 are decreased in C9FTD/ALS patients [2,3]; and through gain of function, as RNA transcripts containing the (GGGGCC) HRs are accumulated in nuclear foci in the frontal cortex and spinal cord, leading to the sequestration of RNA-binding proteins [2]. Contributing to the complexity of the pathological mechanisms, there is also evidence that antisense transcripts of the expansion, i.e., (CCCCCGG) expanded repeats resulting from the bidirectional transcription of the DNA HR expansions, also form nuclear RNA foci [4, 5]. Third, translated repeats can also cause toxicity in the corresponding protein and its interaction partners. Even though the hexanucleotide expansions reside in a non-coding region of the C9ORF72 gene, it has been shown that these expansions can trigger protein translation in the absence of the start ATG codon, giving rise to the unconventional repeat-associated non-ATG (RAN) translation [5–8]. RAN translation of the (GGGGCC) expansion can lead to Gly-Ala, Gly-Pro, and Gly-Arg poly-dipeptide expansions, while RAN translation of the antisense (CCCCCGG) expansion can lead to Pro-Gly, Pro-Ala, and Pro-Arg poly-dipeptide expansions. Generically, these are known as “C9RAN” proteins, and have been detected in C9FTD/ALS patients.

For simple-sequence nucleotide repeat diseases, an important breakthrough has been the recognition that stable atypical DNA secondary structure in the expanded repeats is “a common and causative factor for expansion in human disease” [9]. Chemical and enzymatic probing of a r(GGGGCC) repeat expansion points to a general scenario where the repeat expansion adopts a hairpin structure with G-G mismatches in equilibrium with a quadruplex structure [10]. Recently, CD, optical melting and 1D 1H NMR spectroscopy, combined with chemical and enzymatic analysis reveal that r(GGGGCC) expansions
fold into a G-quadruplex in equilibrium with hairpin structures [11]. The equilibrium is temperature dependent, with \( T = 37^\circ \) favoring hairpins and higher annealing temperatures favoring quadruplexes. The equilibrium is also controlled by the type of ion (with K\(^+\) ions favoring G-quadruplexes and Na\(^+\) ions favoring hairpins), and ionic strength [11].

In this paper, we present an extensive and detailed analysis *in silico* of the structure and dynamics of all possible DNA and RNA duplexes that can be formed from the GGGGCC sense and CCCCGG antisense HRs. First, we carefully generate the possible duplexes that can be obtained by shifting the reading frame in the HRs. This results therefore in three duplexes for each G-rich or C-rich DNA or RNA helical duplex, for a total of twelve duplexes. Then we carry out 1 \( \mu \)s classical molecular dynamics simulations to investigate the global conformational space of dodecamers and associated dynamics; characterize the local conformation of the mismatches and the ion distribution and binding; and when possible, compare the relative stability of the double helices.

### 5.2 Results

**Initial structures**

We start with ideal B-DNA and ideal A-RNA. In order to achieve a more comprehensive structural characterization at the atomic level, we specifically consider all the possible linear duplexes built from sense GGGGCC sequences with G-G mismatches and antisense CCCCGG sequences with C-C mismatches in all possible reading frames. A scheme for 2 repeats is shown in Figure 5.1. In order to identify the independent structures, it is necessary to consider the pattern of steps. To illustrate this, consider a duplex formed by two strands of two HRs each with sequence 5’-GGGGCC-GGGGCC-3’. Now define “steps” as steps with Watson-Crick bonding, \( A=\text{GG/CC=CC/GG}, \ D=\text{CG/C=CG/GG}, \ E=\text{GC/CG=CG/GG}, \ \text{or steps that include mismatches, m}_1=\text{GG/GC=GC/GG}, \ m_2=\text{GG/CG=CG/GG}, \ \text{and M=GG/GG}. \) The step or stacking pattern in this sequence becomes \( \text{Am}_1\text{Mm}_1\text{AADam}_1\text{Mm}_1\text{A} \), and one can see that it repeats periodically. Now consider the sequence 5’-GCCGGG-GCCGGG-3’. The step pattern in this case becomes \( \text{m}_1\text{ADAm}_1\text{Mm}_1\text{ADAm}_1 \). For long sequences, this is exactly the same pattern as before (same reading frame), and the only differences are simply due to end effects. Thus, both sequences give rise to the same type of helix, shown as duplex “DG-2” (DG stands
for double G mismatch) in the scheme in Figure 5.1. Now consider the sequence 5’-GGGCCG-GGGCCG-3’ as shown in duplex “DG-1” in Figure 5.1. The step pattern is $m_2AEm_2Mm_2AEAm_2$, which is different from the previous two. All in all, neglecting end effects, there are three independent duplexes for each sequence, as shown in Figure 5.1. The third type, SG-3 (SC-3) stands for single G-G (C-C) mismatch (of course, the number of mismatches is the same in all cases).

In addition to these duplexes, we created two more duplexes for testing. After 1 µs, we found that the duplex DNA SC-3 became unstable in both sets of simulations: in set A, the DNA helix slowly unfolded while in set B, the second strand shifted one base, transforming it into DNA DC-1. We hypothesized that this behavior reflected a lesser stability for DNA SC-3, and decided to test this in three ways: (i) re-running SC-3 with different random seed; (ii) adding a small restrain (1 kcal/(mol Å)) between bases C2 and G24 as well as G12 and C14 (we call the new restrained SC-3, DNA R-SC-3); (iii) adding a C at the 3’ end of first strand and the 5’ end of second strand to form a single mismatching model without hanging bases (Duplex SC-3+C). Finally, all the duplex structures were built by using the NAB and LEAP programs in AMBER v.14, and Na$^+$ ions were added for neutralization. A summary of all these models is presented in Table 5.1.

<table>
<thead>
<tr>
<th>structure</th>
<th>model name</th>
<th>sequence</th>
<th>arrangement</th>
<th>form</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duplex</td>
<td>DG-1</td>
<td>(GGGCCG)$_2$</td>
<td>antiparallel</td>
<td>DNA,RNA</td>
</tr>
<tr>
<td></td>
<td>DG-2</td>
<td>(GCCGGG)$_2$</td>
<td>antiparallel</td>
<td>DNA,RNA</td>
</tr>
<tr>
<td></td>
<td>SG-3</td>
<td>H-(GGGGCC)$_2$</td>
<td>antiparallel</td>
<td>DNA,RNA</td>
</tr>
<tr>
<td></td>
<td>DC-1</td>
<td>(CCCCGGC)$_2$</td>
<td>antiparallel</td>
<td>DNA,RNA</td>
</tr>
<tr>
<td></td>
<td>DC-2</td>
<td>(CGGCCC)$_2$</td>
<td>antiparallel</td>
<td>DNA,RNA</td>
</tr>
<tr>
<td></td>
<td>SC-3</td>
<td>H-(CCCCCGG)$_2$</td>
<td>antiparallel</td>
<td>DNA,RNA</td>
</tr>
<tr>
<td></td>
<td>R-SC-3</td>
<td>H-(CCCCCGG)$_2$</td>
<td>antiparallel</td>
<td>DNA</td>
</tr>
<tr>
<td>(special)</td>
<td>SC-3+C</td>
<td>(CCCCCGG)$_2$-C</td>
<td>antiparallel</td>
<td>DNA</td>
</tr>
</tbody>
</table>

One final consideration for defining the initial structures is the initial glycosidic angle $\chi$ of the mismatched base pairs. Cytosine tends to always favor the anti state, even
Figure 5.1 Structures of duplex configurations. Gs are in blue and Cs in red.
in C-C mismatches [12] (unless it is a terminal mismatch, where it can rotate more freely). Guanine, on the other hand, can rotate around the glycosidic bond adopting the syn conformation. G_{syn}-C_{anti} Watson-Crick base pairs are observed in both Z-DNA and Z-RNA duplexes [13–17]. More importantly, G_{syn}-G_{anti} mismatches have been observed in the crystal structures of CGG RNA repeats [18,19]. There is, however, a considerable energy barrier for this rotation [17], and we have chosen to insert the right symmetry in the initial structures. Thus, we have set all the initial C-C mismatches in the anti-anti conformation. For DNA and RNA G-rich duplexes, we have chosen G_{syn}-G_{anti} mismatches alternating so as to preserve the symmetry of the strands (and therefore, the symmetry of the steps as described above). This is possible in duplex types DG-1 and DG-2, that exhibit neighboring G-G mismatches. In model SG-3, an anti/syn distribution that preserves the symmetry of the strands results in alternating G_{anti}-G_{anti} and G_{syn}-G_{syn} mismatches, which are energetically less favorable. We have therefore chosen for the initial structures alternating G_{syn}-G_{anti} mismatches, even though the initial symmetry of the strands is broken.

**G- and C-rich duplexes**

Figure 5.2 shows the distribution of conformations over the second half of the simulation projected onto the first three principal components obtained from the PCA [20], which correspond to the largest positional deviation and most of the positional fluctuations of the DNA and RNA duplexes. The G-rich structures in both DNA and RNA (DG-1, DG-2 and SG-3) in Figure 5.2 show Gaussian distributions corresponding to independent harmonic motions that have converged. For the DNA models, the first principal component corresponds to bending around the central hinge and the second, corresponds to bending in DG-1, but to groove breathing in DG-2 and SG-3. For RNA, the first component corresponds to bending, while the second component also corresponds to bending in DG-1 and DG-2, but to groove breathing in SG-3.

In contrast to the G-rich duplexes, non-Gaussian distributions are observed for C-rich duplexes. In particular, for DNA, we find that DC-1 is Gaussian-distributed, with the first two modes corresponding to groove breathing motions. However, DC-2 is not: the first component corresponding to bending exhibits considerable spread, and the second component corresponding to groove breathing has a dominant and a secondary peak. To track the origin of the non-Gaussian distribution, we plotted the time behavior of the first
Figure 5.2  PCA distributions of G-G (1st column) and C-C (4th column) mismatched duplexes over the 500 ns - 1 µs interval projected onto the first three principal components for DNA models (a) and RNA models (b). And the corresponding visualization of the atomic motions involved in the 1st (blue) and 2nd (red) principal components.
Figure 5.2 shows that DNA DC-2 exhibits an end effect at later times where the 3’ end of one of the strands (C24 here) bends over and interacts with the C6 base on the other strand, which has flipped out. This end effect also considerably distorts the main chain and is reflected in the non-Gaussian distribution of the second principal component. The three different versions of SC-3: SC-3, shown in Figure 5.2, and R-SC-3 and SC-3+C, shown in Figure 5.S3 all display a bimodal distribution for the first principal component, but are Gaussian-distributed for the next two components. The time behavior of DNA SC-3 (Figure 5.S4) shows two main conformations: In both cases the helical structure is unwound at the top, but the later one (820-1000 ns) is also unfolding at the bottom. Due to the restraint at the terminal bases, DNA R-SC-3 strives to maintain a relative stable structure during the 1 μs simulation, facing different stages of unwinding at both ends (Figures 5.S3 and 5.S5). Similar behavior is observed in DNA SC-3+C. In Figure 5.S6 we show a state (500-880 ns) corresponding to a half unwound structure and a later state (880-1000 ns) where the duplex has healed its helical structure. All in all, the data for the three versions of DNA SC-3 seem to indicate oscillations between unfolding and partially refolding events.

The C-rich RNA duplexes in Figure 5.2 show the following behavior: SC-3 is Gaussian-distributed while DC-1 and DC-2 are not. Tracking the time behavior (Figure 5.S7), we see that there are two conformations in DC-1: In conformation A, described in more detail below, the C6-C19 mismatches stay inside the helix and occasionally form a hydrogen bond, while the C7 and C18 bases point towards the minor groove, facing each other. In this case, the major groove is less occupied and the duplex is bent towards the major groove. In conformation B, C7 and C19 stay inside the helix, stacking on top of each other, while the complementary mismatches C18 and C6 form an N3-N4 cross hydrogen bond and turn to the major groove, unbending the duplex (see also Figure 5.S10). Both cases seem to correspond to equilibrium distributions, as the C6 and C18 bases slowly turn towards the minor and major groove, slightly bending or unbending the duplex. Finally, RNA DC-2 stays only in conformation B, with a small amplitude oscillation motion (Figure 5.S8) of the mismatched bases, which results in the perfectly bimodal distribution in Figure 5.2.

These general results can be further quantified by analyzing the behavior of backbone sugar torsion δ and the glycosyl torsion χ, as well as the “open angle”, as defined in the SI. Figure 5.3 shows the χ and δ torsion angle distributions of duplex models in DNA form.
and RNA form (right) corresponding to internal G-G mismatches (top three panels) and internal C-C mismatches (bottom three panels; end mismatched bases are excluded). The salient features of these graphs are that the distributions associated with RNA are relatively narrow, while there is considerably dispersion with those associated with DNA. For the G-G mismatches, the two distributions associated with G<sub>syn</sub> and G<sub>anti</sub> are very distinct. The bases in the C-C mismatches are all in some form of anti conformation.

The time evolution graphics of the χ torsion angle for the internal mismatches for the DC-1 and DC-2 forms of DNA and RNA are shown in Figure 5.S9. For the DNA DC-2 model, the distributions are centered around -106° and -130°, corresponding to anti-anti configurations. For the DNA DC-1 model, the χ angles of C6-C19 are Gaussian distributed at -102° and -80°, which correspond to anti and high anti configurations. The pair C7-C18, on the other hand, has a bimodal distribution with values around -112° and -68° for C7; and -120° and -70° for C18. This reflects the presence of the “e-motif”, described below, where the C7 and C18 bases flip out into the minor groove. The distributions for the RNA forms are centered around -160° (corresponding to anti periplanar conformations).

The open angle (see definition in SI) distributions of DNA and RNA are shown in Figure 5.4. Both C and G bases that participate in Watson-Crick hydrogen bonds have a Gaussian distribution around 60°. Distributions for G-G mismatches are centered at 30-35° and 75°, indicative of flipping towards minor and major grooves, respectively. Distributions for C-C mismatches are spread-out and generally multimodal, reflecting the rotating activity of these bases. The time evolution of the open angle for the mismatches and the associated conformations are described below.

Now that we have an overall view of the main structural features of the duplexes and the time behavior of the first principal component and glycosyl angle χ, both shown in the SI, we can concentrate on more particular structural features. First, we consider the single C (SC-3) mismatch models. For DNA, as discussed previously, the various conformations associated with the non-Gaussian peaks in the first principal component distribution (shown in Figures 5.2 and 5.S3) suggest that the various versions of SC-3 are not as stable as those for DC-1 and DC-2. Figure 5.5 shows this in more detail. In the “set A” simulations (corresponding to the data presented here, see the Methods section) we see partial unfolding of the structure: (i) the base C1 at the 5’ end of one strand bends over and interacts with the base moiety of C4; (ii) the mismatched C10 base in the same strand flips into the major groove; (iii) Watson-Crick pairs G11-C15 and G12-C14 have been
Figure 5.3 Backbone sugar torsion $\delta$ and the glycosyl torsion $\chi$ angle distributions for DNA duplexes (left) and RNA duplexes (right) corresponding to G mismatches (top three rows) and C mismatches (bottom three rows).
Figure 5.4  Open angle distribution of G mismatching models (left two columns) and C mismatching models (right two columns) in DNA (odds columns) and RNA (even columns) forms. Black and red lines show the distribution of G and C bases, respectively, that participate in Watson-Crick bonding. The green line shows the distribution for the internal mismatched bases (G or C).
broken; (iv) ending bases C13 and C14 shift towards the 3’-5’ direction, and bend and get trapped in the minor groove. In the “set B” simulations (corresponding to a slightly less extensive equilibration process), one of the strands shifts, transforming the duplex into a partial DC-1 structure with reconstitution of four successive Watson-Crick hydrogen bonds, as shown in Figure 5.5. For both sets of simulations, SC-3+C shows clear signs of unwinding. Finally the “restricted” simulation R-SC-3 shows to be relatively stable for set A but mutates into DC-1 for set B. Figure 5.5 compares the late structures of R-SC-3 and SC-3+C for set A. Interestingly, the RNA SC-3 duplex—in spite of exhibiting a perfect Gaussian distribution for the first principal component, with a very narrow distribution in the $\chi-\delta$ map—also undergoes hydrogen bond shifting, as explained below.

**Figure 5.5** Simulations results for DNA SC-3. (a) This figure shows the DNA SC-3 evolution in the simulation set A from 0 ns (after equilibration) to 1 $\mu$s (final simulation time). First, one strand shifts by one residue in the 3’-5’ direction, then the free ending bases bond to the Watson-Crick paired bases, causing the breaking of the hydrogen bonds. (b) In the simulation set B, one strand shifts by one residue in the 3’-5’ direction, transforming the duplex into a partial DC-1 structure. The flipping out of a C basis allows the retention of the SC-3 configuration in the second half of the duplex.
Next, we consider the double C mismatches, DC-1 and DC-2, which are stable up to 1 µs for both DNA and RNA (except for end effect fluctuations). Let us first consider DNA. Figure 5.6 shows the time evolution of the open angle for the internal mismatches in DC-1 and DC-2 respectively, along with associated conformations. The internal mismatches correspond to the pairs C6-C19 and C7-C18. In the DC-1 structure in Figure 5.6 (a,b), the mismatched bases C7 and C18 start to turn towards the minor groove early in the simulation: around 160 ns, C7 flips out and towards the 5’ direction in the strand, interacting with G5; and C18 flips out and towards the 5’ direction of its own strand at about 300 ns, interacting with G16. Both G5-C18 and C18-G16 form G(N3)-C(N4) hydrogen bonds. This gives rise to the rarely seen “e-motif”, where the C bases (i residue) in a mismatch symmetrically flip out in the minor groove, pointing their base moieties in the direction of the i – 2 residue (i.e., towards the 5’ direction in each strand), as shown in Figure 5.6 (a,b). Once formed, the “e-motif” is stable from 300 ns to the end of the simulation at 1 µs. The other pair, C6-C19, initially shows some flipping activity (C19 turning towards the minor groove and C6 turning towards the major groove), but after 300 ns they both settle inside the helical structure. Now we turn to DNA DC-2 in Figure 5.6 (c,d). In this case, the C7-C18 mismatch mainly stays inside the helix and the conformation of the C6-C19 mismatch evolves slowly. During the first 560 ns, C6-C19 and C7-C18 either form an internal pair or slightly turn to the major and minor grooves as shown in Figure 5.6(d). After 560 ns, the bases exhibit different behavior. During 560-820 ns, C6 turns towards the minor groove and C19 flips out into the major groove. During the last 180 ns, C6 flips completely out into the minor groove, and upwards towards the 5’ direction of its strand, while C19 stays in the major groove. During 560-1000 ns, C18 still stays inside the helix, while C7 turns towards the major groove forming an O2-N4 hydrogen bond with C18 (Figure 5.11). Finally, a comparison between the two DNA DC duplexes shows that the e-motif structure is better stacked, more compact than the structure in DNA DC-2. This can be observed, for instance, in Figure 5.6 and can be quantified by measuring the distance $d_i$ between neighboring layers for each base pair (i.e., for base pair $i$, we compute the distances between base pairs $i - 1$ and $i + 1$). This distance is mainly the same for layers 3, 4, 5, 9, and 10, but is shorter around the mismatches for layers 6, 7 and 8 for DC-1, which is a measure of the better packing of the e-motif. (The average distances for DC-1 are $d_6 = 5.15\text{Å}$, $d_7 = 3.13\text{Å}$, and $d_8 = 4.90\text{Å}$; and for DC-2 are $d_6 = 6.52\text{Å}$, $d_7 = 7.11\text{Å}$, and $d_8 = 7.81\text{Å}$.)
Figure 5.6  (a) Open angle of DNA DC-1 as a function of time (left) and corresponding distribution (right). (b) The e-motif configuration. (c) Open angle of DNA DC-2 as a function of time (left) and corresponding distribution (right). (d) Evolution of the C6-C19 mismatch (sideviews) and top view of the C6-C19 evolution.
Here, we make some observations with respect to the flipping of the bases. An analysis of all the trajectories for C-rich DNA shows that when a C base completes its flipping into the minor groove, it folds back and interacts with bases towards the 5' direction on its own strand. This directionality has also been observed in careful MD studies of base flipping with conformational flooding (Figure 7(a) in Ref. [21]). On the other hand, when the C base flips towards the major groove, the bases have more space and therefore are more flexible, interacting less with the rest of the helix. We have plotted the backbone dihedral angles as a function of the open angle and found that most angles do not give a clear indication of the transition but instead reflect increased fluctuations and transitions. Indeed, same values of the open angle can correspond to completely different values of the backbone torsion angles. However, by plotting the backbone torsion angles as a function of time, we were able to pick up two distinctive trends for flipping into the major versus the minor grooves. Figure 5.7 shows the torsion angles $\alpha$ and $\gamma$ for the C7-C18 mismatch in DNA DC-1. While most backbone angles do not reflect the transition, $\alpha$ decays at the transition points, and then it assumes the characteristic anticorrelation with $\gamma$, where $\alpha + \gamma$ stays constant. Alternatively, if the sum $\alpha + \gamma$ is depicted, it decreases approximately by 100°, when the corresponding base, C7 at around 160 ns and C18 at around 300 ns, flips out. The torsions $\epsilon$ and $\zeta$ also show small jump, synchronized transitions at these times, with constant difference $\epsilon - \zeta \simeq -90°$, corresponding to a BI backbone conformation for this particular mismatch. For the other mismatch, C6-C19, $\alpha$ and $\gamma$ stay constant, but the C6 angles $\epsilon$, defined as the C4'-C3'-O3'-P(i+1) torsion; and $\zeta$, defined as the C3'-O3'-P(i+1)-O5'(i+1) torsion, show an increase at 160 ns from $\epsilon = 188°$ to $\epsilon = 290°$, and a decrease from $\zeta = 277°$ to $\zeta = 81°$, clearly due to the coupling of these angles to the configuration of C7. For C6, this represents a transition from $\epsilon - \zeta \simeq -90$ (BI) to $\epsilon - \zeta \simeq 210$. The flipping out into the major groove, as occurs with C19 after 560 ns in DNA DC-2, is signaled by an increase in $\epsilon$ and a decrease in $\zeta$: $\epsilon$ increases from $\epsilon = 188°$ to $\epsilon = 283°$ and $\zeta$ decreases from $\zeta = 277°$ to $\zeta = 81°$ ($\epsilon - \zeta \simeq -90$ to $\epsilon - \zeta \simeq 200$). C18 also turns towards the major groove but it never flips out: its $\epsilon$ angle increases from $\epsilon = 193°$ to $\epsilon = 283°$, but $\zeta$ stays the same.

Now, we consider the RNA C-rich duplexes. Typical conformations are depicted in Figures 5.8 and 5.9. There is a clear coupling between mismatches that shuttle between major and minor grooves and bending of RNA. Figure 5.S10 shows the open angle of C18 superposed to the radius of gyration ($R_g$) of the DC-1 RNA duplex as a function of time.
Figure 5.7  (a) Torsion angle $\alpha$ (black) and $\gamma$ (red) of C7 and C18 as a function of time for DC-1 in DNA form. Sum of torsion angles ($\alpha + \gamma$) for C7 (blue) and C18 (orange) as a function of time for DC-1 in DNA form. (b) Torsion angle $\epsilon$ (black) and $\zeta$ (red) of C18 and C19 as a function of time for DC-2 in DNA form. Difference of torsion angles ($\epsilon - \zeta$) for C18 (blue) and C19 (orange) as a function of time for DC-2 in DNA form.
(bending is hard to characterize for such a short duplex, so we show radius of gyration instead). We observe two main mismatch conformations in RNA DC-1 (labeled A and B in Figure 5.8). During the 0-180 ns, 310-600 ns and 880-1000 ns time intervals, DC-1 is under configuration A, where C6-C19 mismatches stay inside the helix and occasionally form a N3-N4 hydrogen bond (Figure 5.11), while the C7 and C18 bases point towards the minor groove, facing each other with the open angle of C7 and C18 dropping to 0°. Since both C7 and C18 are in the minor groove in conformation A, the major groove is less occupied and bases i − 1 and i + 1 can interact pulling the backbones closer so that the duplex exhibits a smaller \( R_g \), or is bent, as shown in Figure 5.S10. In conformation B, both C7 and C19 stay inside the helix stacked together, but C6 and C18 turn to the major groove forming an O2-N4 hydrogen bond (Figure 5.11) in their corresponding C6-C19 and C7-C18 mismatches. In this conformation, the occupation of the major groove by C6 and C18 pushes the strands away, unbending the duplex, which adopts a slightly larger \( R_g \) value. In the RNA DC-2 duplex, only conformation B is observed, as shown in Figure 5.8, with C6 and C18 in the major groove. Figure 5.S10 shows that the radius of gyration stays at a constant, “unbent” value of approximately 14 Å (there is an abnormal \( R_g \) value from 375 ns to 500 ns corresponding to one end bending towards the main chain; the end effect disappears after 500 ns).

For the RNA SC-3 duplex shown in Figure 5.9, the C4-C22 and C7-C19 pair bases favor a synchronized oscillation motion between minor and major grooves, where if one base of the mismatched pair turns towards the minor groove, the other turns towards the major one. The pair C10-C16 is under the same oscillation motion before 450 ns, but after that, the Watson-Crick hydrogen bonds in base pair G11-C15 and G12-C14 are broken, and new Watson-Crick hydrogen bonds are formed in the base pairs G11-C16 and G12-C15, while C10 remains trapped inside the helix.

By contrast, the G-rich duplexes are all stable, both for DNA and RNA, with internal mismatches always inside the helix. In particular, the \( G_{syn} - G_{anti} \) conformations for the internal G-G mismatches are stable throughout the simulations (of course, end mismatches are subject to fluctuations). The alternating \( syn/anti \) conformation causes \( G_{syn} \) to point towards the minor groove, with an open angle value centered around 35°, while \( G_{anti} \) points towards the major groove around an open angle 75°, while still remaining inside the helix. Figure 5.10 shows the twist angle based on the C1'-C1' vectors (see SI for a definition of this twist). Following the definition of “steps” given in the “Initial Structures”
Figure 5.8  Open angle of RNA DC-1 (a) and DC-2 (b) as a function of time (left) and corresponding distribution (right). (c) Configurations A and B corresponding to the average structures of DC-1 at 400-500 ns and 650-700 ns, and typical hydrogen bonds associated with them.
Figure 5.9  (a) Open angle of RNA SC-3 as a function of time (left) and corresponding distribution (right). (b) Typical structures of C10-C16 pair base.
section, the steps 2, ..., 10 correspond to \( \text{AEAm}_2\text{Mm}_2\text{AE}_2 \) for DG-1, \( \text{ADAm}_1\text{Mm}_1\text{ADA}_2 \) for DG-2 and \( \text{m}_1\text{m}_1\text{Am}_2\text{m}_2\text{Am}_1\text{m}_1\text{A} \) for SG-3. Converged structures result in a twist that reflects the inversion symmetry of the sequence. Notice that the middle mismatches result in local unwinding of the helix. The G-G mismatch stays inside the helix and is stabilized by two hydrogen bonds, N7-N2 and O6-N1, present in both DNA and RNA; and a third hydrogen bond, N2-OP2, only present in RNA (Figure 5.11).

![Graph showing twist angles](image)

**Figure 5.10** Average over last 100 ns of twist angles of DNA duplexes (top) and RNA duplexes (bottom) as a function of step numbers for G-rich sequences.

Figure 5.11 shows the most common hydrogen bond patterns for the mismatches.
Figure 5.11  Typical hydrogen bond of G bases (a) and C bases (b) in mismatched bases. Due to the syn-anti configuration, a G-G pair forms two hydrogen bonds in DNA form: N7-N2 and O6-N1; and forms three hydrogen bonds in RNA form: N7-N2, O6-N1 and OP2-N2. The C-C pair forms either N3-N4 or O2-N4 hydrogen bonds. (c) Presence of hydrogen bonds in only N3-N4 (red), in only O2-N4 (blue), or formed simultaneously (green), as a function of time for DC-1 (top 6) and DC-2 (bottom 6) in DNA (left) and RNA (right).
Generally in both DNA and RNA there are two hydrogen bonds between the Watson-Crick edge of G anti and the Hoogsteen edge of G syn: O6-N1H and N7-N2H. In addition, RNA mismatches are further stabilized by a hydrogen bond between between N2H in G syn and its phosphate oxygen atom OP2 (see Table 5.S1 for hydrogen bond populations). The population of the hydrogen bonds in the C-C mismatches, on the other hand, varies wildly (see Table 5.S2). As Figure 5.11(c) shows, there are few hydrogen bonds in DNA DC-1, reflecting the tendency of the bases to flip out. The presence of hydrogen bonds increases for DNA DC-2, especially for O2-N4 in C7-C18. Hydrogen bonds in RNA DC-1 also seem to form a considerable amount of the time, and RNA DC-2 is remarkable in the sense that there are no N3-N4 hydrogen bonds, but the O2-N4 bonds are populated all the time. The hydrogen bonds populations are given in Tables 5.S1 for G-G mismatches and 5.S2 for C-C mismatches.

Now we turn to ion occupation whose definition is given in the SI (assignation to major or minor groove is slightly tricky around the mismatches). Figures 5.S11 and 5.S12 show the neutralizing Na$^+$ ion occupation as a function of base pairs in the major and minor grooves, and Figure 5.12 shows some typical patterns for ions involved in direct binding to the mismatched bases. We know from studies of ion distributions in the sequence (CG)$_6$ [22] that neutralizing Na$^+$ ions sit preferentially around the G bases in the major grooves, both in B-DNA and A-RNA forms. This trend is clearly seen in the G-rich duplexes in Figures 5.S11 and 5.S12 where the amount of ions in the minor groove is completely negligible and most of them sit in the major groove, especially around the G-G mismatches. Notice that for both DNA and RNA, the ion distributions are nicely converged reflecting the symmetry of the molecules. A Na$^+$ ion can stabilize a G-G mismatch by forming an ion bridge connecting the G(O6) atoms in the four G bases in a double G-G mismatch (Figure 5.12(a) ); or the two bases in a single G-G mismatch and the two bases in the adjacent G-C base pair (Figure 5.12(b) ). These pockets are very similar to the Na$^+$ ion trapped in a (CG)$_6$ duplex, where the ion is held by two G(O6) and two C(N4), in both B-DNA and A-RNA (see Figure 14(a) and (b) in Ref. [22]).

For the C-rich motifs the story is slightly more complicated. While a single Na$^+$ ion is needed to bridge four G bases in a double mismatch, two Na$^+$ ions are required for a double C-C mismatch, as shown in Fig.5.12(c). This situation occurs in C-C mismatches that tend to form an N3-N4 hydrogen bond. Figure 5.12(d) shows an ion bridge between two bases that are involved in O2-N4 hydrogen bonds. Figure 5.12(e) shows an ion trapped
Figure 5.12 Atomistic details of direct binding sites of $\text{Na}^+$ for different mismatches. (a) G-G mismatch in DG models; (b) a G-C base pair adjacent to a G-G mismatch (c) C-C mismatch where the bases stay inside the helix and form N3-N4 H-bonds; (d) C-C mismatches where the bases form O2-N4 H-bonds; (e) a G-C base pair adjacent to a C-C mismatch can also contribute to trap the ion. In addition, the figure shows two examples of ions binding to the helix backbone. (f) ion bridging four OP1 atoms in DNA DC-1; (g) ion linking the ends of a DG-2 RNA helix.
between a Watson-Crick bonded base pair and the immediately following C-C mismatch. For RNA, ions still prefer to sit in the major groove. While in RNA DC-2 the ions are still more localized in the central mismatches in the major groove, in both RNA DC-1 and RNA SC-3 the ion distribution in the major groove favors the Watson-Crick pairs and there is a non-negligible distribution of ions in the minor groove around the C-C mismatches. RNA DC-2 exhibits only conformation B, where one of the mismatches turns towards the minor groove, and the other stays inside the helix, both bases linked by an O2-N4 hydrogen bond. The internal pairing of the C bases reduces the distance between the two strands of the duplex resulting in a narrower minor groove that admits no ions (Figure 5.12(d,e)). In this configuration, the ion trapped at the major groove around the C mismatches is also held in place by the adjacent G-C base pair, as shown in Figure 5.12(e). In RNA DC-1, where both conformations A and B exist, the ion can be trapped at the minor groove when the bases form a N3-N4 hydrogen bond, as shown in Figure 5.12(c). However, if the two bases form an O2-N4 hydrogen bond, the ion favors the major groove (Figure 5.12(d)), as in the RNA DC-2 model. By contrast, in C-rich DNA the ion population in the minor groove considerably exceeds the population in the major groove, and it is distributed around the C-C mismatches (the DC-1 population is not completely symmetric due to the presence of the e-motif). A distribution where the minor groove Na\(^+\) ion population is larger than that in the major groove is curiously reminiscent of the ion distribution around (CG)\(_6\) Z-DNA [22]. In addition, Figures 5.S13 and 5.S14 present radial distribution functions for the various C-C mismatches for C(O2) atoms in the minor groove and C(N4) atoms in the major groove. Here we have concentrated on the main differences between C-C and G-G tracts, but it is obvious that ions are attracted to other electronegative groups in the sugar rings and backbones. Figure 5.12 showcases two such cases, (f) shows two Na\(^+\) ions directly bound to OP1 atoms in the minor groove of DC-1 DNA, while (g) shows a bridge ion linking the ends of the helix and thus aiding in the bending of a DG-2 RNA duplex. Situations like this are also observed, for instance, in an A-RNA (CG)\(_6\) helix [22].

Finally, for the sake of reference, we include in Table 5.S3 results for the structure prediction of 10 hexanucleotide repeats and their corresponding free energy as calculated by the Mfold web server [23]. Figure 5.S15 shows the predicted minimum free energy structures. The results are only for reference, as Mfold does not consider duplexes but single strands and therefore the resulting structures are hairpins, and the particular
The geometry of the mismatches is not considered.

5.3 Discussion

The three possible G-rich and C-rich double helices are shown in Figure 5.1. As described in the Methods section, for either the sense or antisense sequences, the double helices (when repeated periodically to avoid differences due to end effects) differ in the pattern of “steps” (that also include mismatches), and therefore are not equivalent. All structures, when repeated periodically, initially have the same number of G-G or C-C mismatches, and the same number of Watson-Crick bonded G-C base pairs. One obvious difference is that the “double G” DG and “double C” DC helices have adjacent, double mismatches, separated by 4 Watson-Crick G-C base pairs, while the “single G” SG and “single C” SC helices have single mismatches, separated by 2 Watson-Crick G-C base pairs.

Before discussing specific results, it is worthwhile to discuss two general trends. First, for a given sequence, mismatched RNA duplexes are more stable than mismatched DNA duplexes, as supported by the fact that the few crystal structures of G-G and C-C (and also A-A and U-U) mismatched double helices are only available for RNA structures. For example: (i) crystal structures with G-G mismatches (GCGGCGGC duplex [18], PDB id: 3R1C; GC(8BrG)GCGGCGGC duplex [18], PDB id 3R1D; copies of the 5’-CGG/3’-GGC motif, [19], PDB id: 3JS2); (ii) crystal structures with C-C mismatches (GCCGCCGC duplex [12], PDB id: 4E59); (iii) crystal structures with A-A mismatches (GGCAGCAGCC duplex [24], PDB ids: 3NJ6, 3NJ7). Second, the G-G mismatches are more stable than the C-C mismatches [12,18,25]. This is also not surprising, as the G-G mismatches favor both good hydrogen bonds and good stacking within the helix, but the C-C have weak hydrogen bonding and considerably less favorable stacking. This is what makes G-quadruplexes relatively abundant in the PDB, while continuous C-quadruplexes are non-existent. These general trends are also supported by detailed free energy maps that we are at present computing in the context of trinucleotide GGC and CCG repeats, and by the general trends in the least Gibbs free energy structures as calculated with the Mfold server and presented in Table 5.S3.

G-rich double helices

There are several features that the six DNA and RNA helices with G-G mismatches share:
1) The G-rich double helices are stable. This seems to be the case, at least up to 1µs, for the set of simulations presented here (set A) and for a second set of 1µs simulations that followed a shorter equilibration cycle (set B). There are occasional end effects, where the terminal mismatched bases flip back, but these do not affect the structure of the inner helix. Indeed, for the six DNA and RNA cases, the PCA reveals smooth Gaussian distributions for the first three principal components in each case, that correspond to bending or to groove breathing modes.

2) The preferred conformation for the internal G-G mismatches is \( G_{\text{syn}} - G_{\text{anti}} \). This is clearly illustrated in the map for the backbone sugar torsion \( \delta \) and the glycosyl torsion \( \chi \), shown in Figure 5.3, where distinct distributions are associated with \( G_{\text{syn}} \) and \( G_{\text{anti}} \). The syn and anti conformations correlate with the value of the open angle associated with the corresponding base. The open angle distributions of DNA and RNA are shown Figure 5.4. Both C and G bases that participate in Watson-Crick hydrogen bonds have a Gaussian distribution around 60°. In the mismatches, \( G_{\text{syn}} \) points towards the minor groove, with an open angle value centered around 35°, while \( G_{\text{anti}} \) points towards the major groove with an open angle around 75°.

3) The inner G-G mismatches stay inside the helix core and form hydrogen bonds. Figure 5.11 shows the most common hydrogen bond patterns for the mismatches. For both DNA and RNA, there are two hydrogen bonds between the Watson-Crick edge of \( G_{\text{anti}} \) and the Hoogsteen edge of \( G_{\text{syn}} \): O6-N1H and N7-N2H. In addition, RNA (but no DNA) mismatches are further stabilized by a hydrogen bond between between N2H in \( G_{\text{syn}} \) and its phosphate oxygen atom OP2. These hydrogen bonds have been observed experimentally for (CGG) trinucleotide repeat RNA duplexes [18] (the rather high bond populations are listed on Table 5.S1).

4) The inner G-G mismatches cause local unwinding of the helix. The twist angle based on the C1′-C1 vectors in Figure 5.10 reflects the inversion symmetry of the sequences. A local unwinding of the helix is reflected in a considerable but localized decay in twist in the inner mismatches.

5) Ions sit in the major groove, especially around the G-G mismatches and help stabilize G-G double mismatches via ion bridges. This trend occurs for both DNA and RNA, while ion occupation in the minor groove is completely negligible. In addition, ion bridges such as those shown in Figure 5.12 help stabilize the mismatches; in this case, by connecting the G(O6) atoms of the four mismatched G bases, or the G(O6) atoms of a G-G pair.
with the G(O6) and C(N4) atoms of an adjacent Watson-Crick base pair.

Finally, for either DNA or RNA, the current simulations are not enough to predict which model (DG-1, DG-2 or SG-3) is more stable.

**C-rich double helices**

Unlike their G-rich counterpart, the C-rich helices show a wide variety of conformations, with the one common characteristic that the inner mismatched C bases are all in some form of anti conformation. We thus need to consider the duplexes separately, and first consider DNA.

1) *C mismatched base flipping.* Unlike Watson-Crick hydrogen-bonded base pairs where base flipping takes of the order of milliseconds or longer, and enhanced sampling techniques are needed in simulation studies [21], base flipping in C-C mismatches occur in the microsecond time scale and can be observed in regular molecular dynamics simulations. While RNA C-C mismatches remain inside the helix, inclined towards the major or minor groove, DNA mismatched Cs tend to flip out of the helix core. An analysis of all the trajectories for C-rich DNA shows that when a C base completes its flipping into the minor groove, it folds back and interacts with bases towards the 5’ direction on its own strand. This directionality has also been observed in careful MD studies of base flipping with conformational flooding [21]. When the C base flips towards the major groove, the bases have more space, are more flexible and interact less with the rest of the helix. Linear combinations of the backbone torsion angles signal the flipping transition: a decay of $\alpha + \gamma$ of about 100° indicates a flipping towards the minor groove, and a jump of $\epsilon - \zeta$ of about 290° signals a flipping into the major groove.

2) *DNA DC-1 is stable and accommodates mismatches via the formation of e-motifs.* PCA shows nice Gaussian distribution for the three principal components, with the first two modes corresponding to groove breathing, as the shortening of the helix due to the e-motif hinders bending for the dodecamer. Figure 5.6 (d) tracks the evolution of the open angle for the internal mismatches C6-C19 and C7-C18. The mismatched bases C7 and C18 start to turn towards the minor groove early in the simulation: C7 flips out and towards the 5’ direction in the strand, interacting with G5; and C18 flips out and towards the 5’ direction of its own strand, interacting with G16. Eventually, both G5-C18 and C18-G16 form G(N3)-C(N4) hydrogen bonds. This gives rise to the rarely seen “e-motif”, first described in a solution conformation of a DNA CCG triplet repeat sequence in NMR
experiments [26], where the C bases (i residue) in a mismatch symmetrically flip out in the minor groove, pointing their base moieties in the direction of the i – 2 residue (i.e., towards the 5’ direction in each strand). Once formed around 300 ns, the “e-motif” is stable to the end of the simulation at 1 µs. After some initial flipping activity (C19 turning towards the minor groove and C6 towards the major groove), C6-C19 both settle inside the helical structure at around 300 ns, with occasional hydrogen bonding (Table 5.S2).

3) **DNA DC-2 is stable with the bases of a mismatch alternating between the minor and major grooves.** Figures 5.6 (c,d) show that the C7-C18 mismatch mainly stays inside the helix and the conformation of the C6-C19 mismatch evolves slowly. During 500-820 ns, C6 turns towards the minor groove and C19 flips out into the major groove. During the last 180 ns, C6 flips completely out into the minor groove, and upwards towards the 5’ direction of its strand, while C19 stays in the major groove.

4) **DNA SC-3 is unstable.** This was confirmed through six 1µs runs, with different conditions on the terminal bases to make sure that this is not due to the presence of a single hanging base at the end. The duplexes in these runs either unfold or convert into DC-1. Thus it seems that C-rich DNA needs at least four consecutive Watson-Crick base pairs in order to preserve stability (notice that Mfold also favors DC-1, and when SC-3 is enforced, it still switches to DC-1, see Table 5.S3).

5) **The DC-1 e-motif structure is better stacked than the DNA DC-2 structure.** A comparison of the DC-1 and DC-2 duplexes after 500 ns shows that the e-motif structure is better stacked, more compact than the structure in DNA DC-2. This can be observed in Figure 5.6 and can be quantified, for instance, by measuring the distance between neighboring layers for each base pair (i.e., for base pair i, we compute the distances between base pairs i – 1 and i + 1). This distance is the same for both duplexes except near the mismatches, where it is shorter for layers 6, 7 and 8 for DC-1. Thus the distance between basepairs 5 and 9 is 10.05 Å for DC-1 and 14.33 Å for DC-2. This is a measure of the better stacking of the e-motif.

6) **For the DNA anti-sense (CCCCCGG) HR duplexes, the stability is ranked as DC-1>DC-2>SC-3 (unstable).** We have shown that the e-motif is better packed than the DC-2 resulting structure, and probably it is the optimum structure for the C-rich hexanucleotide repeats. We know that the DC-1 sequence can adopt this motif, but we have not shown that this motif could not be present (if the simulations were long enough) in the other
two models, DC-2 and SC-3. Fortunately, in this case it is not necessary to run any longer to realize that the e-motif has to be less stable, or unstable, in the other two structures. This is due to the nature of the steps surrounding the mismatches. Indeed, going back to Figure 5.1, one can see that the inner mismatched bases in DC-1 can flip towards the 5’ strand and encounter a G base with which to form a G(N3)-C(N4) hydrogen bond. On the other hand, in DC-2 there are no G bases in the 5’ direction to help stabilize the flipping mismatches. Finally, in SC-3, a flipping C in the 5’ direction does encounter a G base but contributes to the destabilization of the helix, as it compromises one of the two hydrogen bonds that separate the mismatches.

7) For the DNA anti-sense (CCCCGG) HR duplexes, the ions occupy the minor groove and are centered around the mismatches. A Na\(^{+}\) ion distribution around the minor groove has also been observed in (CG)\(_{6}\) Z-DNA [22].

Now we consider the RNA C-rich duplexes.

8) RNA DC-1 displays two stable conformations: A, with one mismatch pair pointing towards the minor groove, and the other mismatch forming a N3-N4 hydrogen bond; and B, with two mismatching bases forming an O2-N4 hydrogen bond. The occupation state of the major groove by mismatched bases is coupled to the helix bending. In conformation A, shown in Figures 5.8 and 5.810, the C6-C19 mismatches stay inside the helix and occasionally form an N3-N4 hydrogen bond (Figure 5.11), while the C7 and C18 bases point towards the minor groove, facing each other. In this case, the major groove is less occupied and bases \(i - 1\) and \(i + 1\) interact pulling the backbones and bending the duplex. In conformation B, C7 and C19 stay inside the helix, stacking on top of each other, while the complementary mismatches C18 and C6 turn to the major groove. The mismatches C6-C19 and C7-C18 interact through O2-N4 hydrogen bonds, and occasionally C6 and C18 form an N3-N4 cross hydrogen bond. In this conformation, the occupation of the major groove by C6 and C18 pushes the strands away, unbending the duplex.

9) RNA DC-2 is stable in conformation B. Figures 5.8 and 5.810 show that only conformation B is observed in the RNA DC-2 duplex, with O2-N4 hydrogen bond forms between C6 and C19 as well as C7 and C18.

10) RNA SC-3 is the least stable RNA conformation, exhibiting some reorganization of hydrogen bonds. For the RNA SC-3 duplex shown in Figure 5.9, the C4-C22, C7-C19 and C10-C16 pair bases favor a synchronized oscillation motion between minor and major grooves, where if one base of the mismatched pair turns towards the minor groove, the
other turns towards the major one. After 450 ns, the Watson-Crick hydrogen bonds in base pair G11-C15 and G12-C14 are broken, and new Watson-Crick hydrogen bonds are formed in the base pairs G11-C16 and G12-C15, while C10 remains trapped inside the helix. Obviously, a re-organization of hydrogen bonds breaks the symmetry that defines SC-3. Interestingly, in the crystallization of RNA C-(CCG)$_2$-C, it was observed that the C-(CCG)$_2$-C oligomer formed a “slippery duplex” with overhanging nucleotides and terminal C bases expelled out of the helix core such that the C-C mismatch was surrounded on both sides by 4 Watson-Crick C-G base pairs [12]. This supports the idea that C-C mismatches in RNA, like those in their DNA counterpart, need to be separated by at least four successive Watson-Crick base pairs in order to preserve helix stability.

11) In RNA DC-2, ions favor the major groove around the C-C mismatches, while in RNA DC-1 and SC-3, ions in the major groove favor the Watson-Crick base pairs, and there is a non-negligible ion occupation in the minor groove around the C-C mismatches. Ion direct binding contributes to the stability of the duplexes. While a single Na$^+$ ion is needed to bridge four G bases in a double mismatch, two Na$^+$ ions are required for a double C-C mismatch (Figure 5.12(c)). This situation occurs in C-C mismatches that tend to form an N3-N4 hydrogen bond. Ion bridges also occur between two bases involved in O2-N4 hydrogen bonds (Figure 5.12(d)) and between mismatched bases and Watson-Crick base pairs (Figure 5.12(e)).

5.4 Conclusion

We have carried out the first comprehensive study based on MD simulations of the twelve dodecamers that result from the three different reading frames in sense (GGGGCC) and antisense (CCCCGG) HRs for both DNA and RNA. Our observations can be summarized as follows. G-rich double helices share common features. The inner G-G mismatches stay inside the helix in G$_{\text{syn}}$-G$_{\text{anti}}$ conformations and form two hydrogen bonds between the Watson-Crick edge of G$_{\text{anti}}$ and the Hoogsteen edge of G$_{\text{syn}}$: O6-N1H and N7-N2H. In addition, G$_{\text{syn}}$ in RNA forms a base-phosphate OP2-N2H hydrogen bond. Inner G-G mismatches cause local unwinding of the helix. Na$^+$ ions sit in the major groove and help stabilize G-G double mismatches via ion bridges that join either four G bases belonging to a double mismatch, or two Gs in a mismatch with the bases in the adjacent Watson-Crick G-C base pair. G-rich double helices are more stable than C-rich helices due to better
stacking and to the hydrogen bonds of the $G_{syn}-G_{anti}$ mismatches.

Unlike their G-rich counterpart, the C-rich helices show a wide variety of conformations, with the one common characteristic that the inner mismatched C bases are all in some form of anti conformation. Least (most) stable C-rich RNA and DNA helices have single (double) mismatches separated by two (four) Watson-Crick basepairs. For the C-rich DNA duplexes, mismatched Cs tend to flip out of the helix core, while RNA C-C mismatches remain inside the helix. A decay of the DNA backbone torsion angles $\alpha + \gamma$ of about 100° indicates a flipping of a C base towards the minor groove, and a jump of $\epsilon - \zeta$ of about 290° signals a flipping into the major groove.

For DNA, the duplex stability is ranked as DC-1>DC-2>SC-3 (unstable). The DNA DC-1 helix accommodates mismatches via the formation of e-motifs, where mismatched bases flip towards the minor groove and point in the 5’ direction of their respective strands. The DNA DC-2 duplex is stable with the bases of a mismatch alternating between the minor and major grooves, but the e-motif in DNA DC-1 provides far better stacking and stability to the double helix. The fact that the flipped C bases at position $i$ in an e-motif form G(N3)-C(N4) hydrogen bonds with the $i-2$ G base makes DC-1 the preferred sequence for the e-motif. There are two stable conformations in the C-rich RNA duplexes: conformation A, with one mismatch pair pointing towards the minor groove and the other mismatch pair forming a N3-N4 hydrogen bond; and conformation B, with two mismatching bases forming O2-N4 hydrogen bonds. The occupation state of the major groove by mismatched bases is coupled to the helix bending. RNA DC-1 displays both conformation A and B. In conformation A, one mismatch remains inside the helix while the other points towards the minor groove, thus leaving the major groove relatively unoccupied and causing the helix to bend towards the major groove. In conformation B, two bases belonging to different mismatches stack on top of each other, while their partners turn to the major groove, causing unbending of the helix. On the other hand, RNA DC-2 only exhibits conformation B. In C-rich DNA helices, Na$^+$ ions occupy the minor groove, centering around the C-C mismatches. In RNA DC-2, ions favor the major groove around the C-C mismatches, while in RNA DC-1 and SC-3, ions in the major groove prefer the Watson-Crick base pairs, and there is a non-negligible ion occupation in the minor groove around the C-C mismatches. Ion direct binding to C-C mismatches contributes to the stability of the duplexes.
5.5 Methods

The simulations were carried out with the AMBER 14 simulation package [27] with the ff12SB force field. The TIP3P water model [28] was used for the explicit solvent simulations with periodic boundary conditions in truncated octahedron water boxes with more than 5000 waters each. Electrostatics were handled by the PME method [29], with a direct space cutoff of 9 Å, and with an average mesh size of approximately 1 Å for the lattice calculations. We used Langevin dynamics with a coupling parameter $\gamma = 2.0 \, ps^{-1}$. The NPT simulations were carried out mainly at 300K and 1 atm with the Berendsen barostat [30] with an isothermal compressibility of $\beta = 44.6 \times 10^{-6} \, bar^{-1}$ and pressure relaxation time $\tau_p = 1.0 \, ps$. The length of the simulation was 1 $\mu$s for each model, and the coordinate information was collected every 5 ps.

The equilibration process took place in five steps. First, we carried out a steepest descent simulation followed by conjugate gradient minimization keeping the nucleic acid and ion atoms fixed at their initial positions. Second, the system was slowly heated up from 0 to 300 K in five steps, for a total of 1 ns, under constant volume condition and restraints (100 kcal/(mol Å)) on the nucleic acid and ion atoms. Third, another 1 ns NVT run with restraints on the nucleic acid and ion atoms was carried out at 300 K. Fourth, 1 ns NPT runs with slowly decreasing restraints from 100 kcal/(mol Å) to 0 kcal/(mol Å) in five steps, were carried out for a total of 1 ns. In this step, the system reached a stable density, around 1.0 g/mL. Fifth, 1 ns NPT Langevin dynamics with constant pressure at 300 K without restraints was carried out as a final check. These final structures were used for the 1 $\mu$s simulations. This set of simulations was labeled “set A”. We also have another set of 1 $\mu$s simulations that were carried out with a less extensive equilibration process. These simulations (labeled “set B”, not shown here except for a few cases), give very similar results to the ones presented here, with the main differences due to end effects in DNA, and different behavior in DNA SC-3 (shown in the Results section), which we believe is due to the lesser stability of this arrangement.

Definitions or comments for the principal component analysis, helix open angle, helix twist angle, and ion occupancy are given in the SI.
Acknowledgments

We gratefully acknowledge the financial support of the National Institute of Health (NIH) [R01GM118508]; the National Science Foundation (NSF) [SI2-SEE-1534941]; and the Extreme Science and Engineering Discovery Environment (XSEDE) [TG-MCB160064] for computational support.
5.6 Supporting information

PCA

Principal component analysis (PCA) is a statistical tool used to extract large amplitude motions occurring in the molecular dynamics trajectories through the eigenvectors of the mass-weighted covariance matrix of the atomic positional fluctuations [20]. The dynamics of an N-atom molecule can be expressed by 3N Cartesian coordinates $x(t)$. Before the covariance matrix is built, all the trajectories are fitted to the first frame to remove global translation and rotation. Then the covariance matrix $C$ is constructed from the position coordinates of the selected atoms (here we use all the heavy atoms without the terminal bases):

$$ C = \text{cov}(x) = \langle (x - \langle x \rangle)(x - \langle x \rangle)^T \rangle $$

where T denotes the transposed vector and $\langle \rangle$ denote an average over time. The matrix element can be calculated as:

$$ C_{ij} = \frac{1}{n} \sum_{i=1}^{n} \langle (x_i(t) - \langle x_i \rangle)(x_j(t) - \langle x_j \rangle) \rangle $$

where $n$ is the total number of trajectories, $x_i(t)$ are the position coordinates of $i = 1, 2, ..., 3N$ and $N$ is the number of atoms of the molecule. $\langle x_i \rangle$ is the average coordinate overall the trajectories. After the $3N \times 3N$ time-independent covariance matrix is constructed, a set of eigenvalues ($\lambda_1, \lambda_2, ..., \lambda_{3N}$) and eigenvectors ($v_1, v_2, ..., v_{3N}$), e.g. principal components (PCs), are obtained, which become the new collective coordinates to describe the system. The projection of the original MD trajectory on the corresponding eigenvector (or PC) can be calculated as:

$$ q_i(t) = v_i \cdot (x(t) - \langle x \rangle) $$

The variance of the distribution of $q_i(t)$ is the variance of the atomic fluctuation along the $i^{\text{th}}$ PC, i.e. the corresponding eigenvalue: $\lambda_i = \langle q_i(t)^2 \rangle$. The eigenvalues $\lambda_1, \lambda_2, ..., \lambda_{3N}$ are sorted in order of decreasing value, and the first one $\lambda_1$ represents the largest positional deviation. The original Cartesian coordinates $x(t)$ are transformed into collective coordinates $v(t)$ with the corresponding eigenvalues in decreasing order, and the first few
dimensional subspaces describe the largest variances, thus only the first few subspaces are generally “essential” to check the system dynamics. In this way, the dimension of the system can be reduced to “essential spaces”. To visualize the motion described by a PC, it is useful to project it back to Cartesian coordinates:

$$x_i(t) = q_i(t) \cdot v_i + \langle x \rangle$$

**Duplex open angle**

Open angle (or geometrical quantification of base flipping) has been defined to quantify the base pair opening and flipping motion. Originally, it was defined [31] as the angle between the glycosidic bond (C1’-N9 for G and C1’-N1 for C) and the vector joining the two C1’ atoms of a base pair, as shown in Fig.5.S1. Rather than using this angle directly, we use the projection onto the plane perpendicular to the local helical axis U. This axis is defined as the mean vector of U1 and U2, each of which links the C1’ atoms of the nucleotides preceding and following the open base in the corresponding strand of the duplex. We use C1’(i-1)-C1’(i+1) to define U for Watson-Crick pair bases, and C1’(i-2)-C1’(i+2) to define U for mismatches. The base rotations are positive for a right-hand rotation around the 5’-3’ direction, therefore base flips into the major group correspond to an open angle larger than the open angle for a matched base pair, while flips into the minor group reduce the value of the open angle.
Figure 5.S1 Definition of open angle for a duplex. The local axis U is the average vector of U1 and U2, defined as the vectors of C1’(i-1)-C1’(i+1) (or C1’(i-2)-C1’(i+2) for mismatches) along each strand. The open angle is defined by the vectors C1’-N9(N1) and C1’-C1’ projected on the plane perpendicular to the local axis U.
Duplex twist angle

We use the 3DNA software package [32,33] to calculate the twist angle of the duplexes. Since non-Watson-Crick base pairs are our main object, we choose the “simple” step parameters, which are “intuitive” for non-Watson-Crick base pairs and were introduced into 3DNA as of v2.3-2016jan01. The regular z-axis defined in 3DNA is used here, which is the average of two base normals, taking into consideration the M-N vs M+N base-pair classification. The “simple” inter-base-pair step parameter calculation uses consecutive C1′-C1′ vectors. Since the G-G mismatches have a syn-anti configuration, the z-axis turns out opposite to the normal one. Thus, one must add 180° to the twist angles involving G(syn)-G(anti) mismatches.

Ion occupation in duplex models

The occupation of ions in the major and minor groove side of each base pair has been calculated as follows. We define a cutoff of 3.5Å. For Watson-Crick G-C base pairs, occupancy is defined when there is an ion whose distance to either a major groove atom, C(N4) or G(O6,N7), or a minor groove atom, G(N2,N3) or C(O2), is less than the cutoff. For G-G mismatches, if the distance between an ion and any of the atoms G_{anti}(O6,N7) and G_{syn}(N1,N2,O6) is smaller than the cutoff, this ion is considered to be in the major groove. If the distance between an ion and any of the atoms G_{anti}(N2,N3) and G_{syn}(N7) is smaller than the cutoff, this ion is counted as being in the minor groove.

The base flipping in the C-C mismatches requires a more elaborate criterion to define the ion population. We calculate 4 distances from the ion to the C(O2) and C(N4) atoms in a C-C mismatch as D(O2)_{1}, D(N4)_{1}, D(O2)_{2} and D(N4)_{2}. If any of these distances is smaller than the cutoff (3.5Å), the ion has been localized. If D(O2)_{1}+D(O2)_{2} ≥ D(N4)_{1}+D(N4)_{2}, the ion has been trapped in the minor groove, otherwise it has been localized in the major groove.
Table 5.S1 Hydrogen bond population of G-G mismatches in double (D) and single (S) models for DNA and RNA.

<table>
<thead>
<tr>
<th></th>
<th>6-19(D)/4-22(S) (%)</th>
<th>7-18(D)/7-19(S) (%)</th>
<th>10-16(S) (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DG-1</td>
<td>N7-N2: 87</td>
<td>84</td>
<td></td>
</tr>
<tr>
<td></td>
<td>O6-N1: 81</td>
<td>85</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Both: 73</td>
<td>73</td>
<td></td>
</tr>
<tr>
<td></td>
<td>OP2-N2: 1 (G6)</td>
<td>4 (G18)</td>
<td></td>
</tr>
<tr>
<td>DG-2</td>
<td>N7-N2: 91</td>
<td>91</td>
<td></td>
</tr>
<tr>
<td></td>
<td>O6-N1: 65</td>
<td>64</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Both: 61</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td></td>
<td>OP2-N2: 5 (G6)</td>
<td>5 (G18)</td>
<td></td>
</tr>
<tr>
<td>SG-3</td>
<td>N7-N2: 97</td>
<td>74</td>
<td>94</td>
</tr>
<tr>
<td></td>
<td>O6-N1: 77</td>
<td>89</td>
<td>65</td>
</tr>
<tr>
<td></td>
<td>Both: 75</td>
<td>67</td>
<td>63</td>
</tr>
<tr>
<td></td>
<td>OP2-N2: 3 (G4)</td>
<td>3 (G19)</td>
<td>6 (G10)</td>
</tr>
<tr>
<td>RNA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DG-1</td>
<td>N7-N2: 65</td>
<td>65</td>
<td></td>
</tr>
<tr>
<td></td>
<td>O6-N1: 95</td>
<td>95</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Both: 62</td>
<td>62</td>
<td></td>
</tr>
<tr>
<td></td>
<td>OP2-N2: 97 (G6)</td>
<td>98 (G18)</td>
<td></td>
</tr>
<tr>
<td>DG-2</td>
<td>N7-N2: 80</td>
<td>80</td>
<td></td>
</tr>
<tr>
<td></td>
<td>O6-N1: 90</td>
<td>90</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Both: 72</td>
<td>72</td>
<td></td>
</tr>
<tr>
<td></td>
<td>OP2-N2: 98 (G6)</td>
<td>98 (G18)</td>
<td></td>
</tr>
<tr>
<td>SG-3</td>
<td>N7-N2: 93</td>
<td>62</td>
<td>97</td>
</tr>
<tr>
<td></td>
<td>O6-N1: 72</td>
<td>95</td>
<td>75</td>
</tr>
<tr>
<td></td>
<td>Both: 67</td>
<td>60</td>
<td>73</td>
</tr>
<tr>
<td></td>
<td>OP2-N2: 94 (G4)</td>
<td>96 (G19)</td>
<td>97 (G10)</td>
</tr>
</tbody>
</table>
Table 5.S2 Hydrogen bond population of C-C mismatches in double (D) (G6-G19 and G7-G18) and single (S) (G4-G22, G7-G19, G10-G16) models for DNA and RNA.

<table>
<thead>
<tr>
<th></th>
<th>6-19(D)/4-22(S) (%)</th>
<th>7-18(D)/7-19(S) (%)</th>
<th>10-16(S) (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA</td>
<td>DC-1 N3-N4</td>
<td>14</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>O2-N4</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Both</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>DC-2 N3-N4</td>
<td>38</td>
<td>38</td>
</tr>
<tr>
<td></td>
<td>O2-N4</td>
<td>16</td>
<td>45</td>
</tr>
<tr>
<td></td>
<td>Both</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>R-SC-3 N3-N4</td>
<td>81</td>
<td>83</td>
</tr>
<tr>
<td></td>
<td>O2-N4</td>
<td>30</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Both</td>
<td>13</td>
<td>5</td>
</tr>
<tr>
<td>RNA</td>
<td>DC-1 N3-N4</td>
<td>11</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>O2-N4</td>
<td>51</td>
<td>42</td>
</tr>
<tr>
<td></td>
<td>Both</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>DC-2 N3-N4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>O2-N4</td>
<td>97</td>
<td>98</td>
</tr>
<tr>
<td></td>
<td>Both</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>SC-3 N3-N4</td>
<td>42</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>O2-N4</td>
<td>21</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Both</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Table 5.S3 Structure prediction for 10 repeats of the GGGGCC and CCCCCG hexanucleotide sequences and the corresponding Gibbs energy as calculated by the Mfold web server [23]. Restraints have been imposed in order to obtain the duplex models considered in this paper. The restraint position indicates the pair base where a restraint force is applied on. The folding temperature is fixed at 310 K.

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Name</th>
<th>Restraint position</th>
<th>Final structure</th>
<th>∆G (kcal/mol)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(GGGGGCC)_{10}</td>
<td>dna-free-G</td>
<td>None</td>
<td>SG-3</td>
<td>-25.10</td>
</tr>
<tr>
<td>(GGGCGG)_{10}</td>
<td>dna-DG1</td>
<td>G2-C59</td>
<td>DG-1</td>
<td>-23.61</td>
</tr>
<tr>
<td>(GCGGGG)_{10}</td>
<td>dna-DG2</td>
<td>C2-G59</td>
<td>SG-3</td>
<td>-22.71</td>
</tr>
<tr>
<td>(GGGGCC)_{10}</td>
<td>dna-SG3</td>
<td>G2-C60</td>
<td>SG-3</td>
<td>-25.10</td>
</tr>
<tr>
<td>(CCCCGG)_{10}</td>
<td>dna-free-C</td>
<td>None</td>
<td>DC-1</td>
<td>-17.83</td>
</tr>
<tr>
<td>(CCCGGG)_{10}</td>
<td>dna-DC1</td>
<td>C2-G59</td>
<td>DC-1</td>
<td>-18.06</td>
</tr>
<tr>
<td>(CGGCCG)_{10}</td>
<td>dna-DC2</td>
<td>G2-C59</td>
<td>DC-2</td>
<td>-17.51</td>
</tr>
<tr>
<td>(CCCCCG)_{10}</td>
<td>dna-SC3</td>
<td>C2-G60</td>
<td>DC-1</td>
<td>-14.26</td>
</tr>
<tr>
<td>(GGGGGCC)_{10}</td>
<td>rna-free-G</td>
<td>None</td>
<td>SG-3</td>
<td>-40.80</td>
</tr>
<tr>
<td>(GGGCGLG)_{10}</td>
<td>rna-DG1</td>
<td>G2-C59</td>
<td>DG-1</td>
<td>-42.50</td>
</tr>
<tr>
<td>(GCCGGG)_{10}</td>
<td>rna-DG2</td>
<td>C2-G59</td>
<td>DG-1</td>
<td>-38.50</td>
</tr>
<tr>
<td>(GGGGCC)_{10}</td>
<td>rna-SG3</td>
<td>G2-C60</td>
<td>SG-3</td>
<td>-40.80</td>
</tr>
<tr>
<td>(CCCCCG)_{10}</td>
<td>rna-free-C</td>
<td>None</td>
<td>DC-1</td>
<td>-35.30</td>
</tr>
<tr>
<td>(CCCGGG)_{10}</td>
<td>rna-DC1</td>
<td>C2-G59</td>
<td>DC-1</td>
<td>-35.70</td>
</tr>
<tr>
<td>(CGGCCG)_{10}</td>
<td>rna-DC2</td>
<td>G2-C59</td>
<td>DC-2</td>
<td>-38.30</td>
</tr>
<tr>
<td>(CCCCCG)_{10}</td>
<td>rna-SC3</td>
<td>C2-G60</td>
<td>DC-1</td>
<td>-31.50</td>
</tr>
</tbody>
</table>
Figure 5.S2 Displacement of the first principal component as a function of time, as well as snapshots corresponding to the bimodal distributions. The red structure is an average structure over 600 ns to 650 ns, and the blue structure is an average structure over 885 ns to 935 ns.
Figure 5.S3  Left: PCA distributions of R-SC-3 and SC-3+C DNA duplexes over the 500 ns - 1 µs interval projected onto the first three principal components. Right: Visualization of the atomic motions involved in the 1st (blue) and 2nd (red) principal components.
Figure 5.84 Displacement of the first principal component as a function of time, as well as snapshots corresponding to the bimodal distributions. The red structure is an average structure over 575 ns to 620 ns, and the blue structure is an average structure over 850 ns to 900 ns.
Figure 5.S5 Displacement of the first principal component as a function of time, as well as snapshots corresponding to the bimodal distributions. The red structure is an average structure over 600 ns to 650 ns, and the blue structure is an average structure over 875 ns to 925 ns.
Figure 5.6 Displacement of the first principal component as a function of time, as well as snapshots corresponding to the bimodal distributions. The red structure is an average structure over 500 ns to 550 ns, and the blue structure is an average structure over 950 ns to 1000 ns.
Figure 5.S7: Displacement of the first principal component as a function of time, as well as snapshots corresponding to the bimodal distributions. The red structure is an average structure over 650 ns to 700 ns, and the blue structure is an average structure over 925 ns to 975 ns.
Figure 5.88 Displacement of the first principal component as a function of time, as well as snapshots corresponding to the bimodal distributions. The red structure is an average structure over 725 ns to 735 ns, and the blue structure is an average structure over 740 ns to 750 ns.
Figure 5.S9 $\chi$-angle of DC-1 (odd rows) and DC-2 (even rows) in DNA (top two) form and RNA (bottom two) form as a function of time (left) and the distribution (right).
Figure 5.S10  (a) Open angle (blue line) of C18 superposed to the radius of gyration ($R_g$) (red line) of the RNA duplex as a function of time. (b) Two conformations: A is bent towards the major groove when the mismatching C bases turn to the minor groove. B is an unbent duplex where the mismatching bases are in the major groove.
Figure 5.S11 Occupation (%) of ions trapped in the major (red) and minor (blue) grooves as a function of base pair for G mismatching duplexes (left) and C mismatching duplexes (right) in DNA form.
**Figure 5.S12** Occupation (%) of ions trapped in major (red) and minor (blue) grooves as a function of base pair for G mismatching duplexes (left) and C mismatching duplexes (right) in RNA form.
Figure 5.S13  Radial distribution function of Na\(^+\) at C(O2) (left) and C(N4) (right) for C mismatching helices in DNA form.
Figure 5.S14  Radial distribution function of Na$^+$ at C(O2) (left) and C(N4) (right) for C mismatching helices in RNA form.
Figure 5.S15  Secondary structure prediction with least Gibbs energy computed with Mfold of (a) DNA G mismatching models (b) DNA C mismatching models; (c) RNA G mismatching models; and (d) RNA C mismatching models, resulting in configurations (a) SG-3, (b) DC-1, (c) DG-1 and (d) DC-2.

References


Chapter 6

Conclusions

This thesis presents a study of the structural characterization of selected disordered proteins (including challenging small IDPs gp41$_{659-671}$, N$_{18}$, and N$_{8-N_8}$), ordered amyloids proteins (which resulting from the misfolding of polyQ and polyN that are IDPs in their monomeric forms), as well as DNA and RNA double helices with mismatched base pairs as obtained from the GGGGCC and CCCCGG hexanucleotide repeats.

Intrinsically disordered peptides are oftner hard to adequately characterize experimentally as discussed in Chapter 2. The original experimental data for the solvated gp41$_{659-671}$ structure was highly contradictory. Hence, we investigated the structure of gp41$_{659-671}$ with classical simulation techniques in order to gain insight into the solvated structure of this peptide.

Our investigation has looked at the conformational ensembles of gp41$_{659-671}$ with the AMBER force fields ff99SB and ff12SB, in both implicit and explicit waters via MD for a cumulative time longer than 7.2 s. Our results are consistent with the bulk of the experimental findings. The amount of helical population is important in aqueous solution, but this structure forms part of a flexible ensemble of conformation with a rugged free energy landscape characterized by shallow minima. When the original disordered peptide at the equilibrium end-to-end distance corresponding to the free energy minimum is subject to uniaxial stretching, it first becomes almost entirely helical due to the loss of the conformational entropy of the disordered state under uniaxial tension. Upon further stretching, the helix melts into turns, loops and 3$_{10}$ helices. The explicit waters enable the formation of additional hydrogen bonds that facilitate the “melting” of the helical structure and the lowering of the free energy when compared with the implicit solvent models.
Significantly, populations in the conformational ensemble show epitope conformations that are close to the NMR aqueous solution structure (1LCX) of Biron et al. [7], as well as epitope conformations close to the very different crystal structure of the peptide bound to 2F5 (1TJH) of Ofek et al. [8]. The latter agreement is obtained under weak end-to-end distance restraint that mildly mimics one of the several effects of the binding to 2F5. We find that the newest force field ff12SB seems to perform somewhat better than ff99SB, especially under restraints that take the peptide away from its equilibrium end-to-end distance.

We have also compared the effectiviness of several secondary structure analysis codes: DSSP, STRIDE and KAKSI, as applied to the mainly disordered gp41_{659-671}, N_{18} and N_{8}-N_{8} peptides. As expected, the agreement between the codes for these disordered proteins is considerably poorer than that for regular structures. STRIDE and DSSP define more secondary structural elements than KAKSI, and therefore can provide more details when this information is required. When it comes to distinguishing between 3_10 and \( \alpha \) helices, DSSP performs better than STRIDE. In particular, the DpSn ensembles (the subset of residues found in a given state by DSSP but not by STRIDE) correctly produce a helical rise distribution centered at 2 Å for 3_10 helices and at 1.5 Å for \( \alpha \) helices. However, the DnSp ensembles exhibit distributions centered at shorter helical rises for 3_10 helices and longer helical rises for \( \alpha \) helices. Distributions for the dihedral angle sum \( \psi_i + \phi_{i+1} \) also confirm a better assignment of 3_10 and \( \alpha \) helices by DSSP. The largest source of differences is due to the definition of turns: in DSSP their definition is based on the hydrogen bonds, and include \( \alpha \) turns, \( \beta \) turns, \( \gamma \) turns; in STRIDE only \( \beta \) turns and \( \gamma \) turns are recognized based on the distance \( C_i^\alpha \rightarrow C_{i+3}^\alpha \) and the backbone dihedral angles (\( \phi \), \( \psi \)). This results in a great disparity of these conformations: of all the turns defined by STRIDE, only approximately 40% are recognized as “turns” by DSSP, while the remaining residues are classified as bends, coils, and other structures. DSSP performs somewhat better than STRIDE on IDPs/IDRs.

In a different theoretical thrust, we have carried out MD simulations of a series of N/Q-rich peptides to study and compare the structural characteristics and thermodynamic stability of their protofibril aggregates in prion-like proteins. Specifically, we explored all possible steric zipper interfaces for N-rich and Q-rich oligomers in order to determine which crystallographic class provides the most stability to the N-rich or Q-rich fibril precursor. Our observations can be summarized as follows. First, different crystal classes
maximize stability in N-rich or Q-rich aggregates: N-rich oligomers are most stable in parallel-stranded $\beta$ sheets with 2-by-2 side-chain interdigitation (model 4T3), while polyQ oligomers are most stable in antiparallel-stranded $\beta$ sheets with 1-by-1 steric zippers (model 4O10). Our results show that the polyQ 4O10 model is more stable than the polyN 4T3 model. Enthalpic contributions to the free energy favor the formation of polyQ protofibrils, while entropic contributions favor the formation of polyN protofibrils. The presence of $\beta$-hairpin turns in polyN leading to $\beta$ sheets does not necessarily explain the observed trend of polyN faster aggregation, since N-rich sequences tend to favor parallel $\beta$ sheets, as opposed to antiparallel ones, and $\beta$-hairpin turns are more stable in polyQ than in polyN. The considerably larger phase space that disordered polyQ must sample on its way to aggregation probably is at the root of the associated slower kinetics observed experimentally. When other amino acids are present, such as in the Sup35 PrD, their shorter side chains favor steric-zipper formation for N but not Q, as they preclude the in-register association of the long Q side chains.

Finally, we have carried out the first comprehensive theoretical study based on MD simulations of all the possible linear duplexes built from all possible reading frames in the sense (GGGGGCC) and antisense (CCCCCGG) HRs for both DNA and RNA. Our observations can be summarized as follows. G-rich helices share common features. The inner G-G mismatches stay stable inside the helix in $G_{syn}$-$G_{anti}$ conformations and form two hydrogen bonds (HBs), O6-N1H and N7-N2H in DNA form and one additional HB OP2-N2H within $G_{syn}$ base in RNA form. Inner G-G mismatches cause local unwinding of the helix, and the ions sitting in the major groove help stabilize G-G double mismatches via ion bridges. G-rich double helices are more stable than C-rich helices due to better stacking and HBs of G-G mismatches. Unlike the G-rich counterpart, the C-rich helices show a wide variety of conformations, with the one common characteristic that the inner mismatched C bases are all in some form of anti conformation. For the DNA anti-sense (CCCCCGG) HR duplexes, mismatched C’s tend to flip out of the helix core, while RNA C-C mismatches remain inside the helix. The stability is ranked as DC-1>DC-2>SC-3 (unstable) in DNA form. The DNA DC-1 is stable and accommodates mismatches via the better-stacked formation of e-motifs, where mismatched bases flip towards the minor groove and point in the 5’ direction. The DC-2 is stable with the bases of a mismatch alternating between the minor and major grooves. There are two stable conformations in the RNA anti-sense HR duplexes: A, with one mismatch pair pointing towards the minor
groove and the other mismatch pair forming a N3-N4 hydrogen bond; and B, with two
mismatching bases forming O2-N4 hydrogen bonds. The occupation state of the major
groove by mismatched bases is coupled to the helix bending. RNA DC-1 displays both
conformation A and B, and DC-2 is stable in conformation B. SC-3 is the least stable RNA
formation, exhibiting some reorganization of hydrogen bonds, which supports the idea
that RNA C-C mismatched duplexes need to have at least four successive Watson-Crick
base pairs for helix stability. For the DNA anti-sense HR duplexes, the ions occupy the
minor groove, centering around the mismatches, while ions favor the major groove around
the C-C mismatches in RNA DC-2. In RNA DC-1 and SC-3, ions in the major groove
prefer the Watson-Crick base pairs, and there is a non-negligible ion occupation in the
minor groove around the C-C mismatches. Ion direct binding contributes to the stability
of the duplexes.

References


