Organised Chaos: defining different degrees of intrinsic disorder by molecular dynamics methods

Matthew Nixon B.Sc.

A thesis submitted to Maynooth university in fulfilment of the requirement for the degree of

Doctor of Philosophy

By

Matthew Nixon B. Sc.

Department of Chemistry

Maynooth university

July 2020

Research supervisor: Dr. Elisa Fadda

Head of Department: Dr. Jennifer McManus

Table of Contents

Acknowledgements

Firstly, I would like to thank my supervisor Dr. Elisa Fadda for her help and support over these last four years. Without your advice and guidance this PhD project would not be possible. And for how difficult a process studying for a PhD is, you have made every step as easy as possible.

I would like to thank the Maynooth University Graduate Studies office for their support of my research through the John Hume Scholarship. I would also like to thank the Maynooth University Chemistry Department and all the academic staff and technicians for their help throughout these last four years

I would like to thank all the collaborators involved in this project, namely Prof. Paul Moynagh and his postdoc Dr. Linan Xu, and Dr. Lucia Chemes and her postdoc Dr Lucia Alvarez, for all the hard work and effort that they have put into our collaboration.

To all the postgrads in the Chemistry Department and Hamilton Institute thanks for all the laughs, the infinite cups of tea. The great Halloween and Christmas parties and the fun nights out. Thanks to "Aiofe" Harbison for being a great lab/office mate and all the "quick breaks" that lasted an hour or more.

I would like to thank my parents Mary and Bernard for supporting me throughout all my years of undergrad and postgrad and for letting me move back home when I needed. For their love, support and understand I would like to thank my brothers Paul and Mark, sister, Jennifer and their partners and of course my lovely nephew Theo.

Finally, my greatest thanks go to Craig Gavagan, who has put up with me, the PhD, and all the stress that went with it for the last four years. You have been a source of unmeasurable emotional support and I could not have made it through the process without you.

Declaration

This thesis has not been submitted before, in whole or in part, to this or any other University for any degree, and is, except where otherwise stated, the original work of the author

Matthew Nixon

Abbreviations

AcOH	Acetic acid
AMBER	Assisted Model Building with Energy Refinement
Ala	Alanine
Arg	Arginine
Asn	Asparagine
Asp	Aspartic acid
ATP	Adenosine Triphosphate
Boc	*tert*-butyloxycarbonyl
CD	Circular Dichroism
CHARMM	Chemistry at HARvard Macromolecular Mechanics
Cys	Cystine
DNA	DeoxyriboNucleic Acid
ETC	Electron Transport Chain
ECSIT	Evolutionarily Conserved Signalling Intermediate in Toll pathway
ERCC1	Excision Repair Cross-Complementation group 1
EtOAc	Ethyl Acetate
Fmoc	Fluorenylmethyloxycarbonyl
FRET	Förster Resonance Energy Transfer
fs	femtosecond
GB	Generalised Born
Gln	Glutamine
Glu	Glutamic acid
Gly	Glycine
GROMACS	GROningen MAchine for Chemical Simulations
His	Histidine
HPV	Human Papillomavirus
ICHEC	Irish Centre for High End Computing
IDPs	Intrinsically Disordered Proteins

IDRs	Intrinsically Disordered Region
Ile	Isoleucine
Leu	Leucine
LINCS	LINear Constraint Solver
Lys	Lysine
MD	Molecular Dynamics
Met	Methionine
MM	Molecular Mechanics
MoRFs	Molecular Recognition Features
NER	Nucleotide Excision Repair
nM	Nanomolar
NMR	Nuclear Magnetic Resonance
ns	nanosecond
PB	Poisson-Boltzmann
PCA	Principal component analysis
PDB	Protein Data Bank
Phe	Phenylalanine
PPI	Protein-Protein Interaction
PPII	polyproline type II secondary structure
Pro	Proline
ps	picosecond
P53-CTD	p53 C-Terminal Domain
QM	Quantum mechanics
RAC	RMSD average correlation values
Rg	Radius of Gyration
REMD	Replica Exchange Molecular Dynamics
RMSD	Root Mean Square Deviation
RNA	RiboNucleic Acid
Ser	Serine
SAXS	Small Angle X-ray Scattering
SLiM	Short Linear Motif

TFA	Trifluoroacetic acid
Thr	Threonine
Trp	Tryptophan
Tyr	Tyrosine
μs	microsecond
Val	valine
XPA	xeroderma pigmentosum, complementation group A

Abstract

Intrinsically disordered proteins (IDPs) and regions (IDRs) are defined as proteins, or proteins' regions, that lack a stable 3D structure. The nature of the IDP(R)s amino acid residues and their sequence prevents these systems from folding. IDPs and IDRs are ubiquitous in biology, playing many different roles that range from mechanical, see for example elastin, to scaffolding the structure of macromolecular assemblies, such as the XPA in the Nucleotide Excision Repair (NER) pathway, to regulatory and signalling functions, such as p53 and short linear motifs (SLiMs)-containing regions. Trying to understand the relationships between structure and functions in these highly flexible and dynamic systems is rather challenging, because their intrinsic flexibility makes them extremely difficult to characterize experimentally through structural biology methods, such as X-ray crystallography, NMR, or Cryo-EM. Indeed, the actual classification of protein domains as "intrinsically disordered" derives from the inability of determining 3D atomic coordinates from experimental sources. According to the dogma underlying the whole field of structural biology, all the different functions that IDP(R)s are key player in should be linked to their structure, however this structure is undecipherable within the experimental timescale. The overarching design of my thesis work was to understand these structure-to-function relationships in specific IDP(R)s by molecular simulation techniques. My main goal was to understand how different levels of residual secondary structure propensity in specific IDP(R)s systems of importance in health and disease can explain their macroscopic function and more specifically, how structural disorder can regulate molecular recognition at the atomistic level of details. Molecular simulation techniques have now come to age and are well suited to play a starring role in structural biology, not only as a support for experimental methods, but also as equal partners and/or as the primary research tool for discovery. In this thesis I show how different molecular simulations techniques can be successfully used for discovery in the structure and function of IDP(R)s a) on their own, b) as an equal partners to experiments, c) as tools to understand experimental data and d) for the *ad hoc* design of new experiments. More specifically, we found that in many systems, namely in XPA and in the p53 and ECSIT C-terminal domains, molecular recognition can be explained in terms of a distinct residual secondary structure propensity, which results in an enrichment of the

disordered conformational ensemble with secondary structure motifs that may act as molecular recognition features (MoRFs), or nucleation sites. According to this hypothesis, specific MoRFs are recognized by receptors and fold upon binding, within a scheme that lies between the two extremes of “conformational selection” and “induced fit”. As shown for XPA, the structure of these MoRFs obtained through computing can inform the design of macrocyclic systems that can be used to inhibit specific receptors or pathways in the absence of any other structural information, in view of the development of new diagnostic and/or therapeutic strategies. In this work I also show cases where the propensity to form MoRFs, also termed pre-structuring, is not required. More specifically, in the case of the very short SLiM LxCxE-containing peptides, which can reach low nM binding affinities for the Retinoblastoma (Rb) protein, we have not detected any residual secondary structure propensity (or MoRF) in the unbound peptides. This particular study was conducted in partnership with experiments, through extensive sampling simulations for the bound and unbound peptides, in the presence of counterions or no counterions, and in the case of phosphorylation and in the absence of phosphorylation. Interestingly, in agreement with experiment, our simulations show that phosphorylation increases the level of polyproline II (PPII) structure in the unbound peptides, which has been also recently underlined as an important in signalling pathways. In summary, I believe that this research work on a few examples of structurally disordered systems contributes to shed some light into intrinsic disorder in biomolecular recognition and how intrinsic disorder should not be classed as one grey area, but instead viewed as incorporating many different shades of conformational degrees and diversity that modulate binding affinity through their structure, *via* enthalpy, and relative stability, *via* entropy.

List of Publications

i. Fadda E., and **Nixon M. G.** "The transient manifold structure of the p53 extreme C-terminal domain: insight into disorder, recognition, and binding promiscuity by molecular dynamics simulations" PCCP, 19(32):21287-21296

ii. **Nixon M. G.**, and Fadda E**.** (2019) Binding Free Energies of Conformationally Disordered Peptides Through Extensive Sampling and End-Point Methods. In: McManus J. (eds) Protein Self-Assembly. Methods in Molecular Biology, vol 2039. Humana, New York, NY

iii. Xu L., Humphries F., Delagic N., Wang B., Holland A., Edgar K.S., Hombrebueno J.R., Beer Stoltz D., **Nixon M. G.**, Fadda E., Glezeva N., McDonald K., Watson C. J., Ledwidge M. T., Grieve D. J. and Moynagh P. N *Submitted*. "Human ECSIT protein is critical for assembly of mitochrondrial complex I in the heart: Key limiting factor for cardiac function".

Not yet submitted

i. Alvarez, L.*; **Nixon, M.G.***, Claron, M., Sehr, P., Will, D., Lewis J, Gibson, T. J., Fadda, E. and Chemes, L. B. "Unravelling molecular determinants of the interaction between the Retinoblastoma Tumour Suppressor Protein and Short Linear Motifs"
*Both authors are co-first author

Chapter 1: Introduction

Proteins are an important class of biomolecules, responsible for driving the inner works of the cellular machinery through a multitude of biological processes. Proteins vary in size from short peptides, to large multimeric structures and, within their native environment, their primary structure, or sequence, determines uniquely their propensity to form stable 3D structures, or to fold. Over 50% of proteins encoded in the human genome contain regions that because of their amino acids sequence do not fold[1,2], retaining a very high degree of structural disorder, or intrinsic disorder. Intrinsically disordered regions (IDRs) and intrinsically disordered proteins (IDPs) are characterized by a low sequence complexity with an over representation of hydrophilic and charged residues[3,4]. This trait has been be used quite reliably as a predictor of intrinsic disorder[5–7]. This intrinsic conformational dynamic hinders their characterization through experimental structural biology techniques, leading to sparse and underdetermined data that are very difficult or impossible to resolve. In the last few years computer simulations techniques have shown great promise in providing support for the interpretation of these data[8–10]. Molecular dynamics (MD) is a well - established method for the study the structure and function of biomolecules in their environment[11]. However, simulations of large proteins and IDPs is restricted by the availability of computational resources, which leads to a fundamental limit to the size of the system which can be studied. This also limits the timescales that can be sampled. Recent advances in HPC have allowed the study of larger and more diverse systems and the ever-increasing sophistication of protein force fields, molecular simulations have shown the ability to play a leading role in discovery[12–14]. Some recent and important studies of IDPs include the study of a multichain system of elastin. Elastin is a completely unstructured protein, however its function and elasticity come from its lack of structure[15]. In the past, short simulations in the nanosecond range were carried out on monomer systems or with up to 6 peptides at most. Recent work carried out by Rauscher et al[16] studied a system of 27 elastin chains to probe the aggregation dynamics showed that elastin exhibits a polymer melt-like disordered protein state. Other recent applications of MD in the field of IDPs include showing the disorder to order transition of an IDP upon phosphorylation[17]. This is of particular importance to my work, see **Chapter 5**, where I discuss the phosphorylation of the LxCxE containing HPV E7 peptide and its increase in PPII content upon phosphorylation[18]. Other work has shown that phosphorylation can increase the propensity to form helical[17] and beta structures[19], aside from PPII[20]. The limitation of these studies is that their result may be force field-dependent and this dependence is difficult to assess. Many force fields were benchmarked from single chain systems and work well for structured proteins. However, they have been shown to be somewhat lacking for the study of IDPs, as most ensembles are too compact and, in the specific case of AMBER force fields, they may over stabilise helical motifs[21]. There has been significant effort recently to address this problem and to tailor forcefields to describe IDPs[22]. However, there are example that this tailoring may cause a force field to lose its ability and accuracy to describe folded proteins[23].

In this thesis I will present the work I have done using molecular simulation techniques to understand if and how different degrees of disorder facilitate the many different biological functions that proteins' intrinsically disordered regions (IDR) perform. In particular my work has focused on determining the relative propensity of particular sequences, studied as peptides of different lengths, to form stable secondary structure motifs that can function as molecular recognition features (MoRFs). This 'pre-structuring' propensity, terminology often used in this thesis to signify the tendency of some sequences to form MoRFs, facilitates the recognition and binding from specific receptors, therefore limiting the entropic penalty upon binding[24]. Within the framework of molecular recognition theories[25], a structureless random coil would have to pay a maximum entropy penalty to fold-upon-binding after recognition, through a mechanism known as "induced fit", especially if the bound conformation is a distinct secondary structure such as a helical motif. Meanwhile, according to the "conformational selection" theory the disordered peptide would adopt in solution sub-structures identical to the final bound structure, minimizing the entropic penalty upon binding[26–28]. As we will discuss in detail, we found that in many of the cases we have studied, different degrees of pre-structuring exist and may play a role in recognition, within the framework of a mechanism that sits between conformational selection and induced fit. The regulation of the level of conformational disorder through sequence makes IDRs ideal counterparts in transient and reversible protein-protein interactions (PPIs), thus essential players in signalling and regulatory pathways[29–31] and as scaffolding proteins, coordinating the reversible formation of macromolecular assemblies[32,33].

It has been estimated that approximately 50% of the human proteome contains at least one IDR[34], yet these regions are rarely completely unstructured. As I will show in the work carried out in this thesis and as it has been shown previously[35] varying degrees of prestructuring can be observed in the unbound structure of peptides in solution. This prestructuring, in terms of both its degree and type, may be inherently linked to the function of the peptide[36]. Since IDP binding affinities can range between high µm to low nm, it can be difficult to determine the role of prestructuring in general. However, within the framework of kinetic binding experiments, see **Chapter 5**, prestructuring should affect primarily K_{on} and have little to no effect on K_{off}. The average rate of association of a protein-protein complex is estimated in the 10^4 – 10^6 $M^{-1}s^{-1}$ range[37]. Association rates can be classified as either diffusion limited, or conformationally limited. Below approximately 10^5 $M^{-1}s^{-1}$ the association is said to be conformationally limited and above this limit association is diffusion limited[38]. The association rate of IDPs varies but commonly they have a value in the range of 10^7 $M^{-1}s^{-1}$ [39]. A study of the coupled folding and binding of the IDP cMyb transactivation domain, which binds the KIX protein in a helix, has residual helicity of 30% via CD[37]. The K_{on} for this reaction is 2.2 ± 0.2 10^7 $M^{-1}s^{-1}$. If the 30% helicity corresponds to a fully formed binding motif the K_{on} would still be an order of magnitude lower than the threshold for diffusion limited association for a solely conformational selection-based association. It has also been proposed[37] that this protein may be suitable for probing the effects of disorder as there is little ionic dependence of the system. However, to my knowledge, no such study has been carried out so far.

As I will discuss in **Chapter 3**, sub-structures sampled within prestructuring can be stabilized by restraining the peptide using the “stapling” synthetic methodology. This strategy can in principle improve the peptide binding affinity by locking the active conformation[40]. Stapled peptides have been used to target several different proteins to act as a chemotherapeutics, namely targeting the MDM2 protein, which is over expressed in cancer cells and negatively regulates the p53 tumour suppressor[41], and MCL-1 which binds to an apoptosis inducing protein[42]. A study of stapled peptides derived from IDPs showed an increase in binding affinity of up to 2 orders of magnitude compared to the wild type peptide increasing binding affinity from µm into the nm range[40].

Protein secondary structure motifs are defined by the values of the phi (φ) and psi (ψ) torsion angles as shown in **Figure 1.1**. where their relative stabilities can be represented on a Ramachandran plot[43].

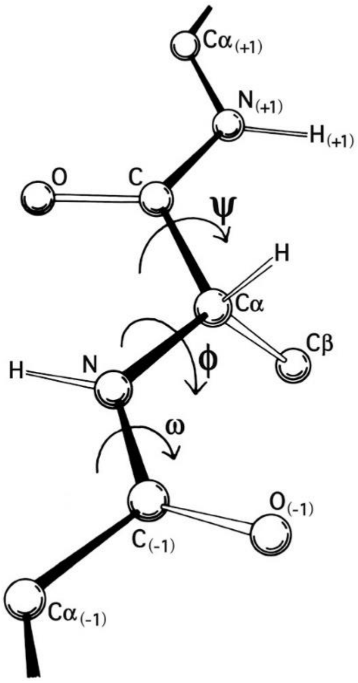


Figure 1.1: Diagram showing the phi psi and omega dihedrals of the protein backbone. ω is almost exclusively 180° except in proline where it adopts an angle of 0°[44]

In the example shown in **Figure 1.2**, the most thermodynamically stable secondary structure motifs correspond to φ/ψ torsions values indicated within specific regions delimited by contours on the diagram. During the study of different sequences discussed in details in the chapters ahead, we have found that MoRFs can be β structures, i.e. hairpins and strands, or helical turns, more commonly α or 3-10 turns, all closely related to the structure of the natural ligand in terms of RMSD values. In the specific case of the LxCxE short linear motif (SLiM)-containing peptides, discussed in detail in **Chapter 5**, we have found that phosphorylation dramatically enhances the propensity to form PPII structures. This is particularly interesting to us as it is in agreement with experimental data obtained for the hyperphosphorylation of Tau[45].

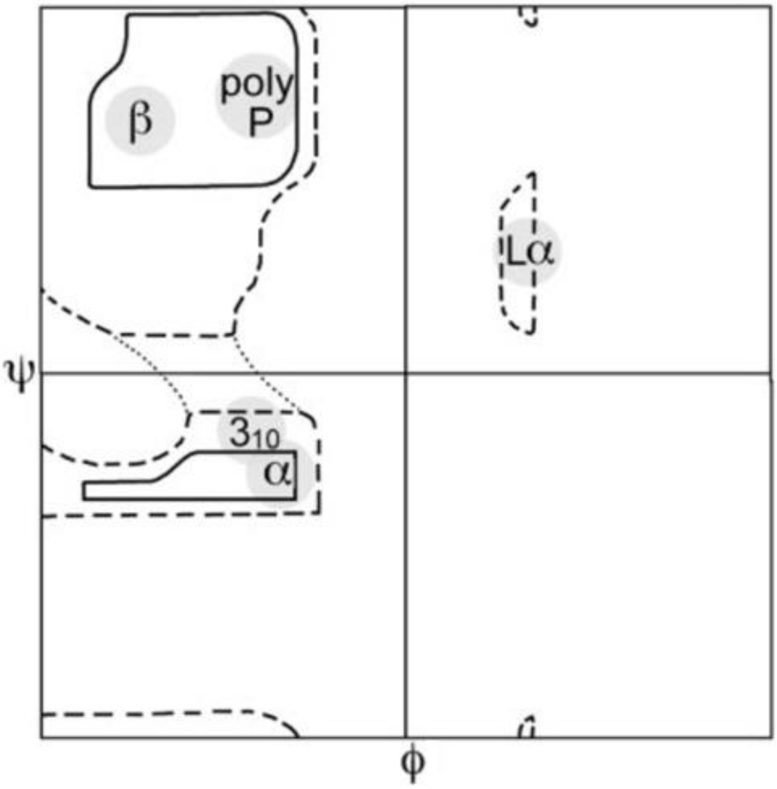


Figure 1.2: Ramachandran plot showing the phi/psi dihedrals values for α-helix (α) , β-sheet (β), 3_{10} helix (3_{10}), left handed α-helix (Lα) and polyproline helix (poly P), which is comprised of polyproline type I and II[46].

The overarching question I have tried to explore throughout my thesis work is, how disordered are disordered proteins? A simple disregard of intrinsic disorder does not justify the abundance of IDPs[47], the wide range of functions they have[3] and their specificity in their particular function[48]. Is it possible that different degrees of disorder facilitate different molecular functions? Through the study of all the different cases discussed in this thesis, it transpires that the high flexibility characterizing IDRs does not translate into a complete lack of structure. Indeed, we and others have found that short-lived structured motifs can be found within the disordered ensemble and can be interconverting too quickly to be clearly identified experimentally[47–49]. For example, we have found that the scaffolding protein XPA N- and C- terminal tails, which are completely disordered, form distinct nucleation sites[32,50–52] which serve as docking points for repair enzymes and proteins in the formation of the molecular assembly responsible for the DNA-damage excision in the Nucleotide Excision Repair (NER) pathway, see **Chapter 3**. The extreme C-terminal region of the p53 tumour suppressor (p53-CTD) is also classified as intrinsically disordered, but targeted by many receptors, making it a classic example of promiscuous binding target. Our work, detailed in **Chapter 4**, shows that the p53-CTD has the distinct propensity to form different secondary structure motifs, each significantly populated for recognition. This degree of diverse conformational propensity confers the p53-CTD its binding promiscuity, while preserving its binding specificity to each receptor. In **Chapter 5** we look at a different type of ID system, namely the short-linear motif (SLiM) LxCxE embedded within different peptide sequences that bind specifically the retinoblastoma protein (Rb). SLiMs are defined as short sequences of residues, usually 3 to 15 residues in length, which consist a few highly conserved residues interspersed with other less conserved or non-conserved residues[53]. Because of their short length, SLiMs generally bind in the lower µM range[54], but in case of the LxCxE motif-containing peptide the affinity for Rb reaches the nM range[18]. In this specific case we found that the LxCxE does not show any evidence of pre-structuring, possibly not seen as an advantageous feature due to the extremely short length of the sequence. Finally, in **Chapter 6** we discuss another case, where the propensity for pre-structuring within the human *vs.* murine ECSIT C-terminal tails is actually a clear defining structural difference between the two sequences, and possibly the feature that explains their very different stabilities. Details on the physics and on the algorithms behind biomolecular simulation methods are discussed in **Chapter 2**. Finally, concluding remarks provide a brief summary of the work discussed in the chapters and attempt to frame all the different degrees of intrinsic disorder within a self-contained interpretation to the classic structure-to-function relationship dogma of structural biology.

References

1. Deiana, A., Forcelloni, S., Porrello, A. & Giansanti, A. Intrinsically disordered proteins and structured proteins with intrinsically disordered regions have different functional roles in the cell. *PLoS One* **14**, e0217889 (2019).
2. Colak, R. *et al.* Distinct Types of Disorder in the Human Proteome: Functional Implications for Alternative Splicing. *PLOS Comput. Biol.* **9**, e1003030 (2013).
3. Dunker, A. K., Brown, C. J., Lawson, J. D., Iakoucheva, L. M. & Obradović, Z. Intrinsic Disorder and Protein Function. *Biochemistry* **41**, 6573–6582 (2002).
4. Romero, P. *et al.* Sequence complexity of disordered protein. *Proteins Struct.*

Funct. Bioinforma. **42**, 38–48 (2001).

5. Ishida, T. & Kinoshita, K. PrDOS: prediction of disordered protein regions from amino acid sequence. *Nucleic Acids Res.* **35**, W460–W464 (2007).
6. Linding, R. *et al.* Protein disorder prediction: implications for structural proteomics. *Structure* **11**, 1453–1459 (2003).
7. Garner, Cannon, Romero, Obradovic & Dunker. Predicting Disordered Regions from Amino Acid Sequence: Common Themes Despite Differing Structural Characterization. *Genome Inform. Ser. Workshop Genome Inform.* **9**, 201–213 (1998).
8. Boomsma, W., Ferkinghoff-Borg, J. & Lindorff-Larsen, K. Combining Experiments and Simulations Using the Maximum Entropy Principle. *PLOS Comput. Biol.* **10**, e1003406 (2014).
9. Lückmann, M. *et al.* Molecular dynamics-guided discovery of an ago-allosteric modulator for GPR40/FFAR1. *Proc. Natl. Acad. Sci.* **116**, 7123 LP – 7128 (2019).
10. Boon, P. L. S. *et al.* Partial Intrinsic Disorder Governs the Dengue Capsid Protein Conformational Ensemble. *ACS Chem. Biol.* **13**, 1621–1630 (2018).
11. Dror, R. O., Dirks, R. M., Grossman, J. P., Xu, H. & Shaw, D. E. Biomolecular Simulation: A Computational Microscope for Molecular Biology. *Annu. Rev. Biophys.* **41**, 429–452 (2012).
12. Zosel, F., Mercadante, D., Nettels, D. & Schuler, B. A proline switch explains kinetic heterogeneity in a coupled folding and binding reaction. *Nat. Commun.* **9**, 3332 (2018).
13. Ithuralde, R. E., Roitberg, A. E. & Turjanski, A. G. Structured and Unstructured Binding of an Intrinsically Disordered Protein as Revealed by Atomistic Simulations. *J. Am. Chem. Soc.* **138**, 8742–8751 (2016).
14. Ball, K. A. *et al.* Homogeneous and Heterogeneous Tertiary Structure Ensembles of Amyloid-β Peptides. *Biochemistry* **50**, 7612–7628 (2011).
15. Muiznieks, L. D., Weiss, A. S. & Keeley, F. W. Structural disorder and dynamics of elastin. *Biochem. Cell Biol.* **88**, 239–250 (2010).
16. Rauscher, S. & Pomès, R. The liquid structure of elastin. *Elife* **6**, e26526 (2017).
17. Rieloff, E. & Skepö, M. Phosphorylation of a Disordered Peptide—Structural Effects and Force Field Inconsistencies. *J. Chem. Theory Comput.* **16**, 1924–1935 (2020).
18. Chemes, L. B., Sánchez, I. E., Smal, C. & de Prat-Gay, G. Targeting mechanism of the retinoblastoma tumor suppressor by a prototypical viral oncoprotein. *FEBS J.* **277**, 973–988 (2010).
19. Wang, K., Ning, S., Guo, Y., Duan, M. & Yang, M. The regulation mechanism of phosphorylation and mutations in intrinsically disordered protein 4E-BP2. *Phys. Chem. Chem. Phys.* **22**, 2938–2948 (2020).
20. Brister, M. A., Pandey, A. K., Bielska, A. A. & Zondlo, N. J. OGlcNAcylation and phosphorylation have opposing structural effects in tau: phosphothreonine induces particular conformational order. *J. Am. Chem. Soc.* **136**, 3803–3816 (2014).
21. Rauscher, S. *et al.* Structural Ensembles of Intrinsically Disordered Proteins Depend Strongly on Force Field: A Comparison to Experiment. *J. Chem. Theory Comput.* **11**, 5513–5524 (2015).
22. Huang, J. *et al.* CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat. Methods* **14**, 71 (2016).
23. Robustelli, P., Piana, S. & Shaw, D. E. Developing a molecular dynamics force

field for both folded and disordered protein states. *Proc. Natl. Acad. Sci.* **115**, E4758–E4766 (2018).

24. Fuxreiter, M., Simon, I., Friedrich, P. & Tompa, P. Preformed structural elements feature in partner recognition by intrinsically unstructured proteins. *J. Mol. Biol.* **338**, 1015–1026 (2004).
25. Csermely, P., Palotai, R. & Nussinov, R. Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. *Trends Biochem. Sci.* **35**, 539–546 (2010).
26. Fadda, E. & Nixon, M. G. The transient manifold structure of the p53 extreme C-terminal domain: insight into disorder, recognition, and binding promiscuity by molecular dynamics simulations. *Phys. Chem. Chem. Phys.* **19**, 21287–21296 (2017).
27. Tompa, P., Szasz, C. & Buday, L. Structural disorder throws new light on moonlighting. *Trends Biochem. Sci.* **30**, 484–489 (2005).
28. Oldfield, C. J. *et al.* Flexible nets: disorder and induced fit in the associations of p53 and 14-3-3 with their partners. *BMC Genomics* **9 Suppl 1**, S1 (2008).
29. Mollica, L. *et al.* Binding Mechanisms of Intrinsically Disordered Proteins: Theory, Simulation, and Experiment. *Front. Mol. Biosci.* **3**, 52 (2016).
30. Wright, P. E. & Dyson, H. J. Intrinsically disordered proteins in cellular signalling and regulation. *Nat. Rev. Mol. Cell Biol.* **16**, 18 (2014).
31. Iakoucheva, L. M. *et al.* The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res.* **32**, 1037–1049 (2004).
32. Fadda, E. Role of the XPA protein in the NER pathway: A perspective on the function of structural disorder in macromolecular assembly. *Comput. Struct. Biotechnol. J.* **14**, (2015).
33. Babu, M. M., van der Lee, R., de Groot, N. S. & Gsponer, J. Intrinsically disordered proteins: regulation and disease. *Curr. Opin. Struct. Biol.* **21**, 432–440 (2011).
34. Niklas, K. J., Dunker, A. K. & Yruela, I. The evolutionary origins of cell type diversification and the role of intrinsically disordered proteins. *J. Exp. Bot.* **69**, 1437–1446 (2018).
35. Kim, D.-H. & Han, K.-H. PreSMo Target-Binding Signatures in Intrinsically Disordered Proteins. *Mol. Cells* **41**, 889–899 (2018).
36. De Sancho, D. & Best, R. B. Modulation of an IDP binding mechanism and rates by helix propensity and non-native interactions: association of HIF1α with CBP. *Mol. Biosyst.* **8**, 256–267 (2012).
37. Shammas, S. L., Travis, A. J. & Clarke, J. Remarkably fast coupled folding and binding of the intrinsically disordered transactivation domain of cMyb to CBP KIX. *J. Phys. Chem. B* **117**, 13346–13356 (2013).
38. Schlosshauer, M. & Baker, D. Realistic protein-protein association rates from a simple diffusional model neglecting long-range interactions, free energy barriers, and landscape ruggedness. *Protein Sci.* **13**, 1660–1669 (2004).
39. Shammas, S. L., Crabtree, M. D., Dahal, L., Wicky, B. I. M. & Clarke, J. Insights into Coupled Folding and Binding Mechanisms from Kinetic Studies. *J. Biol. Chem.* **291**, 6689–6695 (2016).
40. Miles, J. A. *et al.* Hydrocarbon constrained peptides – understanding preorganisation and binding affinity. *Chem. Sci.* **7**, 3694–3702 (2016).
41. Ali, A. M., Atmaj, J., Van Oosterwijk, N., Groves, M. R. & Dömling, A. Stapled Peptides Inhibitors: A New Window for Target Drug Discovery. *Comput. Struct. Biotechnol. J.* **17**, 263–281 (2019).

42. Joseph, T. L., Lane, D. P. & Verma, C. S. Stapled BH3 Peptides against MCL-1: Mechanism and Design Using Atomistic Simulations. *PLoS One* **7**, e43985 (2012).
43. Ramachandran, G. N., Ramakrishnan, C. & Sasisekharan, V. Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.* **7**, 95–99 (1963).
44. Richardson, J. S. Protein backbone PhiPsiOmega drawing. (2011). Available at: https://commons.wikimedia.org/wiki/File:Protein_backbone_PhiPsiOmega_drawing.jpg. (Accessed: 7th October 2019)
45. Bielska, A. A. & Zondlo, N. J. Hyperphosphorylation of Tau Induces Local Polyproline II Helix. *Biochemistry* **45**, 5527–5537 (2006).
46. Richardson, J. S. Ramachandran plot original outlines. (2011). Available at: https://commons.wikimedia.org/wiki/File:Ramachandran_plot_original_outlines.jpg. (Accessed: 7th October 2019)
47. Dunker, A. K., Obradovic, Z., Romero, P., Garner, E. C. & Brown, C. J. Intrinsic protein disorder in complete genomes. *Genome Inform. Ser. Workshop Genome Inform.* **11**, 161–171 (2000).
48. Tsodikov, O. V *et al.* Structural basis for the recruitment of ERCC1-XPF to nucleotide excision repair complexes by XPA. *EMBO J.* **26**, 4768–4776 (2007).
49. Choi, U. B., McCann, J. J., Weninger, K. R. & Bowen, M. E. Beyond the random coil: stochastic conformational switching in intrinsically disordered proteins. *Structure* **19**, 566–576 (2011).
50. Sugitani, N., Sivley, R. M., Perry, K. E., Capra, J. A. & Chazin, W. J. XPA: A key scaffold for human nucleotide excision repair. *DNA Repair (Amst).* **44**, 123–135 (2016).
51. Fadda, E. The role of conformational selection in the molecular recognition of the wild type and mutants XPA 67-80 peptides by ERCC1: Molecular Recognition of XPA 68-80 Peptide Mutants. *Proteins* **83**, (2015).
52. Fadda, E. Conformational Determinants for the Recruitment of ERCC1 by XPA in the Nucleotide Excision Repair (NER) Pathway: Structure and Dynamics of the XPA Binding Motif. *Biophys. J.* **104**, 2503–2511 (2013).
53. Davey, N. E. *et al.* Attributes of short linear motifs. *Mol. Biosyst.* **8**, 268–281 (2012).
54. Diella, F. *et al.* Understanding eukaryotic linear motifs and their role in cell signaling and regulation. *Front. Biosci.* **13**, 6580–6603 (2008).

Chapter 2: Computational Method

In this chapter I describe the methodology which underpins the work I carried out in this thesis. In the first section I will discuss the basic of the molecular mechanics approach to atomistic simulations**.** Secondly, I will discuss the molecular dynamics simulation method and the options available to the user in setting up an MD simulation. Following that I will outline the use of enhanced sampling methods for MD, in particular REMD. Finally, I will discuss binding free energy calculations, and the application of MMGBSA as an end-point free energy calculation method.

Molecular Mechanics and Empirical Force Fields. Within a molecular mechanics (MM) formalism, atoms are described in as 'classical' objects, namely as hard, impenetrable spheres, which behaviour and energetic properties are determined by an empirical force field. In a classical force field covalent (bonded) interactions, namely bond length and bond angles and torsions, are approximated using Hooke's law, where torsions are treated as a sinusoidal function. Non-bonded interactions are divided into electrostatic and van der Waals (vdW), modelled using the Coulomb and Lennard-Jones potentials, respectively. These contributions are summarised in **Equation 2.1**.

$$\begin{aligned} V(r^N) = {} & \sum_{bonds} \frac{K_i}{2}\left(l_i - l_{i,o}\right)^2 + \sum_{angles} \frac{K_i}{2}\left(\theta_i - \theta_{i,0}\right)^2 \\ & + \sum_{torsions} \sum_{n} \frac{V_n}{2}\left[1 + \cos(n\omega - \gamma)\right] \\ & + \sum_{i}^{n} \sum_{j=i+1}^{n} \left\{ 4\varepsilon_{i,j}\left(\left[\frac{\sigma_{ij}}{r_{ij}}\right]^{12} - \left[\frac{\sigma_{ij}}{r_{ij}}\right]^{6}\right) + \frac{q_i q_j}{4\pi\varepsilon_o r_{i,j}} \right\} \end{aligned} \tag{2.1}$$

Atoms in a force field are assigned an "atom type", based on its nature and chemical bonding environment. The atom type is descriptive of the properties of that atom. In an empirical force field, bonding and non-bonding parameters assigned to each atom type are derived from experimental, from quantum mechanical (QM) calculations or from both and are developed to reproduce properties of the atoms. Most force fields, especially the best known and most widely used ones, such as AMBER99SB-ildn[1], have been successful in simulating the properties of folded proteins. However, many of the force fields developed to describe folded proteins have been found not to be suitable to describe unfolded, also known as disordered, proteins, which nowadays are a field of ever-increasing interest in science and biophysics. In fact, a study of the most popular force fields[2] at the time, shows that simulations done using different force fields produced a high variability in chain dimension and that CHARMM36[3] during unusually long simulation times generated left-handed helices, which are rarely or never encountered in known protein structures deposited in the RCSB Protein Data Bank (PDB). During the past few years there has been an increased interest in developing force fields that can accurately reproduce the properties of intrinsically disordered proteins (IDPs). Among those, as part of the AMBER family of force fields, a99SB-disp[4] has been developed specifically to simulate IDPs and has been obtained from a99SB*-ILDN[1] by rescaling the protein torsion parameters, adjusting the carbonyl oxygen amide hydrogen Lennard-Jones potential and adjusting the protein-water vdW interactions. Results show that the a99SB-disp forcefield appears to suitable for simulations of both, structured and disordered proteins[4]. From the CHARMM family of force fields, CHARMM36m[5] has also been developed specifically to describe IDPs and it was obtained from the CHARMM36 parameter set, where the vdW radius of the backbone Cα was adjusted to eliminate the occurrence of left handed α-helices. The CHARMM36m parameters were specifically tuned to closely match the experimental SAXS and NMR data[5]. Several other forcefields that have been developed for IDP simulations[6–10].

In addition to parameter sets that describe protein atoms, the quality of a simulation greatly depends on the choice of water force field. Ideally, the chosen water model should be the one that was used in the parameterization and validation of the protein force field itself, however the use of different combinations of protein and water force fields, which are not necessarily compatible, or which compatibility has not been proven, are commonly used. The most common water models are transferable intermolecular potential with three points and 4 points models, known as TIP3P and TIP4P, respectively[11]. Models derived from these are also frequently used[12,13]. Because of the importance of the water parameters, there has been some effort in the development of water force fields specifically aimed at improving the accuracy of IDPs simulations. For example, work carried out on the unfolded N-terminal Zn-binding domain of HIV-1 integrase[13], showed that the TIP3P and TIP4P-Ew and TIP4P/2005 water models all generate overly compact ensemble averages with R_g

similar to the folded state rather than to the unfolded state. Based on these results, the TIP4P-D water model was developed to determine an increased dispersion interaction potential, which was found to alleviate the over compactness and to reproduces more closely the experimental results[13] . As stated in a personal communication from Dr. Paul Robustelli first author on the original work[13], a significant drawback of the TIP4P-D water model is that the increased dispersion interactions cause the unphysical unfolding of known structured motif. We also demonstrated this shortcoming of the TIP4P-D model with results shown in **Chapter 5**. As the reader will see in the following chapters, rather than testing each possible combination of parameter sets, for the work presented in this thesis we used a limited number of different force fields and of different combinations of water and protein models. Our choice has been fundamentally based on "the best reputed" force field to describe IDPs at the time the work was undertaken, where the study, as a whole, spans literally four years. Nevertheless, because of the ever-growing production of new IDP force fields and specifically tailored approximations, we would have never been able to simulate exhaustively every system we studied with every combination of IDP force field and water model. Rather we focused on performing exhaustive sampling and carefully treating the simulation conditions based on the information we had on the limitation of the models. As shown in **Chapter 6**, we find that through sufficient sampling two historically accurate force field even so slightly adapted for the correct treatment of intrinsic disorder, produce the same results.

Molecular Dynamics. In the work presented in this thesis, we used both classical (conventional) molecular dynamics (MD), as well as enhanced sampling in the form of replica-exchange MD to perform conformational sampling. MD studies the evolution of a system over time. A force field is used to calculate the potential energy of such system using **Equation 2.1,** where the forces acting on the atoms are obtained through **Equation**

2.2**.**

$$F = -\frac{dV}{dr} \qquad 2.2$$

As such, the accelerations of the atoms can be calculated using Newton's second law of motion, shown in **Equation 2.3**

$$F = ma \tag{2.3}$$

Where *m* is the atom mass. In order to calculate how the positions and velocities of particles change over time Newton's equation of motion is to be integrated numerically. As an example of a classic integrator, the Verlet algorithm, which is derived from **Equations 2.4 and 2.5,** is shown in **Equation 2.6**, which uses positions and accelerations at time *t* and positions at time *t-δt* to calculate new positions at time *t+δt.*

$$r(t + \delta t) = r(t) + v(t)\delta t + \frac{1}{2}a(t)\delta t^2 + \cdots \tag{2.4}$$

$$r(t - \delta t) = r(t) - v(t)\delta t + \frac{1}{2}a(t)\delta t^2 - \cdots \tag{2.5}$$

$$r(t + \delta t) = 2r(t) - r(t - \delta t) + \ a(t)\delta t^2 \tag{2.6}$$

As an alternative, in the leapfrog integrator[14], shown in **Equations 2.7 - 2.8**, the velocities are calculated at time *t+1/2δt* and the accelerations are calculated at time *t*, while the positions are calculated at *r(t+δt)* using *r(t)* and *v(t+1/2δt)*, so that velocities and positions "leap-frog" over each other.

$$r(t + \delta t) = r(t) + v\left(t + \frac{1}{2}\delta t\right)\delta t \tag{2.7}$$

$$\mathrm{v}\left(\mathrm{t} + \frac{1}{2}\delta\mathrm{t}\right) = \mathrm{v}\left(\mathrm{t} - \frac{1}{2}\delta\mathrm{t}\right) + \mathrm{a}(\mathrm{t})\delta\mathrm{t} \tag{2.8}$$

The advantage of the Leap-Frog over the Verlet integrator is that the velocities are explicitly calculated, even if not at the same time as the positions. In this work we use a stochastic dynamics leap-frog integrator (sd)[15], which includes into the integrator a friction term. Other integrators are available such as the velocity Verlet algorithm[16] or the Runge-Kutta method[17]. As we sampled in the NPT ensemble, the system needs to be brought at target pressure and temperature, thus a suitable thermostat and barostat must be used to equilibrate and maintain the system at the desired temperature and pressure. The leap-frog stochastic dynamics integrator also acts as the thermostat and we used a Berendsen barostat to regulate pressure. The Berendsen barostat has been shown to have issues with calculating compressibility, it has been described as "simply wrong" in simulations where the fluctuation of volume is important[18] . As such, use of a different barostat, such as the Parrinello-Rahman barostat[19], which more accurately calculates compressibility[18], is recommended. All REMD simulations were run in the NVT ensemble, where we only required a thermostat.

A crucial aspect of MD simulations is the size of the time-step as it determines how long the simulation will take to run based on the computational resources available. The time step must be an order of magnitude shorter than the shortest vibration of the system. This is usually the vibration of the covalent bond between a hydrogen and a carbon atom, which is approximately 10 fs, corresponding to a time step of 1 fs. Most simulation setups constrain this vibration allowing a longer time step of 2 fs. Two methods to constrain these bonds are the SHAKE[20] and LINCS[21] algorithms, implemented in AMBER and GROMACS MD packages, respectively. While the use of a time step that is too short will be inefficient and slow down the simulations unnecessarily, provided LINCS or SHAKE algorithms are used, a time-step that is too long will cause overlap of the vdW spheres or large electrostatic repulsions that will lead to instability and erroneous behaviours.

A general protocol for running an MD simulation begins with the choice of an initial conformation of the system under study, which would ideally correspond to an NMR or a crystal structure. In case of disordered peptides, we do not have experimental structures to rely on, therefore the peptide should be built in a chosen conformation or in the fully extended conformation. As described in each specific chapter, we chose to build peptides in extended conformations and to use short simulations to generate a single, or multiple starting structures for the MD study. After the initial conformation is chosen, the target molecule(s) is placed in a box with water, where ions are added to neutralise the system and to reach the desired ionic strength. The positions of the water molecules and ions are then minimised, where the structure of the target molecule(s) is constrained. This step allows the water molecules to reorient themselves around the solute molecule. Different equilibration steps follow where the system is brought up to the desired temperature and pressure and the conformation is has reached as stable starting point. The production phase of the conformational search begins here when the properties of interest can be collected and analysed.

Enhanced sampling. The two main factors that limit the accuracy of conventional sampling through MD simulations are the accuracy of the force field, which I discussed briefly earlier within the context of IDPs, and the computational cost of simulations that can greatly limit their duration thus search power. When the potential energy landscape of biomolecules presents many local minima with potentially large energy barriers separating them, sufficient sampling of the whole conformational space of the target biomolecule may require the implementation of enhanced sampling simulation methods. Enhanced sampling methods are a class of molecular simulation techniques developed to provide the target system with enough energy to escape potential energy wells and overcome barriers. Examples of these techniques include REMD[22], metadynamics[23], and umbrella sampling[24]. In this work we used REMD, for which the basics are outlined below.

The REMD algorithm developed by Sugita and Okamoto[22] is a widely used enhanced sampling method that combines the parallel tempering method used in Monte Carlo simulations[25] with an MD approach. The general REMD protocol is summarised in **Figure 2.1**.

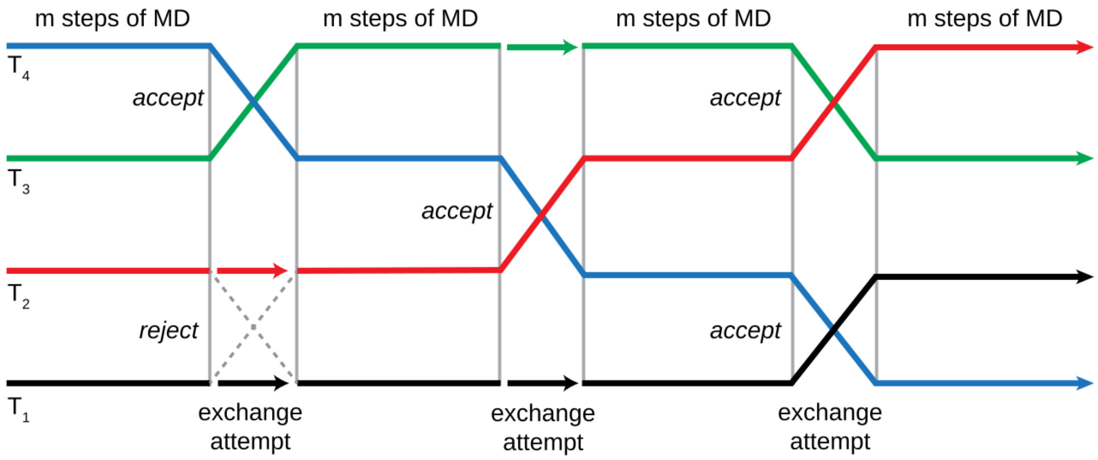


Figure 2.1: Overview of temperature REMD, showing exchanges being attempted after every m simulation steps. Exchanges can only happen between adjacent temperatures[26].

REMD is usually implemented using temperature as an exchange variable, but other types exist, such as Hamiltonian replica exchange[27]. Within the framework of temperature replica exchange, multiple isothermal (conventional) MD simulations are run in parallel sorted by increasing temperatures in the range (T_0,T_1,…,T_{n-1}). After every *m* number of steps the system attempts to swap replicas *i* and *j* with an acceptance ratio dependent on the difference in energy between replicas, as shown in **Equation 2.9**.

$$\min\left\{1, \mathrm{e}^{(\mathrm{E_i}-\mathrm{E_j})*\left(\frac{1}{kT_i}-\frac{1}{kT_j}\right)}\right\} \tag{2.9}$$

The larger the difference between temperatures the least likely will be the exchange probability. When deciding the temperature distribution, it is best to have a geometric series of temperatures, so that the ratio of temperatures between all pairs of adjacent replicas is the same and each replica should visit every temperature during the simulation[28,29]. The advantage of REMD over other enhanced sampling techniques is that it is relatively easy to set up compared to metadynamics and umbrella sampling, where one has to consider their collective variables and bias potentials carefully. A disadvantage of temperature REMD is the relatively high simulation cost, due to a large number of replicas being required for protein simulations in explicit solvent[30].

Binding Free Energy Calculations. In some of the studies presented in this thesis we have calculated binding free energies, see Chapter 5. Binding free energy calculations are extremely informative within the context of providing support in the interpretation and design of experiments. Free energy perturbation (FEP) and thermodynamic integration (TI) are highly accurate binding free energy calculation methods, but very computationally expensive and restricted to cases where the differences in conformation between initial (unbound) and final (bound) states are negligible. Because of the large conformational ensemble commonly sampled by IDPs, the use of binding free energy calculations based on perturbation theory is unfeasible in terms of convergence. End point methods such as MM/PB(GB)SA offer a more viable, less computationally expensive alternative for the calculation of relative binding affinities of interactions involving IDPs. In Chapter 5 we present work where binding free energies of short peptides were estimated by MM/GBSA. The general thermodynamic cycle on which the MM/GBSA approach is based is shown in

Figure 2.2, with the corresponding contributions to the free energy shown in **Equation 2.10**.

$$\Delta G_b^s = \Delta G_b^g + \Delta G_{sol}^C - (\Delta G_{sol}^L + \Delta G_{sol}^R) \tag{2.10}$$

The MM/GBSA calculation consists in three parts: 1) a molecular mechanics (MM) contribution that gives the total energy for the bonded and non-bonded interactions

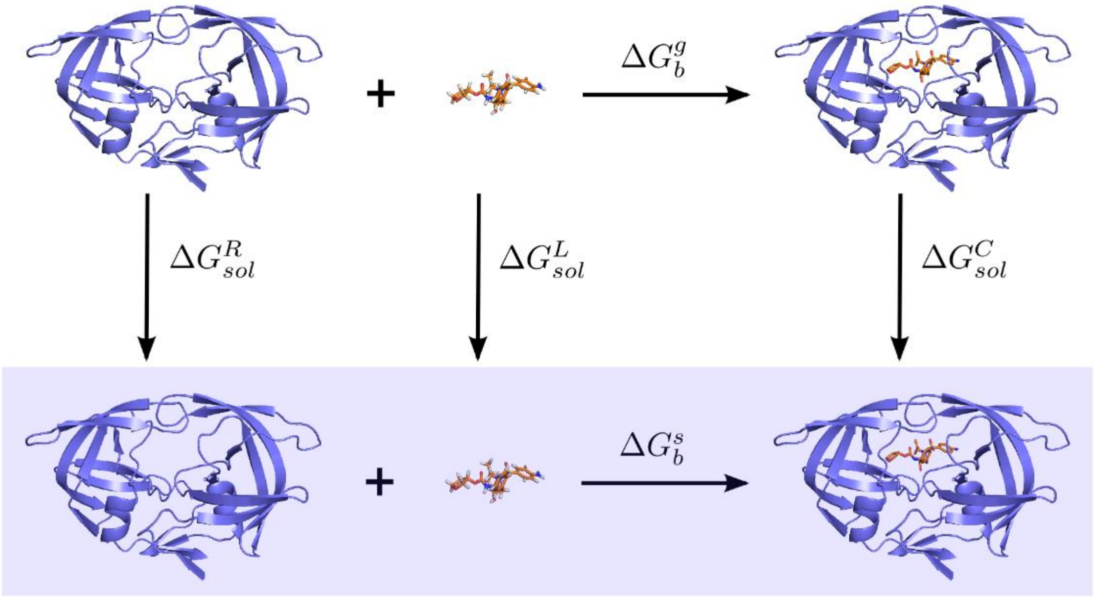


Figure 2.2: General thermodynamic cycle for MM/GB(PB)SA showing the ligand and receptor, and complex in the gas phase and solution phase. Also shown are the free energies of solvation of the ligand receptor and complex and the binding free energy of the complex in gas phase and solution phase[31].

from **Equation 2.1**, 2) the desolvation free energy calculated as a polar contribution given by solving the Poisson Boltzmann (PB) equation, or the generalised born (GB) equation, 3) the non-polar contribution to desolvation free energy is obtained from a linear equation which is dependent on solvent accessible surface area (SASA). The solvation binding free energy, in **Equation 2.10**, is calculated as the sum of these contributions, namely a gas phase enthalpic contribution from the MM forcefield and a solvation free energy contribution from the sum of the PB or GB free energy and the SA free energy.

References

1. Lindorff-Larsen, K. *et al.* Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* **78**, 1950–1958 (2010).
2. Rauscher, S. *et al.* Structural Ensembles of Intrinsically Disordered Proteins Depend Strongly on Force Field: A Comparison to Experiment. *J. Chem. Theory Comput.* **11**, 5513–5524 (2015).

3. Huang, J. & MacKerell, A. D. J. CHARMM36 all-atom additive protein force field: validation based on comparison to NMR data. *J. Comput. Chem.* **34**, 2135–2145 (2013).

4. Robustelli, P., Piana, S. & Shaw, D. E. Developing a molecular dynamics force field for both folded and disordered protein states. *Proc. Natl. Acad. Sci.* **115**, E4758–E4766 (2018).

5. Huang, J. *et al.* CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat. Methods* **14**, 71 (2016).

6. Song, D. *et al.* ff14IDPs force field improving the conformation sampling of intrinsically disordered proteins. *Chem. Biol. Drug Des.* **89**, 5–15 (2017).

7. Ye, W., Ji, D., Wang, W., Luo, R. & Chen, H.-F. Test and Evaluation of ff99IDPs Force Field for Intrinsically Disordered Proteins. *J. Chem. Inf. Model.* **55**, 1021–1029 (2015).

8. Wang, W., Ye, W., Jiang, C., Luo, R. & Chen, H.-F. New Force Field on Modeling Intrinsically Disordered Proteins. *Chem. Biol. Drug Des.* **84**, 253–269 (2014).

9. Liu, H., Song, D., Lu, H., Luo, R. & Chen, H.-F. Intrinsically disordered protein-specific force field CHARMM36IDPSFF. *Chem. Biol. Drug Des.* **92**, 1722–1735 (2018).

10. Best, R. B., Zheng, W. & Mittal, J. Balanced Protein-Water Interactions Improve Properties of Disordered Proteins and Non-Specific Protein Association. *J. Chem. Theory Comput.* **10**, 5113–5124 (2014).

11. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926–935 (1983).

12. Horn, H. W. *et al.* Development of an improved four-site water model for biomolecular simulations: TIP4P-Ew. *J. Chem. Phys.* **120**, 9665–9678 (2004).

13. Piana, S., Donchev, A. G., Robustelli, P. & Shaw, D. E. Water Dispersion Interactions Strongly Influence Simulated Structural Properties of Disordered Protein States. *J. Phys. Chem. B* **119**, 5113–5123 (2015).

14. Goga, N., Rzepiela, A. J., de Vries, A. H., Marrink, S. J. & Berendsen, H. J. C. Efficient Algorithms for Langevin and DPD Dynamics. *J. Chem. Theory Comput.* **8**, 3637–3649 (2012).

15. Van Gunsteren, W. F. & Berendsen, H. J. C. A Leap-frog Algorithm for Stochastic Dynamics. *Mol. Simul.* **1**, 173–185 (1988).

16. Swope, W. C., Andersen, H. C., Berens, P. H. & Wilson, K. R. A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. *J. Chem. Phys.* **76**, 637–649 (1982).

17. Janezic, D. & Orel, B. Implicit Runge-Kutta method for molecular dynamics integration. *J. Chem. Inf. Comput. Sci.* **33**, 252–257 (1993).

18. Shirts, M. R. Simple Quantitative Tests to Validate Sampling from Thermodynamic Ensembles. *J. Chem. Theory Comput.* **9**, 909–926 (2013).

19. Parrinello, M. & Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.* **52**, 7182–7190 (1981).

20. Ryckaert, J.-P., Ciccotti, G. & Berendsen, H. J. C. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* **23**, 327–341 (1977).

21. Hess, B., Bekker, H., Berendsen, H. J. C. & Fraaije, J. G. E. M. LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem.* **18**, 1463–1472 (1997).

22. Sugita, Y. & Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **314**, 141–151 (1999).

23. Laio, A. & Parrinello, M. Escaping free-energy minima. *Proc. Natl. Acad. Sci.* **99**, 12562 LP – 12566 (2002).

24. Torrie, G. M. & Valleau, J. P. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.* **23**, 187–199 (1977).

25. Swendsen, R. & Wang, J.-S. Replica Monte Carlo Simulation of Spin-Glasses. *Phys. Rev. Lett.* **57**, 2607–2609 (1986).

26. Rowley, C. Schematic of a replica exchange molecular dynamics simulation.

27. Okamoto, Y. Generalized-ensemble algorithms: enhanced sampling techniques for Monte Carlo and molecular dynamics simulations. *J. Mol. Graph. Model.* **22**, 425–439 (2004).

28. Ballard, A. J. & Jarzynski, C. Replica exchange with nonequilibrium switches. *Proc. Natl. Acad. Sci.* **106**, 12224 LP – 12229 (2009).

29. Rosta, E. & Hummer, G. Error and efficiency of replica exchange molecular dynamics simulations. *J. Chem. Phys.* **131**, 165102 (2009).

30. Meli, M. & Colombo, G. A Hamiltonian replica exchange molecular dynamics (MD) method for the study of folding, based on the analysis of the stabilization determinants of proteins. *Int. J. Mol. Sci.* **14**, 12157–12169 (2013).

31. Kuhn, O. Mmpbsa cycle. Available at: https://commons.wikimedia.org/wiki/File:Mmpbsa_cycle.png.

Chapter 3: Conformational analysis of XPA67-80 peptide homologues and design of high-affinity macrocyclic XPA67-80 derivatives

3.1 Introduction

The nucleotide excision repair (NER) pathway is responsible for repairing bulky DNA lesions, such as pyrimidine dimers caused by UV damage and crosslinked adducts caused by alkylating agents[1]. These lesions all destabilise the DNA double helix[2]. This damage is detected through stalling of RNA polymerase, which initiates transcription-coupled NER (TC-NER), or through damage sensing proteins, which initiates global-genomic NER (GG-NER)[2]. After damage recognition, both GG-NER and TC-NER proceed through the same pathway. NER begins with recognition of damage, followed by unwinding of the DNA surrounding the lesion and incision at both 3′ and 5′ ends. The lesion containing strand is then removed and the missing bases are replaced and ligated[2,3]. The main steps leading to the damage excision in the GG-NER pathway are summarized in **Figure 3.1**. Helical distortions resulting from bulky DNA adducts are detected by XPC–RAD23B complex and GG-NER is initiated[2–4]. As shown in **Figure 3.1**, panel b, TFIIH, which is composed of two helicases XPB and XPD, is recruited by XPC–RAD23B to the damage site[5]. TFIIH unwinds the DNA helix, exposing the strand carrying the lesion, a structure known as a DNA bubble[1]. XPA is recruited at this stage along with RPA through an interaction with the XPC N-terminal domain[6,7], as shown in **Figure 3.1**, panel c[9]. XPA interacts with TFIIH and TFIIH may be involved in XPA recruitment, or in the recruitment of the XPA–RPA complex[10]. The XPA–RPA causes the dissociation of the XPC–RAD23B from the damage site[11]. As shown in **Figure 3.1**, panel d, XPA performs a scaffolding role by docking the 5′ endonuclease, ERCC1-XPF dimer, to the junction of the single and double strand DNA[1,2,13,14]. The XPG endonuclease (not shown) which cleaves DNA at the 3′ end is likely recruited to the damage site by TFIIH[2,16,17]. Excision begins with ERCC1–XPF at the 5′ side followed by the XPG at the 3′ side[19]. The damaged oligonucleotide is released along with TFIIH[20], which causes XPA to unbind and the missing bases to be replaced and ligated[2].

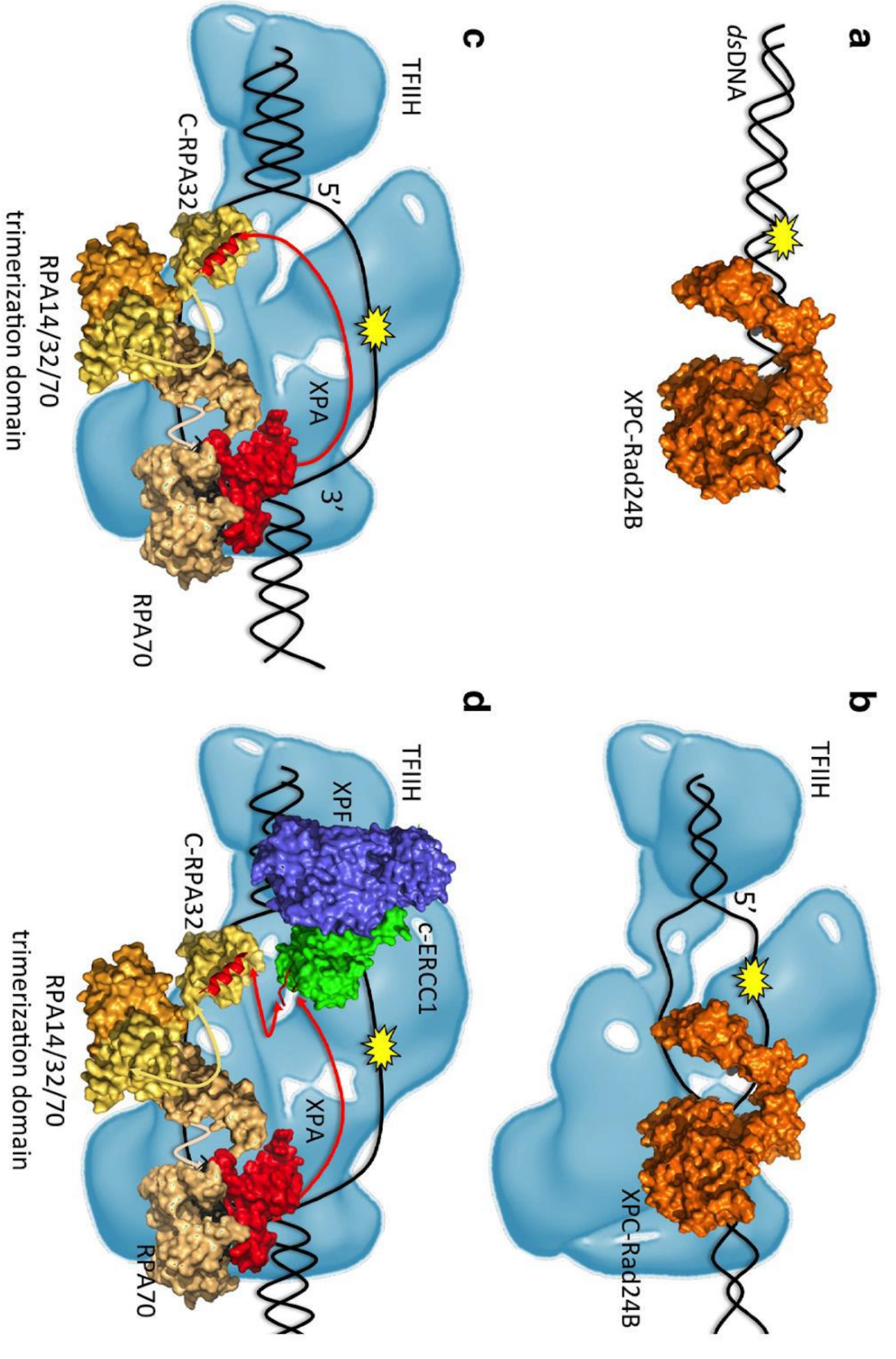


Figure 3.1: Graphical rendition the macromolecular assembly of NER proteins during the pre-excision phase of the DNA damage, shown in yellow. a) DNA damage is detected by the XPC-RAD24B dimer. b) TFIIH is recruited by XPC-RAD24B which opens the DNA bubble to expose the DNA damage. c) XPA shown in red and RPA shown in tan are recruited, releasing XPC-RAD24B. The disordered region of XPA is represented by a red line connecting it to RPA32. d) XPA recruits the ERCC1-XPF endonuclease, shown in green and purple respectively, to the 5′ side of the lesion. Figure adapted from reference 21

The ERCC1-XPF heterodimer is responsible for cleavage of bulky DNA lesions at the 5’ end of the DNA[1]. *Xeroderma pigmentosum* group A (XPA) is a key protein involved in the NER pathway. Its role is to provide the scaffolding for the assembly of the pre-excision multi-protein complex[2]. In this context XPA is known to bind the DNA double helix at the 3’ junction through its globular, folded core, i.e. between residues 98 to 219[22]. The ERCC1 binding region of XPA is located between residues 67 and 80, which is part of the unstructured N-terminal region of the protein[23]. A peptide with sequence corresponding to this the ERCC1 binding region of XPA, namely XPA_{67-80}, was found to be sufficient to competitively inhibit the XPA-ERCC1 interaction[23]. NMR data show that the XPA_{67-80} peptide is unstructured in solution[23], meanwhile, as shown in **Figure 3.2**, when bound to ERCC1 is in a β-hairpin conformation. Cisplatin is an anti-cancer drug used in the treatment of different type of cancers, such as testicular cancer, and it acts by forming an adduct with DNA by cross linking two guanine bases[24]. NER is the repair pathway responsible for excision of Pt adducts. ERCC1 concentration is related to cisplatin drug resistance, and the overexpression of ERCC1 is linked to cisplatin resistance in cancer cells[25]. The use of microarray technology with a high affinity binder for ERCC1 could be used to qualitatively determine ERCC1 expression levels in cells and, as such, as a diagnostic tool for whether to prescribe Pt-based drugs to a patient.

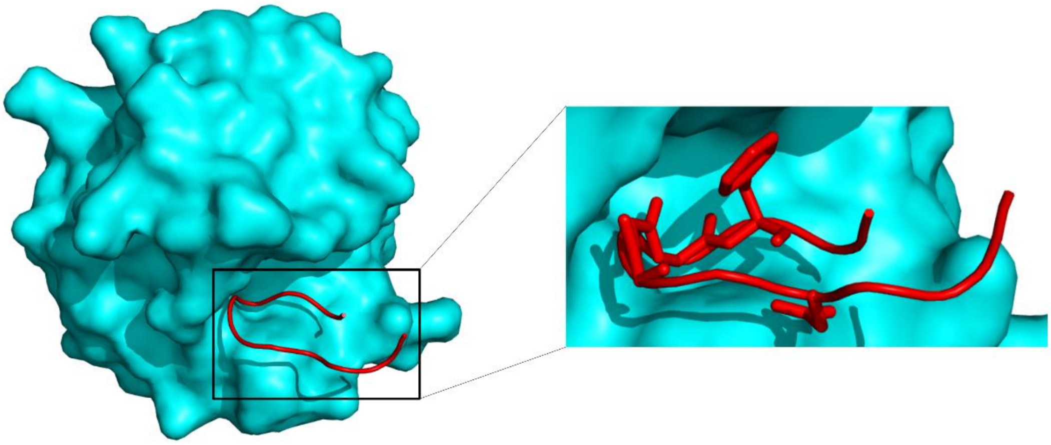


Figure 3.2: NMR structure (PDBid 2JNW) of XPA_{67-80} shown in red bound to ERCC1, represented as its solvent accessible surface area, shown in cyan. XPA peptide in the NMR structure includes only residues 67 to 77, as residues 78 to 80 are too flexible to be localized. The inset image shows the residues of XPA that are involved in interactions with ERCC1 as sticks, namely Asp 79, Gly 72 to 74 and Phe 75.

Conformational sampling of the free peptide through conventional MD simulations shows that the XPA$_{67-80}$ peptide can adopt β-hairpin structures stable at the low μs timescale[26,27]. This β-hairpin may behave as a MoRF, where this hairpin motif can be recognised by ERCC1 causing XPA to adopt its binding conformation[28,29]. Previous simulation work[26,27] also highlighted the role of residues Asp 70, Gly 72 to 74 and Phe 75 in establishing direct contacts with the ERCC1 binding site, as well as the role of Lys 67 and Glu 78 to 80 in stabilizing the β-hairpin conformation of the free peptide in solution. More specifically, Phe 75 interacts with the ERCC1 Asn 110 sidechain forming a π-stacking interaction. Gly 72 forms a hydrogen bond between its backbone carbonyl oxygen and the sidechain of ERCC1 Arg 156, Gly 73 forms a hydrogen bond between its backbone carbonyl oxygen and ERCC1 Gln 107 sidechain amide, Gly 74 forms a hydrogen bond between its backbone carbonyl oxygen and ERCC1 Ser 142 backbone amide nitrogen. Asp 70 forms hydrogens bonds with both ERCC1 Tyr 145 sidechain hydroxyl and with ERCC1 His 149 Nε. In previous work from the group[27] different mutants were designed to investigate further the role of key residues in the recognition and binding of the XPA$_{67-80}$ peptide by ERCC1. More specifically, the F75W mutation was designed to increase the aromatic surface area and potentially improve on the stacking interaction with the ERCC1 Asn 110, meanwhile the I68K mutant was designed to increase the stability of the β-hairpin conformation of the peptide when free in solution by enhancing the electrostatic interaction between the N and C termini of the peptide[27]. However, extensive sampling simulations showed that both of these mutations changed the conformational propensity of the peptide in solution, dramatically reducing the stability of the recognized β-hairpin conformer[27]. Although these results do not indicate that the mutations won't be tolerated at all, they do not suggest either that the binding affinity for ERCC1 would be increased relative to the wild type XPA$_{67-80}$. With these results in mind we set out to evolve this XPA$_{67-80}$ motif into a high affinity binder with the potential for use in a diagnostic tool to determine cisplatin resistance. To evolve this sequence from a low affinity binding peptide to a high affinity binder we looked at two different strategies to modify the peptide. According to the first strategy, we studied XPA sequences from selected species with functional NER pathways, shown in **Table 3.1**, to evaluate the effects of mutations relative to *H. sapiens* on the structural propensity of the peptide and on its binding interaction with ERCC1. The second strategy concerns chemically restricting the degrees of freedom of the XPA_{67-80} peptide to decrease its entropic penalty upon binding by designing suitable linkers that would restrain its conformation in its β-hairpin MoRF conformation. Both of these strategies have highlighted potential ways to increase the peptide's binding affinity for ERCC1

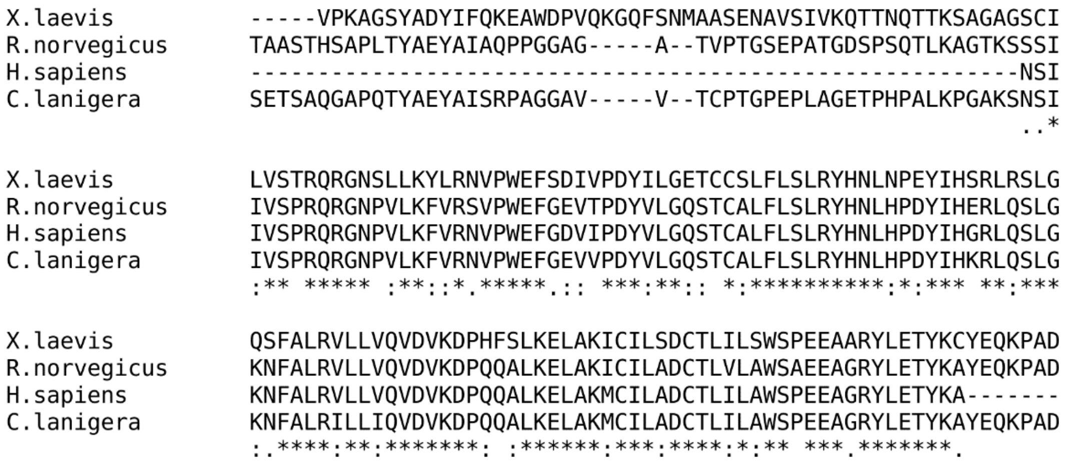
Table 3.1: Wild type sequences of XPA_{67-80} from select species with mutated residues relative to the H. sapiens sequence highlighted in red. Complementary mutations found in the ERCC1 protein sequences from the same species relative to the H. sapiens are also indicated.

Species	XPA_{67-80} sequence	Complementary ERCC1 mutation
H. sapiens	KIIDTGGGFILEEE	-
R. norvegicus	KVIDTKGGFILEEE	G155E
C. lanigera	KIIDTEGGFILEEE	G155K
X. laevis	KVIDSGGGFFIEEE	-

3.2 Computational method

The following protocol was used to set-up and run the MD simulations of both, the XPA_{67-80} peptides unbound in solution and in complex with ERCC1.

As starting structure for the simulation of the *H. sapiens* ERCC1/XPA_{67-80} complex we used the first conformation form the NMR ensemble (PDBid 2JNW)[23]. The starting structures for the simulations of the complexes between ERCC1 and the XPA_{67-80} from *X. laevis, C. lanigera* and *R. norvegicus* were obtained with by structural alignment of the highly populated β-hairpin motifs identified from the MD simulations of the free peptide the *H. sapiens* ERCC1/XPA67-80 structure from NMR. We carried out all the simulations of the wild type peptides using GROMACS v 4.6.3[30]. We used the AMBER99SB-ILDN[31] force field to represent the protein atoms and counterions and TIP4P-Ew[32] for the water. The complex was placed in a rhombic dodecahedral simulation box with minimum distance between the protein and the sides of the box of 1.2 nm. We solvated the system and neutralised with either Na+ or Cl- ions. We carried out an energy minimisation through 500,000 steps of steepest descent, with a force-based convergence threshold of 100 kJ $mol^{-1}nm^{-1}$. Long range electrostatics were represented through periodic boundary conditions within the Particle Mesh Ewald (PME) framework with a switch of from real space to reciprocal space at 1.2 nm. Van der Waals interactions were calculated using a cut-off method, with a cut-off of 1.2 nm. All hydrogen bonds were constrained using the LINCS algorithm. After minimization, we carried out an equilibration of 500 ps in the NVT ensemble restraining the position of all solute heavy-atoms atoms. To integrate the equation of motion we used a leap-frog stochastic dynamics (sd) integrator, with a friction coefficient corresponding to the inverse of tau-t equal to 0.1 ps, where tau-t is the time constant for coupling. The *sd* integrator was set to maintain a target temperature of 300 K. We performed a second restrained equilibration of 500 ps in the NPT ensemble, with a Berendsen barostat set to a target pressure of 1 bar. For the ERCC1/XPA$_{67-80}$ complex, the production stage involves three consecutive equilibration steps of 5 ns each, first with the heavy atoms of the receptor and the ligand backbone atoms restrained, then restraining only the backbone atoms of ligand and receptor, and finally with the ligand atoms free and the receptor backbone atoms restrained. All ERCC1/XPA$_{67-80}$ complexes were simulated for 2 μs of conventional MD. In case of the XPA$_{67-80}$ peptides free in solution only one 5 ns of unrestrained MD equilibration was considered as sufficient, followed by a production step of 100 ns, which was used to isolate ten uncorrelated snapshots (one every 10 ns) that were used as starting point of ten independent MD simulations that were run in parallel, from which we collected data. Each of these ten simulations was run for 1 μs. We carried out a clustering analysis of the backbone atoms of each of the ten XPA peptide free in solution simulations with the first NMR structure in 2JNW used as a reference using the *gromos* method[33] and a RMSD cut-off of 0.15 nm. This value considered as optimal after testing cut-offs in a range between 0.05 nm and 0.2 nm, as it allowed to obtain the highest number of clusters while avoiding redundancy. In the set-up of the simulations of the XPA peptide from other species bound to ERCC1, we also mutated the ERCC1 so that the sequence of both ERCC1 and XPA match that of the corresponding animal. While there are more mutations than just those shown in **Table 3.1**, as shown in **Figure 3.3**, the mutations are either between residues with similar properties or the site of mutation is sufficiently far from the binding site so as to not affect it, as such we are only interested in the mutations listed in **Table 3.1**



```
X.laevis          -----VPKAGSYADYIFQKEAWDPVQKGQFSNMAASENAVSIVKQTTNQTTKSAGAGSCI
R.norvegicus      TAASTHSAPLTYAEYAIAQPPGGAG-----A--TVPTGSEPATGDSPSQTLKAGTKSSSI
H.sapiens         ---------------------------------------------------------NSI
C.lanigera        SETSAQGAPQTYAEYAISRPAGGAV-----V--TCPTGPEPLAGETPHPALKPGAKSNSI
                                                                           ..*

X.laevis          LVSTRQRGNSLLKYLRNVPWEFSDIVPDYILGETCCSLFLSLRYHNLNPEYIHSRLRSLG
R.norvegicus      IVSPRQRGNPVLKFVRSVPWEFGEVTPDYVLGQSTCALFLSLRYHNLHPDYIHERLQSLG
H.sapiens         IVSPRQRGNPVLKFVRNVPWEFGDVIPDYVLGQSTCALFLSLRYHNLHPDYIHGRLQSLG
C.lanigera        IVSPRQRGNPVLKFVRNVPWEFGEVVPDYVLGQSTCALFLSLRYHNLHPDYIHKRLQSLG
                  :** ***** :**::*.*****.:: ***:**:: *:**********:*:*** **:***

X.laevis          QSFALRVLLVQVDVKDPHFSLKELAKICILSDCTLILSWSPEEAARYLETYKCYEQKPAD
R.norvegicus      KNFALRVLLVQVDVKDPQQALKELAKICILADCTLVLAWSAEEAGRYLETYKAYEQKPAD
H.sapiens         KNFALRVLLVQVDVKDPQQALKELAKMCILADCTLILAWSPEEAGRYLETYKA-------
C.lanigera        KNFALRILLIQVDVKDPQQALKELAKMCILADCTLILAWSPEEAGRYLETYKAYEQKPAD
                  :.****:**:*******: :******:***:****:*:** ***.*******.
```

Figure 3.3: Sequence alignment of H. sapiens ERCC$_{99-214}$ to corresponding sequences from X. laevis, R. norvegicus and C. lanigera. Alignment obtained with Clustal-Omega[34]

We built the macrocycles based on the XPA$_{67-80}$ structure from the first entry of 2JNW using the molecular builder MAESTRO[35]. All macrocycles MD simulations were ran using AMBER v.12[36]. The AMBER99SB-ILDN force field was used to represent protein atoms and counterions, while we atoms of the linkers were represented by the Generalized Amber Force Filed (GAFF)[37]. The topology files were prepared using *tleap* tool, which is part of the AMBER v.12 distribution. We solvated the peptide with TIP4P-Ew[32] water model and placed it in an octahedral periodic box, as tleap doesn't support a dodecahedral box, with minimum distance between the peptide and the box sides of 12 Å. The system was minimised with 500k steps of steepest descent with protein heavy atoms restrained with a restrain weight of 5.0 kcal mol^{-1} Å^{-2}. Hydrogen bond lengths were restrained using the SHAKE algorithm[38]. According to the AMBER standard protocol, the temperature of the system is raised to 300 K gradually over two 500 ps NVT equilibration steps, first heating from 0 K to 100 K and then from 100 K to 300 K. An NPT equilibration followed to reach the equilibrium pressure of 1 bar. The equilibrated systems were then simulated for 1 μs in the NPT ensemble with all atoms unconstrained. For the simulation of the macrocycles in complex with ERCC1, the starting structures were obtained by structural alignment of a snapshot from the trajectories of the macrocycles unbound in solution onto the first structure from the NMR ensemble (PDBid 2JNW). For the equilibration of the system we followed the same protocol as the one used in the simulations of the unbound macrocycles, with four additional equilibration steps of 5 ns each, 1) with only the macrocycle sidechains unrestrained, 2) with all sidechains (macrocycle and protein) unrestrained, 3) with all atoms of the macrocycle unrestrained, and 4) and finally with all atoms unrestrained. This was followed by production runs of 1 μs for all four complexes.

MM/GB(PB)SA Calculations. The calculations were run by using the *MMPBSA.py* script[39] as part of the AMBER v.12 distribution. For the wild type peptides, we converted pdb files to AMBER topology and structure files, namely prm7 and rst7 files, respectively, using the *tleap* tool. The trajectories were converted from GROMACS format to AMBER format using VMD (http://www.ks.uiuc.edu/Research/vmd/)[40].

Conformational Entropy. We carried out Principal Component Analysis (PCA) calculations in GROMACS v 4.6.3 to obtain the conformational entropy contribution to the binding free energy not included in the MMGB(PB)SA free energy estimate. All calculations were done based on the peptide's backbone atoms. The total entropy was estimated based on the quasiharmonic formula and Schlitter's method[41] at 300K. We calculated the entropy penalty upon binding as the difference in entropy between the bound and free states.

3.3 Results

XPA$_{67\text{-}80}$ unlinked peptides. The identification of MoRFs within the conformational ensembles produced by the MD simulations and the corresponding stability in terms of relative populations were estimated by clustering analysis. The results obtained from the simulations of the *R. norvegicus*, *C. lanigera* and *X. laevis* XPA$_{67\text{-}80}$ peptides are shown in **Table 3.2**, **Table 3.3** and

Table 3.4, respectively. To determine if a cluster is a MoRF, the representative structure of the cluster was superimposed into the binding pocket of ERCC1 and was visually inspected for clashes with ERCC1. If the cluster has a hairpin motif and does not clash with ERCC1 it is considered a MoRF. The clusters classified as "maybe" would have a backbone conformation matching the binding conformation requirement, but with the Phe 75 sidechain in an unsuitable orientation, which would cause a clash with ERCC1. As the reorientation of the Phe 75 sidechain into its binding conformation would feasible upon recognition this can be seen in **Figure 3.4** panels c, f and i.

Table 3.2: Clustering analysis results for the three highest populated clusters from the 10 MD simulations (S1 to S10 in the first column) of the R. norvegicus XPA_{67-80} peptide. Binding conformations (MoRFs) are highlighted in green. The number of clusters generated through the clustering method is shown in the second column. The % population of each of the three most populated clusters is also shown. Middle Å indicates the RMSD of the representative structure of a cluster compared to the first structure of XPA in 2JNW.

	Cluster	Pop #1	middle (Å)	Binding	Pop #2	middle (Å)	Binding	Pop #3	middle (Å)	Binding
S1	136	50.1	4.140	Maybe	13.7	3.706	N	4.1	4.014	N
S2	357	10.3	2.971	Y	2.8	3.551	Y	2.8	3.036	N
S3	330	14.8	3.964	Maybe	4.7	3.428	N	3.3	3.505	N
S4	330	10.0	2.811	N	2.5	2.759	Y	2.2	4.797	N
S5	123	67.2	3.608	Maybe	3.3	3.082	N	2.9	3.498	Y
S6	45	63.4	3.992	Maybe	12.4	3.082	N	5.6	3.039	Y
S7	134	36.3	3.623	N	9.4	2.870	N	6.2	4.237	N
S8	298	9.9	4.991	N	5.8	3.002	N	3.9	3.489	Y
S9	244	25.6	5.806	N	10.3	4.659	N	3.7	5.267	N
S10	322	5.8	4.556	Y	5.0	1.940	N	2.9	4.025	N

Table 3.3 Clustering analysis results for the three highest populated clusters from the 10 MD simulations (S1 to S10 in the first column) of the C. lanigera XPA_{67-80} peptide. Binding conformations (MoRFs) are highlighted in green. The number of clusters generated through the clustering method is shown in the cluster column. The % population of each of the three most populated clusters is shown. Middle Å indicates the RMSD of the representative structure of a cluster compared to the first structure of XPA in 2JNW.

	cluster	Pop #1	middle (Å)	Binding	Pop #2	middle (Å)	Binding	Pop #3	middle (Å)	Binding
S1	275	40.5	2.575	Maybe*	2.2	3.142	Maybe*	2.1	3.886	N
S2	189	46.1	4.096	Y	6.3	4.160	N	2.1	2.616	Maybe*
S3	203	13.5	3.384	Maybe*	7.2	4.127	N	5.4	4.396	N
S4	197	11.4	3.725	N	7.2	5.470	N	5.8	3.449	Y
S5	229	21.2	3.042	Maybe*	10.3	1.801	Maybe*	5.2	3.495	N
S6	242	13.6	4.312	N	10.2	2.060	Maybe*	9.6	3.962	N
S7	298	5.3	3.940	N	5.0	2.880	Y	4.1	4.189	N
S8	320	8.1	4.413	Y	7.5	4.378	N	3.1	2.208	N
S9	100	37.9	2.768	Y	28.6	4.203	N	3.4	4.058	Y
S10	127	24.0	3.618	N	17.5	4.116	N	14.1	4.360	N

Table 3.4: Clustering analysis results for the three highest populated clusters from the 10 MD simulations (S1 to S10 in the first column) of the X. laevis XPA$_{67\text{-}80}$ peptide. Binding conformations (MoRFs) are highlighted in green. The number of clusters generated through the clustering method is shown in the cluster column. The % population of each of the three most populated clusters is shown. Middle Å indicates the RMSD of the representative structure of a cluster compared to the first structure of XPA in 2JNW.

	Cluster	Pop #1	middle (Å)	Binding	Pop #2	middle (Å)	Binding	Pop #3	middle (Å)	Binding
S1	317	8.0	3.225	N	7.3	2.872	Y	6.6	3.172	Y
S2	137	34.5	4.308	Maybe*	10.4	4.856	N	6.3	3.885	N
S3	151	51.8	4.784	N	3.0	4.230	N	2.9	4.433	Y
S4	459	9.7	3.559	Maybe*	2.8	2.694	Maybe*	1.9	4.473	Maybe*
S5	83	64.0	3.500	Y	7.6	3.067	Maybe*	4.0	4.182	N
S6	156	59.0	3.193	Y	5.4	3.685	N	4.5	2.942	Y
S7	170	36.6	3.240	Y	8.8	4.150	N	4.5	5.330	N
S8	222	10.4	2.114	Maybe*	8.7	2.824	N	4.4	3.390	N
S9	238	33.3	4.161	Maybe*	16.0	2.875	Y	3.9	3.929	Maybe*
S10	112	39.0	3.950	Maybe*	34.0	3.760	Maybe*	3.0	2.830	Maybe*

Because the MD simulations were started from different conformers it is possible that within the 1 μs sampling some simulations would end up exploring the same conformational space and some will not. For there we cross analysed the clusters identified for each trajectory in terms of structural similarity by backbone RMSD. The results are shown in **Table 3.5**, **Table 3.6** and **Table 3.7** for the simulations of the *R. norvegicus*, *C. lanigera* and *X. laevis* XPA$_{67\text{-}80}$ peptides, respectively. Conformations with a difference in backbone RMSD below 1.5 Å, which was cut-off value used in the clustering analysis, were considered as part of the same cluster. The total populations of the XPA$_{67\text{-}80}$ peptides from different species corresponding to MoRFs are shown in **Table 3.8**.

Table 3.5: Similarity matrix of the highest populated cluster from each MD simulation (S1 to S10) of the *R.norvegicus* XPA$_{67\text{-}80}$ peptide with identical clusters highlighted in green.

Population	50.1	10.3	14.8	10.0	67.2	63.4	36.3	9.9	25.6	5.8
	s0	s1	s2	s3	s4	s5	s6	s7	s8	s9
s0	0.0									
s1	3.4	0.0								
s2	2.6	3.3	0.0							
s3	3.8	3.7	3.1	0.0						
s4	4.1	2.6	3.8	3.4	0.0					
s5	1.1	3.7	0.5	2.6	3.3	0.0				
s6	4.7	3.7	4.4	3.9	3.6	4.8	0.0			
s7	5.2	6.1	6.3	5.8	5.9	5.9	3.8	0.0		
s8	3.6	4.3	3.3	4.3	4.7	3.9	4.8	5.0	0.0	
s9	3.0	4.1	3.1	4.5	4.5	3.4	4.8	5.7	3.7	0.0

Table 3.6: Similarity matrix of the highest populated cluster from each MD simulation (S1 to S10) of the *C. lanigera* XPA$_{67-80}$ peptide with identical clusters highlighted in green.

Population	40.5	46.1	13.5	11.4	21.2	13.6	5.3	8.1	37.9	24.0
	s1	s2	s3	s4	s5	s6	s7	s8	s9	s10
s0	0.0									
s1	3.8	0.0								
s2	4.7	4.2	0.0							
s3	5.5	3.8	3.0	0.0						
s4	2.0	3.4	4.7	4.9	0.0					
s5	4.0	2.3	4.0	3.9	3.2	0.0				
s6	6.5	5.5	5.1	5.4	6.2	6.3	0.0			
s7	4.9	3.4	2.9	2.6	4.3	3.7	5.2	0.0		
s8	3.5	4.1	4.6	4.2	3.1	4.4	5.5	4.8	0.0	
s9	2.8	3.5	4.0	4.5	2.9	4.2	5.3	4.2	3.7	0.0

Table 3.7: Similarity matrix of the highest populated cluster from each MD simulation (S1 to S10) of the *X. laevis* XPA$_{67-80}$ peptide with identical clusters highlighted in green.

Population	8.0	34.5	51.8	9.7	64.0	59.0	36.6	10.4	33.3	39.0
	s1	s2	s3	s4	s5	s6	s7	s8	s9	s10
s0	0.0									
s1	4.2	0.0								
s2	3.7	4.1	0.0							
s3	3.4	3.5	4.2	0.0						
s4	3.5	2.7	4.1	1.6	0.0					
s5	2.6	2.9	3.9	2.4	2.1	0.0				
s6	3.2	2.3	4.4	3.0	2.5	1.5	0.0			
s7	3.4	4.9	3.9	4.6	4.4	4.4	4.6	0.0		
s8	3.5	2.2	3.4	3.2	2.6	2.8	2.6	3.8	0.0	
s9	3.5	2.0	4.0	3.2	2.5	2.2	2.0	5.0	2.5	0.0

Table 3.8: Total populations (%) of conformations corresponding to MoRFs identified from the MD simulations of the XPA$_{67-80}$ peptides. The corresponding populations calculated considering the conformation for the Phe 75 sidechain unable to reorient upon binding, thus excluding the conformations flagged as *MAYBE, are shown in parentheses.

	H. sapiens	***R. norvegicus***	***X. laevis***	***C. lanigera***
MoRFs	36.4	24.0 (3.4)	37.7 (19.7)	16.4 (8.7)

Structural alignment of the representative (middle) conformers of the highest populated MoRF-like cluster, non-MoRF-like cluster and potential MoRF-like cluster

(*MAYBE) for each wild type XPA$_{67-80}$ peptide sequence to the first conformer from the NMR ensemble (PDBid 2JNW) from the *H. sapiens* XPA$_{67-80}$ are shown in **Figure 3.4**. The positions of Phe 75 sidechains relative to the bound conformation of the *H. sapiens* XPA$_{67-80}$ are highlighted.

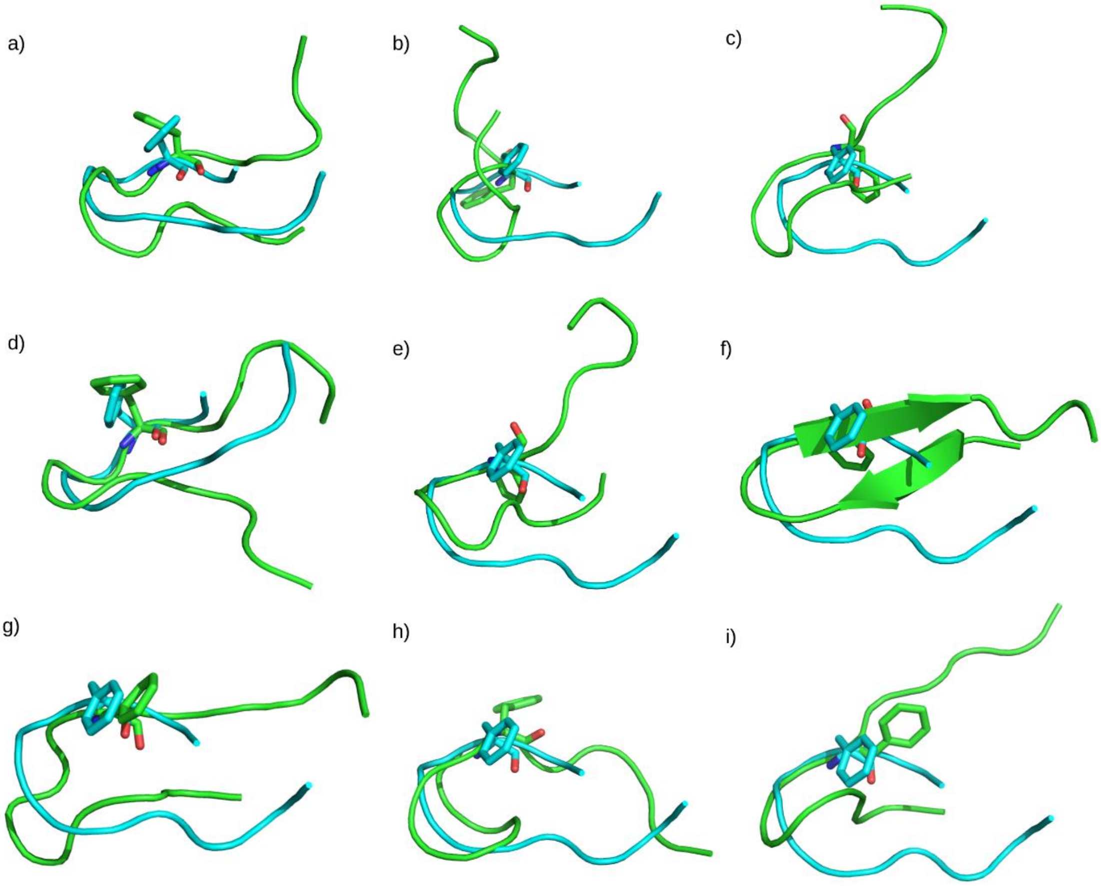


Figure 3.4: Structural alignments of highest populated a) MoRF-like cluster b) non-MoRF-like cluster and c) potential MoRF-like cluster from the simulations of the *R. norvegicus* XPA$_{67-80}$ peptide are shown in green to the H. sapiens XPA$_{67-80}$ peptide from PDBid 2JNW in cyan. Alignments of highest populated d) MoRF-like cluster e) non-MoRF-like cluster and f) potential MoRF-like cluster from the simulations of the *C. lanigera* XPA$_{67-80}$ peptide are shown in green. Alignments of highest populated g) MoRF-like cluster h) non-MoRF-like cluster i) potential MoRF-like cluster from the simulations of the *X. laevis* XPA$_{67-80}$ peptide are shown in green.

We evaluated the binding of the XPA$_{67-80}$ peptide variants to ERCC1 by looking at specific contacts between residues that in the different species we studies are mutated relative to *H. sapiens*, namely a salt bridge between Glu 155 of ERCC1 and Lys 72 of XPA$_{67-80}$ for *R. norvegicus* a salt bridge between Lys 155 and/or Arg 106 of ERCC1 and Glu 72 of XPA$_{67-80}$ for *C. lanigera*, and a cation–π interaction between Phe 76 and Arg 144 for X. laevis. The results in terms of distances between the heavy atoms involved in these interactions are shown in **Table 3.9**. The complexes binding affinities values calculated by MM-GBSA with entropies from PCA are shown in **Table 3.10**.

Table 3.9: Average distances between the heavy atoms involved in interactions between the mutated residues in the XPA$_{67-80}$/ERCC1 variants. Error shown in parentheses and the population below a cut-off of 5 Å is shown in brackets. Error is calculated as the standard deviation of the average value

	R. norvegicus (Å)	*C. lanigera* (Å)
K72-E155	9.42 (2.65) [10]	-
E72-K155	-	12.70 (2.86) [0]
	C. lanigera (Å)	*X. laevis* (Å)
E72-R106	5.48 (1.49) [46]	-
F75-R144	-	4.21 (0.53)

Table 3.10: Complexes binding free energies are shown in kcal mol^{-1}. Conformational entropies at 300 K evaluated by the Schlitter's method are shown. Total ΔG is calculated as the sum of ΔG_{MMGBSA} and -TΔS from PCA. Errors are indicated in parentheses. Error is calculated as the standard deviation of the average value

	H. sapiens	***R. norvegicus***	***X. laevis***	***C. lanigera***
ΔG_{MMGBSA}	-43.39 (3.59)	-43.27 (4.54)	-41.87 (4.52)	-49.14 (3.47)
-TΔS	14.33	20.13	17.90	30.29
Total ΔG	-29.06	-23.14	-23.97	-18.85

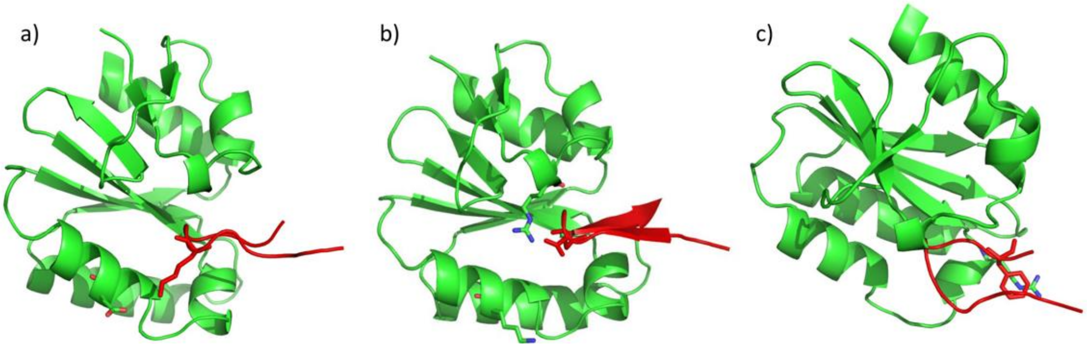


Figure 3.5: Interactions between ERCC1 (shown in green) and XPA$_{67-80}$ peptides (shown in red) in a) *R. norvegic*us, b) *C. lanigera* and c) *X. laevis*

Macrocyclic peptides. We looked at rigidifying the motif through formation of a macrocycle. The design of which based on modifying naturally occurring amino acids and using click chemistry to simplify the synthetic route and minimise the difference between the macrocyclic peptide and wild-type peptide. The sequences of the macrocyclic peptides are shown in

Table 3.11 and modified amino acids can be seen in **Figure 3.6**. MMGBSA data for the bound peptides and PCA results are shown in **Table 3.12**.

Table 3.11: Sequences of the macrocyclic peptides, * indicates the position of residues which we modified to form the macrocycle. In the case of MC004 the first and last residues have a peptide bond between their backbone atoms to form a fully cyclic peptide.

sequence	XPA$_{67-80}$ sequence
MC001	KIID*GGG*ILEEE
MC002	KIIDT*GG*ILEEE
MC003	KIID*GGG*ILEEE
MC004	KIIDTGGGFILEEEPAAP

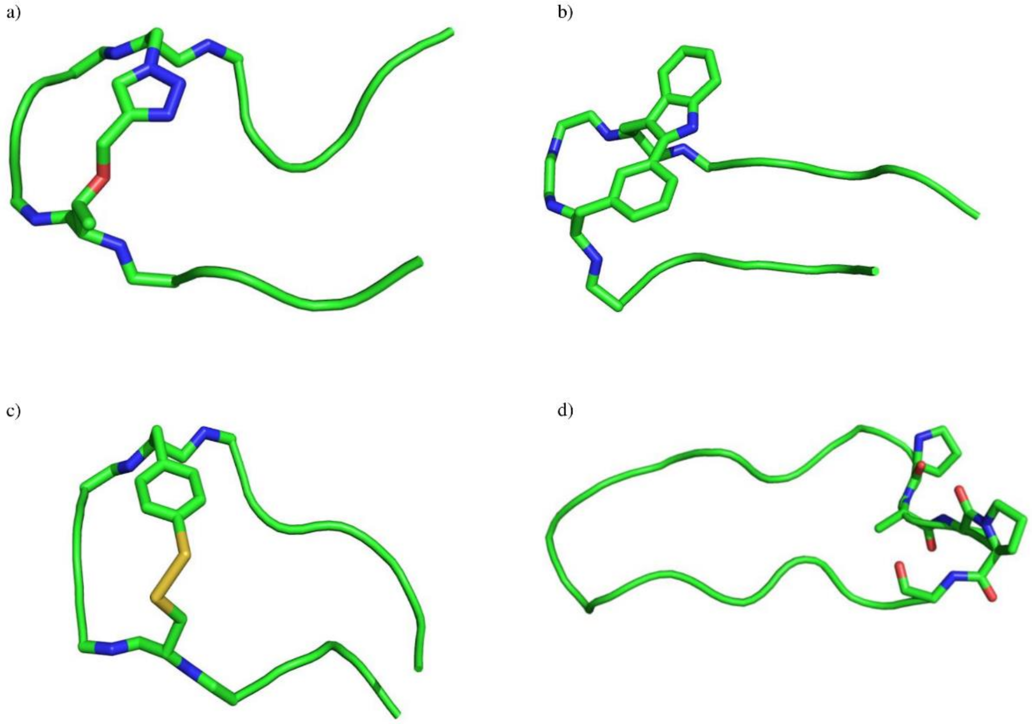


Figure 3.6: Macrocyclic peptides a) MC001 b) MC002 c) MC003 and d) MC004 with the modified residues shown as sticks and naturally occurring residues shown as a cartoon.

Table 3.12: Complexes binding free energies are shown in kcal mol^{-1}. Conformational entropies at 300 K evaluated by the Schlitter's method are shown. Total ΔG is calculated as the sum of ΔG_{MMGBSA} and -TΔS from PCA. Errors are indicated in parentheses. Error is calculated as the standard deviation of the average value

	MC001	**MC002**	**MC003**	**MC004**
ΔG_{MMGBSA}	-32.75 (4.19)	-42.33 (4.33)	-47.47 (4.99)	-38.51 (6.26)
-TΔS	6.77	12.43	8.56	10.93
Total ΔG	-25.98	-29.90	-38.91	-27.58

3.4 Discussion

The XPA$_{67\text{-}80}$ peptide homologues we have chosen to analyse have subtle but interesting mutations in their sequence relative to *H.sapiens*, see **Table 3.1**. The G72K and G72E mutations are particularly striking as they appear at the head of the β-hairpin MoRF conformation, which is occupies a sterically constrained region of the ERCC1 binding pocket. Our simulations show that in fact these mutations that involve residues with large sidechains from a glycine, with no sidechains can be well accommodated in the ERCC1 binding site. Additionally, in both *R. norvegicus* and *C. lanigera*, the XPA$_{67\text{-}80}$ mutations are accompanied by mutations in the ERCC1 sequences, namely G155E and G155K, respectively. In *R. norvegicus* G155E forms a weak salt bridge with the G72K of XPA$_{67\text{-}80}$ stable for 10% of the simulation time, see **Table 3.9**. Meanwhile, the inverse of these mutations in *C. lanigera*, namely the ERCC1 G155K and the XPA$_{67\text{-}80}$ G72E, are not seen to interact through a salt bridge, as the distance between heavy atoms is always significantly above a cut-off of 5 Å throughout the simulation. However, our results show that the Glu 72 of the *C. lanigera* XPA$_{67\text{-}80}$ forms a stable salt bridge with the ERCC1 at Arg 106, a residue that is also present in the human ERCC1. The contribution of this new interaction is reflected by the binding affinity values calculated by MMGBSA, see **Table 3.12**, for the peptide from *C. lanigera* relative to the human counterpart. Therefore, based on this result a G72E mutation could be introduced to improve the binding affinity of the peptide. Another interesting mutation is found in *X. laevis* XPA$_{67\text{-}80}$, namely I76F shown in **Table 3.1**. The additional Phe in position 76 forms a cation-π interaction with the ERCC1 Arg 144 that is stable throughout the simulation. Arg 144 is also present in the human ERCC1. However, this mutation does not affect the binding free energy of the peptide relative to *H. sapiens*, this may be due to Arg 144 being on the surface of ERCC1 and being exposed to the solvent. It is important to note that all these mutations contribute to enhance the conformational entropy of the peptide in solution, thus the penalty upon binding evaluated through PCA, in particular the ones where the Gly 72 is replaced by a charged residue, either Lys in *R. norvegicus* or Glu in *C. lanigera*. Indeed, as shown in **Table 3.9**, the introduction of these mutations at the head of the β-hairpin destabilizes the MoRF. Although all the mutations do not seem to enhance the linear XPA$_{67\text{-}80}$ peptide's binding affinity, mostly because of their effect on the entropy, we think that it could be useful to consider these mutations, and particularly the G72E mutation, within the context of the design of conformationally restrained macrocycles, as any conformational entropy penalty will be limited by the cyclization.

As shown in **Table 3.12**, all the macrocycles we have designed have a lower entropic penalty upon binding relative to the linear XPA$_{67-80}$, proving the introduction of chemical linkers to restrain the peptide’s dynamics is a promising strategy for the generation of high affinity binders. However, the macrocycles studied here do not show an improvement in the binding enthalpy and in fact, as shown in **Tables 3.12**, all have a slightly lower ΔG_{MMGBSA} contribution relative to the linear peptides. Nevertheless, rounds of improvement in the macrocyclic peptide sequence, potentially guided by the results obtained for the linear peptides from the XPA homologues, could very well lead to high-affinity macrocyclic XPA$_{67-80}$ peptide candidates.

3.5 Conclusion

From this work we see that nature can be a useful resource of inspiration for the identification of viable mutations of short IDR. This strategy is viable as we know that those mutations lead to functional proteins in the respective species they are in and thus must be functional. Because of this, it is important that when looking at these alternative sequences from homologues, complementary mutations in the receptor must be also considered, as the same and/or additional interactions may not be possible in the human (or targeted) receptor. Our results show that interestingly all the mutations we have looked at can modify the dynamics of the peptide balancing out the binding enthalpy gain with a higher entropic penalty upon binding. Macrocyclisation is therefore a valuable tool to rigidify a highly flexible peptide and to reduce its conformational entropy. We think that combining these two approaches, i.e. introduction of point mutations that enhance the enthalpic contribution from other species and chemically restraining the dynamics of the peptide through cyclization, represent an interesting and viable strategy to develop higher affinity molecules from IDR scaffolds.

References

1. Shell, S. M. & Chazin, W. J. XPF-ERCC1: on the bubble. *Structure* **20**, 566–568 (2012).
2. Schärer, O. D. Nucleotide Excision Repair in Eukaryotes. *Cold Spring Harb. Perspect. Biol.* **5**, (2013).
3. Shuck, S. C., Short, E. A. & Turchi, J. J. Eukaryotic nucleotide excision repair: from understanding mechanisms to influencing biology. *Cell Res.* **18**, 64–72 (2008).
4. Compe, E. & Egly, J.-M. TFIIH: when transcription met DNA repair. *Nat. Rev. Mol. Cell Biol.* **13**, 343–354 (2012).
5. Volker, M. *et al.* Sequential Assembly of the Nucleotide Excision Repair Factors In Vivo. *Mol. Cell* **8**, 213–224 (2001).
6. Krasikova, Y. S., Rechkunova, N. I., Maltseva, E. A., Petruseva, I. O. & Lavrik, O. I. Localization of xeroderma pigmentosum group A protein and replication protein A on damaged DNA in nucleotide excision repair. *Nucleic Acids Res.* **38**, 8083–8094 (2010).
7. Bunick, C. G., Miller, M. R., Fuller, B. E., Fanning, E. & Chazin, W. J. Biochemical and Structural Domain Analysis of Xeroderma Pigmentosum Complementation Group C Protein. *Biochemistry* **45**, 14965–14979 (2006).
8. Li, L., Lu, X., Peterson, C. A. & Legerski, R. J. An interaction between the DNA repair factor XPA and replication protein A appears essential for nucleotide excision repair. *Mol. Cell. Biol.* **15**, 5396 LP – 5402 (1995).
9. Krasikova, Y. S. *et al.* Interaction of nucleotide excision repair factors XPC-HR23B, XPA, and RPA with damaged DNA. *Biochem.* **73**, 886–896 (2008).
10. Park, C.-H., Mu, D., Reardon, J. T. & Sancar, A. The General Transcription-Repair Factor TFIIH Is Recruited to the Excision Repair Complex by the XPA Protein Independent of the TFIIE Transcription Factor. *J. Biol. Chem.* **270**, 4896–4902 (1995).
11. You, J.-S., Wang, M. & Lee, S.-H. Biochemical analysis of the damage recognition process in nucleotide excision repair. *J. Biol. Chem.* **278**, 7476–7485 (2003).
12. Bergink, S. *et al.* Recognition of DNA damage by XPC coincides with disruption of the XPC–RAD23 complex. *J. Cell Biol.* **196**, 681 LP – 688 (2012).
13. Das, D. *et al.* The Structure of the XPF-ssDNA Complex Underscores the Distinct Roles of the XPF and ERCC1 Helix- Hairpin-Helix Domains in ss/ds DNA Recognition. *Structure* **20**, 667–675 (2012).
14. Sugitani, N., Shell, S. M., Soss, S. E. & Chazin, W. J. Redefining the DNA-Binding Domain of Human XPA. *J. Am. Chem. Soc.* **136**, 10830–10833 (2014).
15. Hermanson-Miller, I. L. & Turchi, J. J. Strand-specific binding of RPA and XPA to damaged duplex DNA. *Biochemistry* **41**, 2402–2408 (2002).
16. Schärer, O. D. The molecular basis for different disease states caused by mutations in TFIIH and XPG. *DNA Repair (Amst).* **7**, 339–344 (2008).
17. Lafrance-Vanasse, J. *et al.* Structural and functional characterization of interactions involving the Tfb1 subunit of TFIIH and the NER factor Rad2. *Nucleic Acids Res.* **40**, 5739–5750 (2012).
18. Araújo, S. J., Nigg, E. A. & Wood, R. D. Strong Functional Interactions of

TFIIH with XPC and XPG in Human DNA Nucleotide Excision Repair, without a Preassembled Repairosome. *Mol. Cell. Biol.* **21**, 2281 LP – 2291 (2001).

19. Fagbemi, A. F., Orelli, B. & Schärer, O. D. Regulation of endonuclease activity in human nucleotide excision repair. *DNA Repair (Amst).* **10**, 722–729 (2011).

20. Kemp, M. G., Reardon, J. T., Lindsey-Boltz, L. A. & Sancar, A. Mechanism of release and fate of excised oligonucleotides during nucleotide excision repair. *J. Biol. Chem.* **287**, 22889–22899 (2012).

21. Fadda, E. Role of the XPA protein in the NER pathway: A perspective on the function of structural disorder in macromolecular assembly. *Comput. Struct. Biotechnol. J.* **14**, (2015).

22. Buchko, G. W., Ni, S., Thrall, B. D. & Kennedy, M. A. Structural features of the minimal DNA binding domain (M98-F219) of human nucleotide excision repair protein XPA. *Nucleic Acids Res.* **26**, 2779–2788 (1998).

23. Tsodikov, O. V *et al.* Structural basis for the recruitment of ERCC1-XPF to nucleotide excision repair complexes by XPA. *EMBO J.* **26**, 4768–4776 (2007).

24. Siddik, Z. H. Cisplatin: mode of cytotoxic action and molecular basis of resistance. *Oncogene* **22**, 7265–7279 (2003).

25. Du, P., Wang, Y., Chen, L., Gan, Y. & Wu, Q. High ERCC1 expression is associated with platinum-resistance, but not survival in patients with epithelial ovarian cancer. *Oncol. Lett.* **12**, 857–862 (2016).

26. Fadda, E. Conformational Determinants for the Recruitment of ERCC1 by XPA in the Nucleotide Excision Repair (NER) Pathway: Structure and Dynamics of the XPA Binding Motif. *Biophys. J.* **104**, 2503–2511 (2013).

27. Fadda, E. The role of conformational selection in the molecular recognition of the wild type and mutants XPA 67-80 peptides by ERCC1: Molecular Recognition of XPA 68-80 Peptide Mutants. *Proteins* **83**, (2015).

28. Tompa, P., Szasz, C. & Buday, L. Structural disorder throws new light on moonlighting. *Trends Biochem. Sci.* **30**, 484–489 (2005).

29. Oldfield, C. J. *et al.* Flexible nets: disorder and induced fit in the associations of p53 and 14-3-3 with their partners. *BMC Genomics* **9 Suppl 1**, S1 (2008).

30. Abraham, M. J. *et al.* GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1–2**, 19–25 (2015).

31. Lindorff-Larsen, K. *et al.* Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* **78**, 1950–1958 (2010).

32. Horn, H. W. *et al.* Development of an improved four-site water model for biomolecular simulations: TIP4P-Ew. *J. Chem. Phys.* **120**, 9665–9678 (2004).

33. Daura, X. *et al.* Peptide Folding: When Simulation Meets Experiment. *Angew. Chemie Int. Ed.* **38**, 236–240 (1999).

34. Madeira, F. *et al.* The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res.* **47**, W636—W641 (2019).

35. Maestro, version 9.2. Schrödinger, LLC; New York, NY, U. 2011. .

36. D.A. Case, T.A. Darden, T.E. Cheatham, III, C.L. Simmerling, J. Wang, R.E. Duke, R.Luo, R.C. Walker, W. Zhang, K.M. Merz, B. Roberts, S. Hayik, A. Roitberg, G. Seabra, J. Swails, A.W. Götz, I. Kolossváry, K.F. Wong, F. Paesani, J. Vanicek, R.M. Wolf, J. L, and P. A. K. AMBER 12. (2012).

37. Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. Development and testing of a general amber force field. *J. Comput. Chem.* **25**, 1157–1174 (2004).

38. Ryckaert, J.-P., Ciccotti, G. & Berendsen, H. J. C. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* **23**, 327–341 (1977).

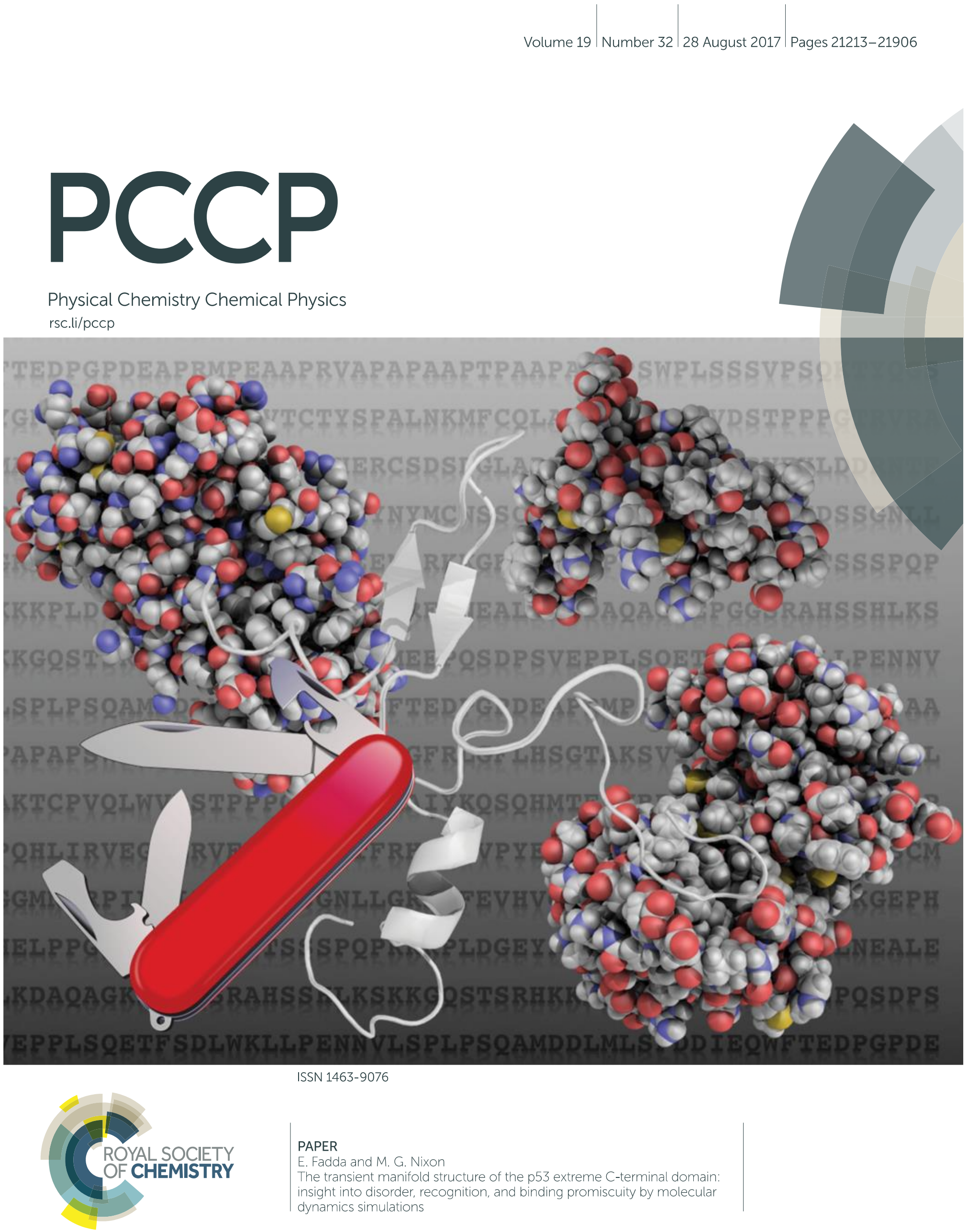
39. Miller, B. R. *et al.* MMPBSA.py: An Efficient Program for End-State Free Energy Calculations. *J. Chem. Theory Comput.* **8**, 3314–3321 (2012).

40. Humphrey, W., Dalke, A. & Schulten, K. {VMD} -- {V}isual {M}olecular {D}ynamics. *J. Mol. Graph.* **14**, 33–38 (1996).

41. Schlitter, J. Estimation of absolute and relative entropies of macromolecules using the covariance matrix. *Chem. Phys. Lett.* **215**, 617–621 (1993).

Chapter 4: The transient manifold structure of the p53 extreme C-terminal domain: insight into disorder, recognition, and binding promiscuity by molecular dynamics simulations

Note: This Chapter is largely based on the paper, Fadda E. and Nixon MG, *Phys Chem Chem Phys* (2017), **19**(32): 21287-21296, which was also highlighted on the journal's front cover.



Volume 19 | Number 32 | 28 August 2017 | Pages 21213–21906

PCCP

Physical Chemistry Chemical Physics

rsc.li/pccp

ISSN 1463-9076

ROYAL SOCIETY OF CHEMISTRY

PAPER

E. Fadda and M. G. Nixon

The transient manifold structure of the p53 extreme C-terminal domain: insight into disorder, recognition, and binding promiscuity by molecular dynamics simulations

4.1 Introduction

The p53 extreme C terminal domain (CTD) is a 30 residue highly basic IDP of the p53 tumour suppressor. The p53-CTD has been shown to have a negative regulatory control of the p53 DNA-binding activity[1–3], with phosphorylation or deletion of the p53-CTD region resulting in a constitutively DNA-binding active p53 molecule[4]. The p53-CTD is also highly targeted for post-translational modifications, which modulate its DNA-binding activity[4,5]. The p53-CTD is poorly structured in solution[6,7], but it adopts a variety of stable secondary structures when bound to different receptors, ranging from α-helices, to coils[6,8,9]. Recognition and binding could follow these different mechanistic scenarios, (a) a non-specific “encounter complex” is initially formed between peptide and receptor, followed by an induced fit phase, where the progressive setting of specific interactions drives folding, or (b) the peptide unbound in solution can access its bound fold, which is selected and bound by the target receptor, *i.e.* by conformational selection, or (c) a mechanism in between these two scenarios, whereby conformational selection and induced fit both play a role[10,11]. In the specific case of the p53-CTD, it has been determined that the folded conformations are stable only when the peptide is bound[6,12,13]. Intricate balances regulating recognition and binding are not unusual for IDPs[6,12–19]. Because of the high degree of intrinsic disorder, structural investigations of the unbound p53-CTD in solution have not been particularly informative in terms of subtleties in its residual secondary structure[2].

Recent work strongly suggests that recognition and binding mechanisms of IDPs depends on their intrinsic secondary structure propensity, dictated by the peptide/protein sequence[17]. In this work we used extensive molecular dynamics (MD) simulations to analyse the conformational propensity of the unbound p53-CTD in solution and to investigate the structure and potential roles of transiently stable structural motifs in recognition and binding[20]. Indeed, these motifs could function as molecular recognition features (MoRFs)[21,22], by being selectively targeted by different receptors, thus working as nucleation sites for the completion of the folding by induced fit[20,23–25]. Also, an intrinsic propensity to form minimal structural motifs that can be specifically recognized by different receptors would explain how the p53-CTD conformational disorder supports its binding promiscuity[26].

The analysis of over 20 μs of cumulative MD simulation trajectories of a 22 residue peptide free in solution, with sequence corresponding to the p53-CTD $_{367}$SHLKSKKGQSTSRHKKLMFKTE$_{388}$ segment, suggest that the p53-CTD conformational disorder also includes in addition to random coils, structures containing specific, relatively stable, localized and reoccurring short secondary structure motifs, which encompass stretches of 3 to 4 residues in case of helical turns. The results also show that the MoRFs we identified occur with higher probability in the C-terminal half of the peptide, while the N-terminal half remains mostly disordered. These findings are in agreement with structure/disorder prediction tools, namely s2D[27] and PONDR-VL-XT[28] that show a different degree of structural propensity along the p53-CTD sequence and supported by very recent experimental results that in agreement with our simulation results suggest the presence of β-turn or helical structures[29]. The identification of such distinct motifs within the disordered ensemble suggests that the p53-CTD may exert its broad binding specificity through minimal structural MoRFs, which are specifically selected and bound by different receptors, fitting within a broader framework of the conformational selection theory[10,11]. We discuss a potential MoRF-based recognition and binding mechanism in the case of the p53-CTD peptide in complex with the Ca^{2+} bound S100B(ββ) dimer[6], and with sirtuin Sir2[9].

4.2 Computational method

A 22-residue peptide corresponding to residues 367–388 of the *H. sapiens* p53-CTD was built in a fully extended conformation with the molecular builder tool in Maestro v.9.7[30]. N- and C-termini were capped with ACE and NME residues, respectively. The fully extended peptide, measuring 8.4 nm, was centred in a truncated dodecahedral simulation box sized so that the minimum distance between the peptide and the box sides would not be lower than 1.2 nm. The total charge of +6 was neutralized with the addition of Cl^{-} counterions. Because the aim of this work is to determine the conformational propensity of the peptide in function of its sequence, the effect of ionic strength in physiological conditions has not been addressed. Protein atoms and counterions were represented with AMBER-99SB-ILDN parameters[31] while TIP4P-Ew[32] was chosen as water model. Convergence of our simulations has been verified by monitoring, (a) the average backbone RMSD values, calculated relative to 2 distinct and highly populated MoRFs we identified, namely an asymmetric β sheet hairpin (cluster 1, MD1) and a conformer containing a 3_{10} helical turn located at $_{376}$STS$_{378}$ (cluster 2, MD3), and (b) the corresponding backbone RMSD running averages, and (c) the RMSD average correlation values (RAC)[33], see **Figure 4.1**.

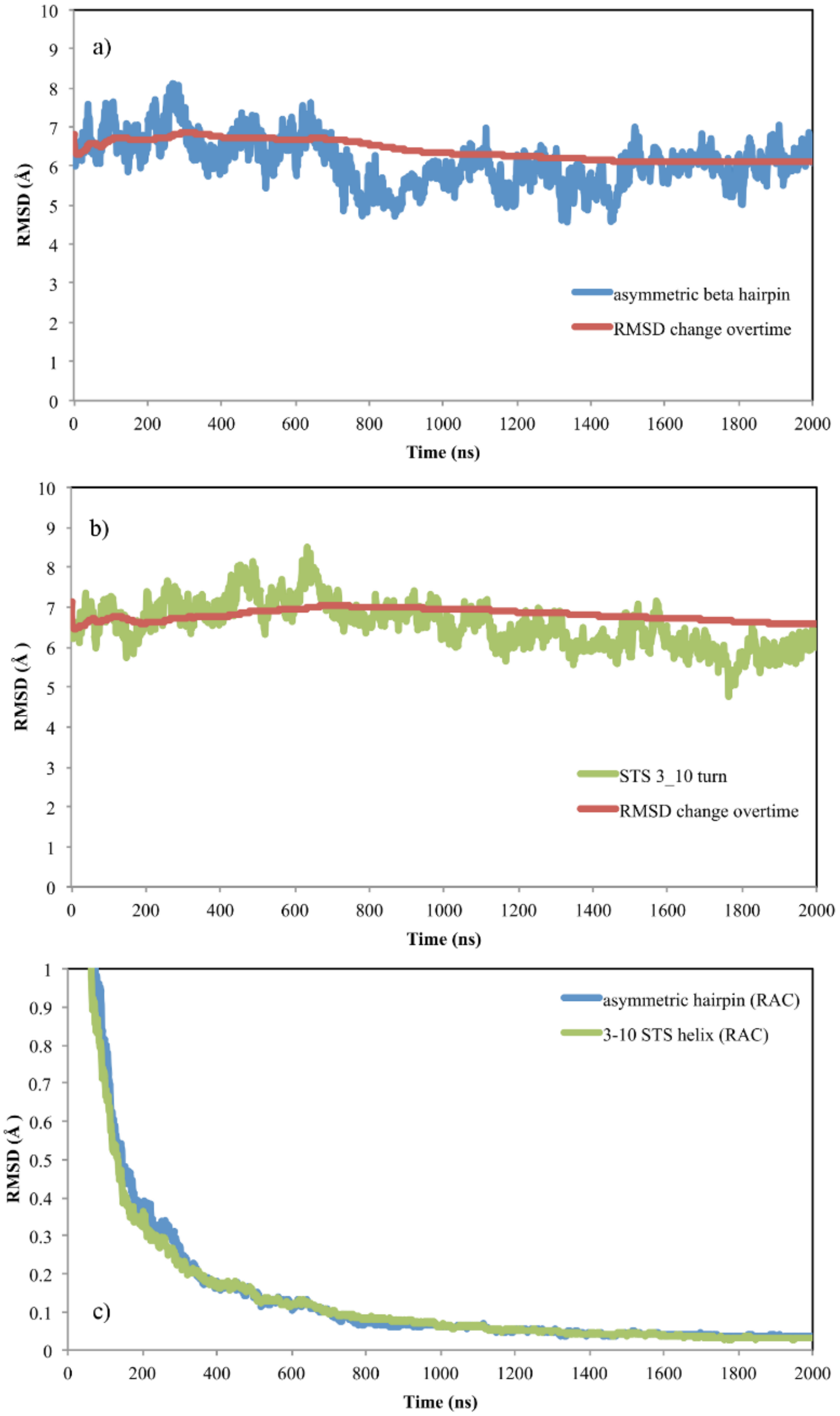


Figure 4.1: Backbone RMSD values for a) a highly populated asymmetric hairpin b) a highly populated 3_{10} helical turn and c) using the RMSD average correlation (RAC) values.

Bond lengths with hydrogen atoms were constrained with the LINCS algorithm[34]. Long range electrostatic interactions were treated with Particle Mesh Ewald (PME), with a switch from real space to reciprocal space at 1.2 nm. Van der Waals interactions were also cut off at 1.2 nm. All MD simulations were run with version 4.6.3 of GROMACS[35]. Two additional simulations, each of 2 μs production, were run starting with the p53-CTD peptide in the helical conformation seen when in complex with S100B(ββ) (PDBid 1dt7). The following protocol was used to set-up and run all simulations in this work. An initial energy minimization of 500k steps of steepest descent were used to prepare the system for the equilibration phase. During the minimization only the positions of the solvent molecules, counterions, and hydrogen atoms was left unconstrained. 500 ps of equilibration in the *NVT* ensemble and subsequently 500 ps in the *NPT* ensemble followed, with a target temperature of 300 K and pressure of 1 bar. Following this stage, a 1 ns equilibration was run with the backbone atoms restrained and the sidechain atoms free. All restraints were then released for 5 ns and a 100 ns production run was recorded for analysis. From this trajectory, snapshots were collected every 10 ns. The 10 uncorrelated peptide structures were removed from their original simulation box and placed in a truncated dodecahedral simulation box of 84 Å sides, sized to leave enough space to accommodate largely extended conformations that may occur during the simulation. The minimization and equilibration protocol described above was repeated for all 10 systems. Production of every trajectory was extended to 2 μs, for a total simulation time of 20 μs. The clustering analysis was performed with the *gromos* algorithm[36] with a cut-off of 0.45 nm. Such large cut-off value is necessary due to the highly dynamic nature of the N-terminal tail of the peptide, and it was found as the minimum cut-off value that allows us to separate and group significant secondary structure motifs. The secondary structure analysis was done on the 10 highest populated clusters obtained from each MD trajectory. These clusters are representative of a minimum of 63% of all conformations accessible, in case of high conformational dynamics, up to a complete coverage of 100%, in case of the formation of stable secondary structure motifs. MoRF populations for each MD trajectory were calculated based on clusters populations, with counts of 1 ns per frame. Populations over the cumulative 20 μs simulation were estimated as sums of the populations during the single trajectories, where the error bars correspond to the standard deviation. Secondary structures were assigned according to the STRIDE online tool[37]. Image rendering, structural alignments, and distance analysis was done with PyMOL v.1.6[38].

4.3 Results

The conformational dynamics of a 22-residue peptide, corresponding to the Ser 367 to Glu 388 section of the p53-CTD, was analysed with molecular dynamics (MD) for a cumulative time of 20 μs. The simulation is an aggregate of ten separate trajectories, started from uncorrelated snapshots, selected from a 100 ns trajectory started from a fully extended backbone conformation. In addition to these ten separate trajectories, we ran two separate, 2 μs long, trajectories, both started from the helical conformation of p53-CTD peptide when in complex with S100B(ββ)[6] (PDBid 1dt7). These calculations were run to assess the stability of the helix in solution and the residual helicity retained in function of the starting structure.

The p53-CTD peptide size was analysed in terms of its R_g. The average R_g calculated over the combined 20 μs trajectories is 0.97 ± 0.07 nm. This value is larger than the R_g value 0.73 nm, predicted for a 22 residue random-flight polymer with link distances of 0.38 nm, representative of a random coil behaviour, but smaller than 1.01 ± 0.04 nm, the R_g measured for the Ace-$(AAKAA)_4$-GY-NH_2 peptide, containing well-structured α helical motifs[39]. The R_g plots obtained from the 10, 2 μs long, trajectories are shown in **Figure 4.2**

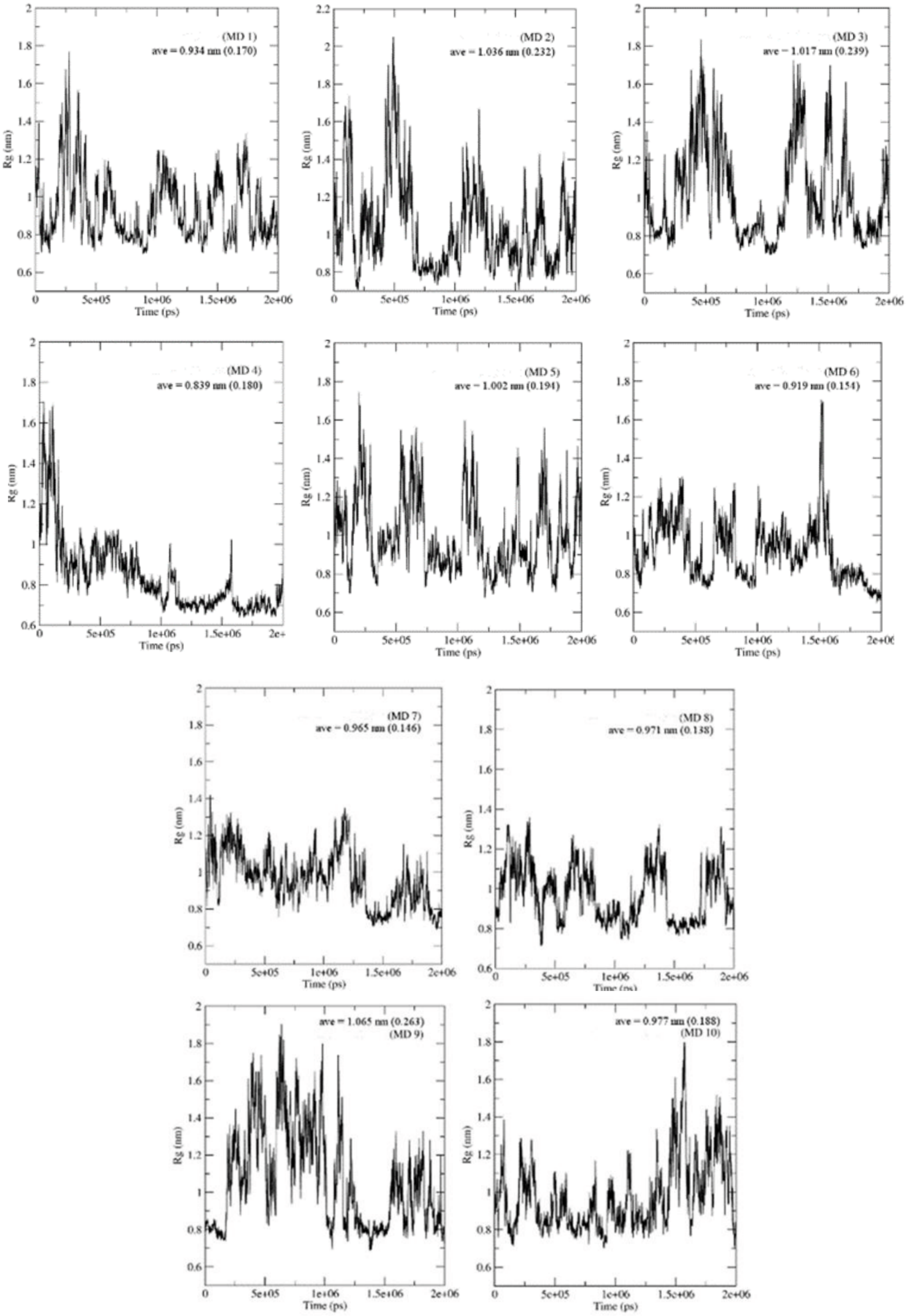


Figure 4.2: Radius of gyration (Rg) values calculated over the 10 (2 μs) MD simulations, namely MD 1 to 10, of the 22 residue p53-CTD peptide unbound in solution. Standard deviations are indicated in brackets.

The trends describe a highly dynamic structure with recurrent signatures of relatively stable conformers. These correspond to localized and short secondary structure motifs, which we have classified through clustering analysis and are described in the sections below. The relative populations of the secondary structure motifs identified over the 20 μs of cumulative simulation time are shown in **Figure 4.3** and in **Table 4.1**. The highest populated structures contain β-bridges (27.3%), and β-sheet hairpins (25.7%). We also identified helical turns (13%), both α-helical, and 3_{10}. Random coils and turns, which contribute the most to the disordered character of the peptide, have a relative population of 15.2%. Interestingly, similar secondary structure motifs form recurrently throughout the dynamics, and are found to involve preferentially the same group of residues within the C-terminal half of the peptide. The relative populations only accounts for the presence of the motif and does not account for the number of residues which comprises the motif, so the true level of secondary structure will be lower than indicated in **Figure 4.3** and in **Table 4.1**.

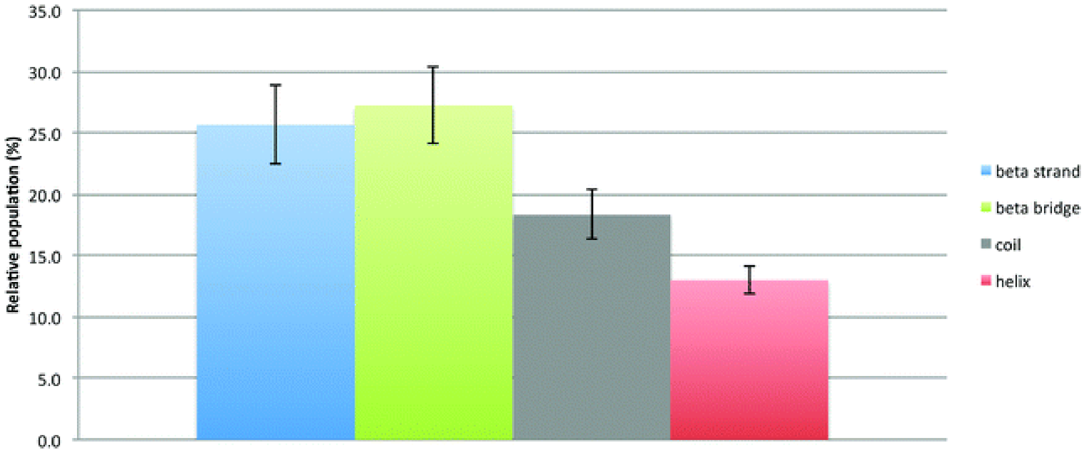


Figure 4.3: Relative populations of secondary structure motifs identified over a cumulative 20 μs MD simulation of the 22 residue p53-CTD peptide.

Table 4.1: Relative populations (%) of the secondary structure motifs identified during each trajectory. MD 1–10 are trajectories started from a common fully extended peptide, see Computational method section, while Helix 1–2 MD indicate trajectories started from the α helical conformation from the complex with S100B(ββ) (PDBid 1dt7). Populations are calculated over 2 μs and account for the 10 highest populated clusters. The total reflects the populations over 20 μs cumulative sampling, where the standard deviation is indicated in brackets

Trajectory	β-Strands	α/3_{10} helices	β-Bridges	Coil/turns
MD 1	55.3	7.7	8.9	13.4
MD 2	0.0	19.1	0.0	53.4
MD 3	0.0	27.8	24.4	17.3

MD 4	0.0	0.0	81.4	14.5
MD 5	3.8	19.9	9.4	14.4
MD 6	63.4	7.1	26.5	4.7
MD 7	17.7	21.1	83.1	0.0
MD 8	89.0	0.0	8.9	2.2
MD 9	0.0	2.9	0.0	57.6
MD 10	27.5	24.8	30.1	6.0
Total (over 20 µs)	**25.7(3.2)**	**13.0(1.1)**	**27.3(3.1)**	**18.3(2.0)**
Helix 1 MD	0.0	52.7	0.0	30.7
Helix 2 MD	0.0	11.7	41.4	29.0

β-Sheet hairpin motifs. As shown in the R_g plot in **Figure 4.2**, the first MD trajectory (MD1) visits a relatively stable conformation between 400 ns and 1.5 µs. In this interval the peptide is in an asymmetric β-sheet conformation, shown in **Figure 4.4**. The core of the β-sheet is held together by hydrogen bonds connecting Ser 376 and Phe 385, while the free N-terminal tail, spanning residues Ser 367 to Gly 374 ($_{367}$SHLKSKKG$_{374}$), is not tied

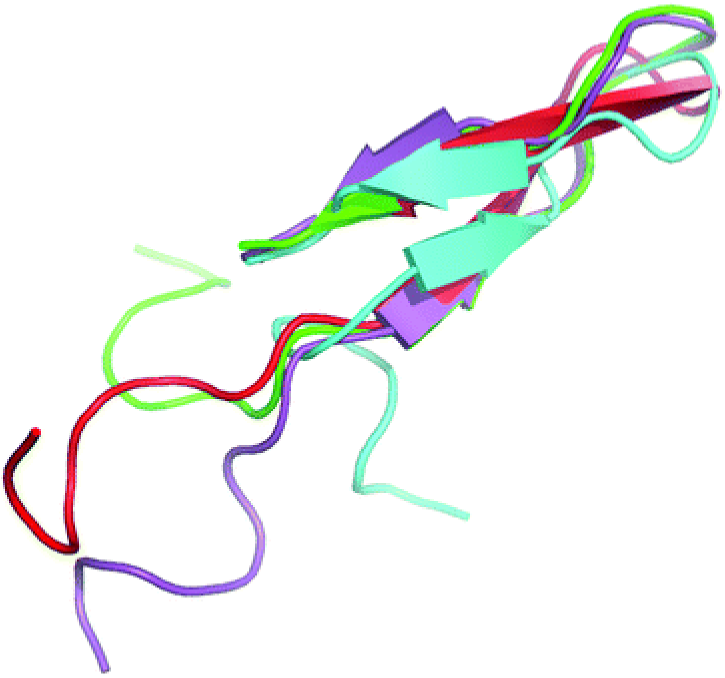


Figure 4.4: Examples of the β-sheet motif identified through the clustering analysis of the 20 µs MD simulation of the p53-CTD. The conformations visited during MD 1, MD 7, and MD 8 (cluster 1 and 3), are shown in cyan, red, green, and purple, respectively. The flexible tail corresponds to the stretch between Ser 367 and Gly 374.

into the hairpin and its dynamics determines the oscillations of the R_g value. The hairpin turn comprises His 380 and Lys 381. This conformation is the highest populated, stable for 55% of the simulation time, or more than half of the MD1 trajectory. The structural alignment of all the β-sheet motifs found during all the other simulations shows that this particular asymmetric hairpin conformation is the highest populated type of β-sheet, also present in MD6 with 63.4% population over 2 μs, MD7 with 17.7% population, MD8 with 89.0% population, and MD10 with population of 23.0%. Two other slightly different, and lower populated, β-sheet conformations have been identified, one in MD5, with a population of 3.8% over 2 μs, and the other in MD10, with a population of 4.4%. As shown in **Figure 4.5**, these two conformers show the same asymmetry,

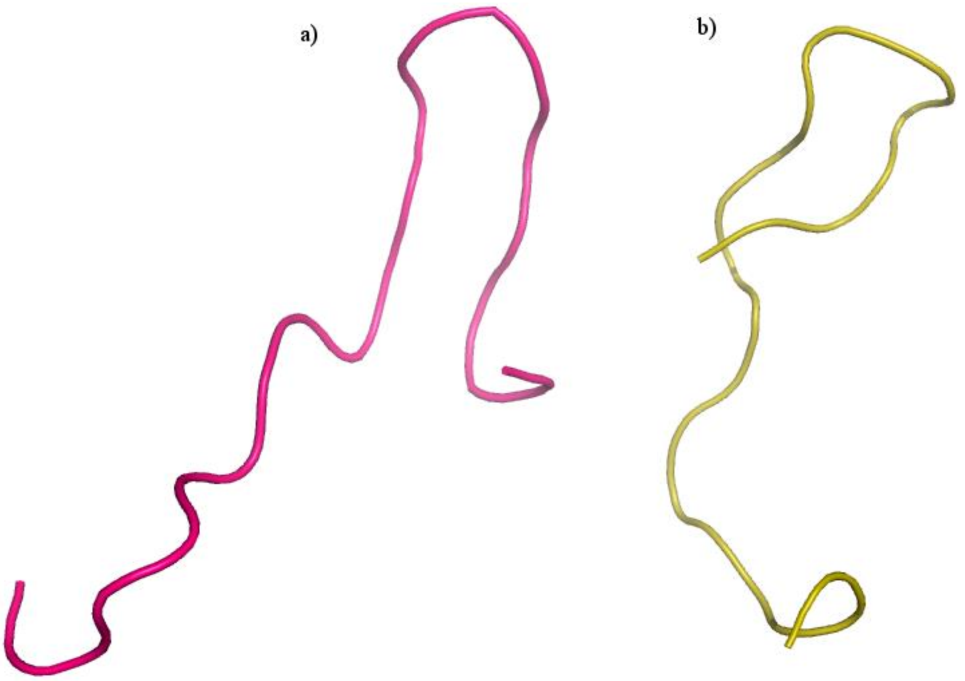


Figure 4.5: Low populated β-sheet conformers identified through the MD simulation that slightly differ from the highest populated and most stable asymmetric fold. In panel a) the conformer corresponding to MD5 cluster 9, in panel b) the conformer corresponding to MD10 cluster 7.

but one (MD10) has a wider hairpin section, formed by Lys 381, Lys 382, and Leu 383, and the other one (MD5) with a slightly different hydrogen bonding pattern relative to the highest populated motif, connecting not only Ser 376 to Phe 385, but also Ser 378 to Met 386. An RMSD matrix obtained through sequence alignment of the 15-residue stretch between Gly 375 and Glu 388, followed by structural alignment of all the β-sheet motifs, is shown in **Table 4.2**.

β-Bridges containing motifs. As shown in **Figure 4.3**, structures containing β-bridges are the highest populated over the 20 μs MD. Because of the degree of conformational flexibility a single hydrogen bond allows, structures containing β-bridges can be quite different, ranging from elongated narrow hairpins, to globular folds containing one or two β-turns, see **Figure 4.6**

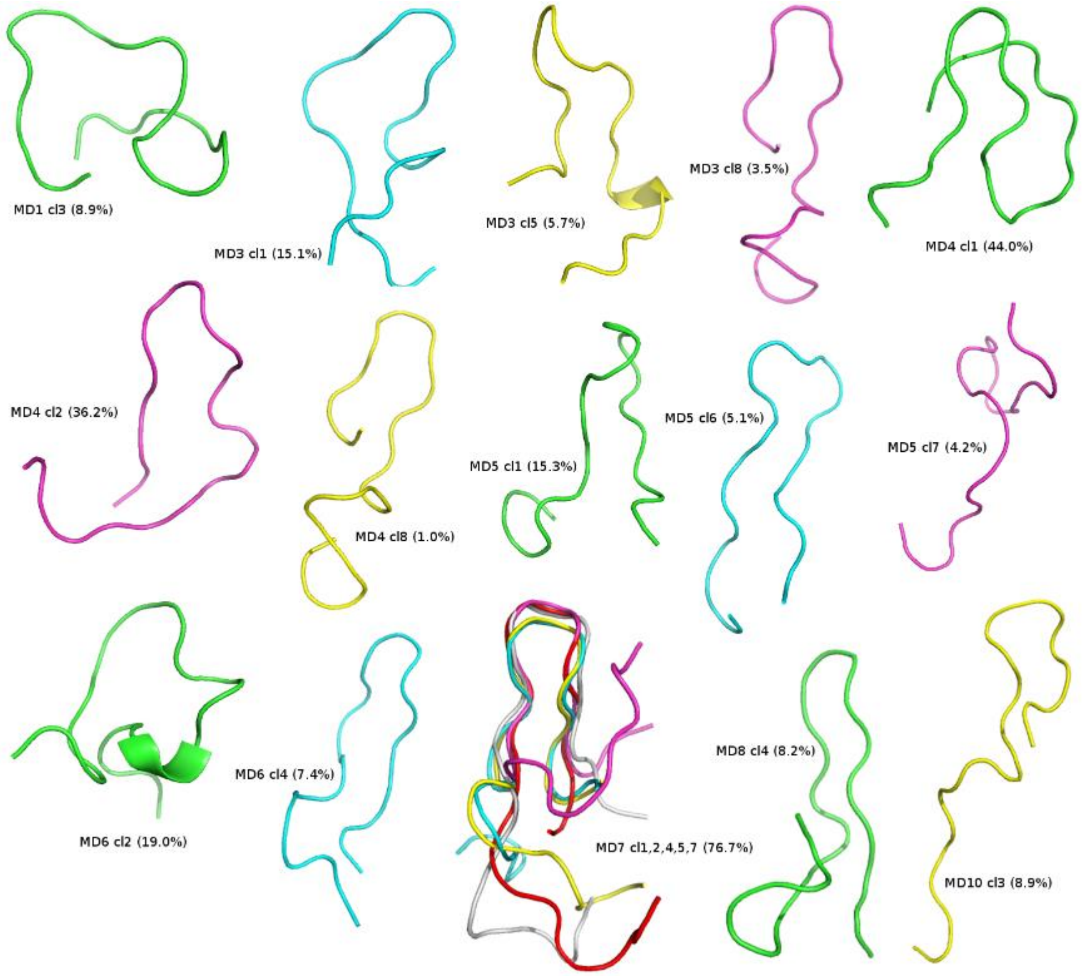


Figure 4.6: Structures containing β-bridges visited during the 20 ms MD simulations. The label beside each structure indicates the MD run and the cluster number, 1 to 10, 1 being the highest populated. The relative population of the structure, or of the structures in case of MD7, over 2 μs is indicated in brackets. Structures are represented with the N-terminal tail on the left-hand side of the image.

Table 4.2: RMSD (Å) values matrix obtained by sequence alignment, followed by structural refinement, of all the β-sheet structural motifs identified during the 20 μs MD simulation. RMSD. The alignment was done with PyMol.

		MD1	MD1	MD1	MD5	MD6	MD6	MD6	MD6	MD7	MD7	MD7	MD8	MD8	MD8	MD8	MD8	MD10	MD10
	b-sheet	cl1	cl2	cl4	cl9	cl1	cl3	cl5	cl8	cl3	cl6	cl9	cl1	cl2	cl3	cl5	cl10	cl1	cl7
MD1	cl1	0.00																	
MD1	cl2	1.02	0.00																
MD1	cl4	0.56	0.90	0.00															
MD5	cl9	4.76	4.60	4.66	0.00														
MD6	cl1	2.24	2.23	2.08	3.87	0.00													
MD6	cl3	2.56	2.51	2.38	3.62	0.67	0.00												
MD6	cl5	2.75	2.79	2.62	4.04	1.63	1.62	0.00											
MD6	cl8	2.85	2.71	2.54	3.46	1.30	0.88	1.82	0.00										
MD7	cl3	2.30	2.33	2.13	4.09	1.40	0.89	1.69	1.79	0.00									
MD7	cl6	2.90	2.98	2.77	3.39	2.32	2.32	2.26	2.26	1.08	0.00								
MD7	cl9	3.47	3.32	3.39	3.23	2.72	2.89	3.52	2.61	3.11	2.91	0.00							
MD8	cl1	2.05	2.59	2.35	4.07	1.31	0.55	1.77	1.72	1.99	2.07	2.47	0.00						
MD8	cl2	2.25	2.09	2.02	4.08	1.69	1.99	2.71	1.55	1.34	2.06	2.84	1.84	0.00					
MD8	cl3	2.29	2.32	2.13	4.04	1.50	1.56	2.26	1.35	1.58	1.93	2.59	0.79	1.31	0.00				
MD8	cl5	2.39	2.42	2.25	3.75	1.33	1.36	1.85	1.47	0.52	2.65	2.56	0.73	1.31	0.72	0.00			
MD8	cl10	2.72	2.70	2.59	4.65	1.39	2.19	2.58	1.78	2.47	3.84	1.37	1.30	1.71	1.46	1.82	0.00		
MD10	cl1	2.31	2.12	2.33	4.59	2.22	3.23	3.45	3.57	3.05	3.55	4.07	3.34	2.87	3.12	3.22	4.43	0.00	
MD10	cl7	5.36	5.30	5.11	4.56	4.74	4.14	3.54	3.68	4.13	4.60	4.74	4.62	4.31	4.30	3.63	4.46	6.02	0.00

The most stable hairpins that contain β-bridges are structurally similar to asymmetric hairpins described in the previous section. In fact, these β-bridge hairpins also present along dynamic tail that comprises residues Ser 367 to Gly 374 ($_{367}$SHLKSKKG$_{374}$) and also have His 380 and Lys 381 at the hairpin turn. The largest group of stable, narrow, asymmetric hairpins was visited during MD7 with 76.7% population over 2 μs MD, see **Figure 4.6**. The structure of the highest populated hairpins containing β-bridges and β-sheet hairpins are structurally very similar; they fall in different categories as the MoRFs structure classification we used is based on the STRIDE definition of the clusters middle structure. Indeed, clusters of narrow hairpins with a β-bridge often also contain β-sheets and *vice versa*.

3$_{10}$ and α helical turns. Over the 20 μs MD we were able to distinguish 3 significantly populated short helical motifs, either 3$_{10}$ or α single helical turns. Representative structures are shown in **Figure 4.7**

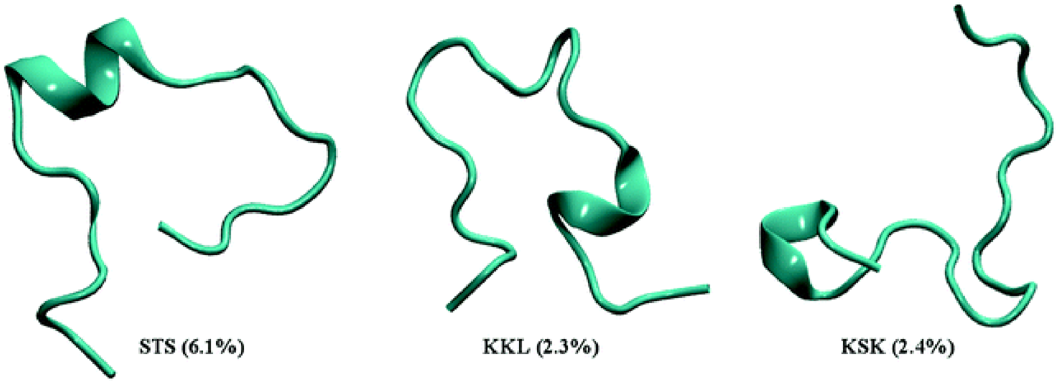


Figure 4.7: Short helical motifs identified during the 20 μs MD simulations. The labels indicate the group of residues where the helical turn is centred, and the minimum stretch for a 3$_{10}$ turn, while the relative populations over 20 μs are indicated in brackets.

The highest populated helical MoRF involves a 3 residue segment, between Ser 376 and Ser 378 ($_{376}$STS$_{378}$), with a relative population of 6.1% (±0.3) over the cumulative 20 μs. Two helical motifs are equally populated; one is located at the N-terminus end of the peptide, stretching across Lys 370 to Lys 372 ($_{370}$KSK$_{372}$) with a relative population of 2.4% (±0.2) over 20 μs, and the other is located at the C-terminus end, between Lys 381 and Leu 383 ($_{381}$KKL$_{383}$), with a relative population of 2.3% (±0.8). Although we found that the single helical turn motifs are in general less stable overtime than the asymmetric β-sheet or β-bridge-containing hairpins, we observed that the highest populated helical turns are stable for between 250 and 450 ns, see **Figure 4.8**.

Stability of the S100B(ββ)-bound conformation in solution. We ran 2 additional 2 μs trajectories to determine the stability in solution of the α/3_{10} helical structure of the 22 residue p53-CTD peptide when in complex with the S100B(ββ) dimer[6]. According to the STRIDE classification, this structure is α helical from Thr 377 to Met 384, and 3_{10} helical from Phe 385 to Thr 387. As shown by the NMR ensemble[6] (PDBid 1dt7), large part of the N-terminal half of the peptide, *i.e.* from Ser 367 to Ser 376, is unbound and highly dynamic. Although the helical structure of the peptide unfolds quite readily within the first 2 ns of both trajectories, the secondary structure analysis in **Table 4.1** shows that a higher degree of helicity remains in one of the two 2 μs trajectories, namely in Helix 1 MD, relative to all other trajectories originated from the common fully extended starting structure. A high degree of helicity in the unbound p53-CTD was described in a recently published computational work[40]. Such strong conformational propensity is inconsistent with circular dichroism data[7,40] and could be

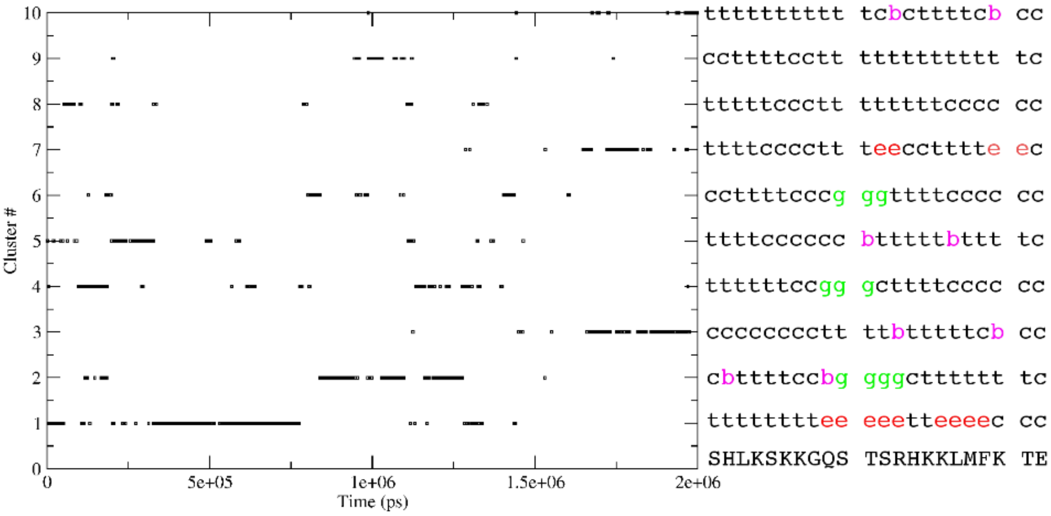


Figure 4.8: Cluster ID overtime calculated for MD10. The secondary structure assignments (STRIDE) for the middle structure of each cluster is indicated on the right-hand side, together with the peptide sequence. B-sheets, 310 helices, b-bridges, coils and turns are indicated with the letters, e (red), g (green), b (pink), c and t (black) respectively.

due not only to force field limitations, but also to the choice of starting structures derived from the S100B(ββ) bound conformation[40]. In 30.7% of the helical structures, the residual helicity spans the $_{381}$KKL$_{383}$ stretch. In all other cases, single helical turns are observed in the disordered, and unbound N-terminal tail, often in addition to the helical turn at $_{381}$KKL$_{383}$. As shown in **Table 4.1**, also during the Helix 2 MD simulation, a residual helical character remains, however much less predominant than in the case of Helix 1 MD. Indeed, the highest populated cluster corresponds to an asymmetric hairpin structure with a free N-terminal tail (Ser 367 to Lys 373), similar to the β-bridge and β-sheet hairpin motifs observed during the 20 μs simulation MD 1 to 10, described in the previous sections. Average R_g values are 9.96 Å for Helix 2 MD and 10.59 Å for Helix 1 MD, see **Figure 4.9**

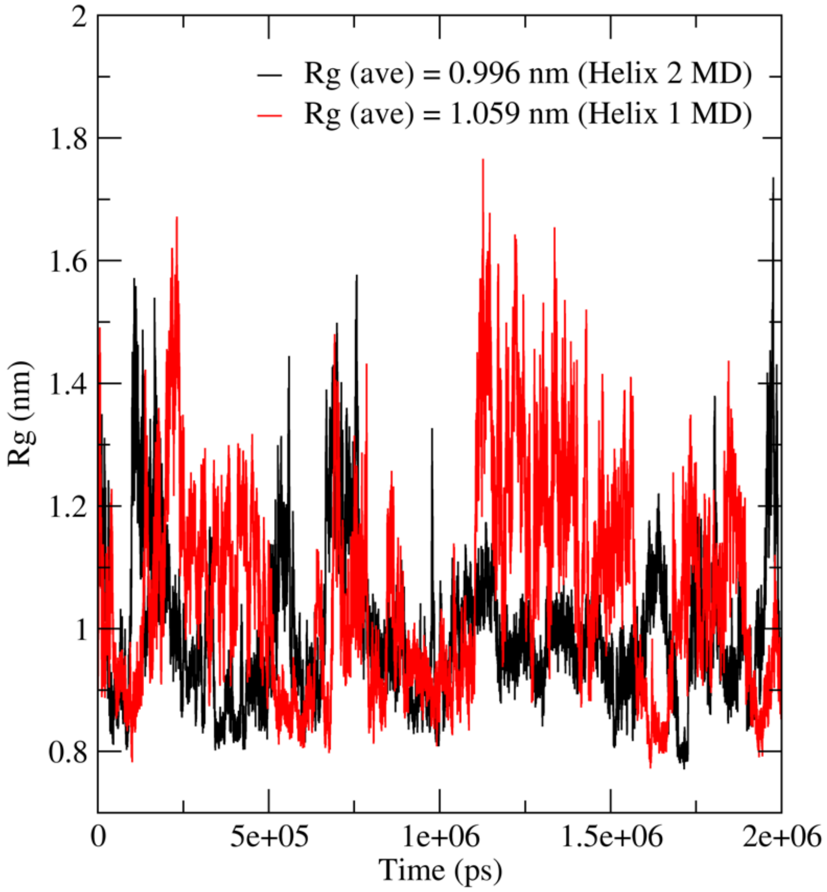


Figure 4.9: Rg values (nm) from the 2 μs trajectories started from the helical conformation of the p53-CTD peptide when bound to the S100B(ββ) (PDBid 1dt7). The Helix 2 MD (black line) was started from structure 6 of the NMR ensemble, while Helix 1 MD (red line) was started from structure 3 of the NMR ensemble.

4.4 Discussion

The extreme C-terminus of the p53 tumour suppressor (p53-CTD) is a 30 residue long conformationally disordered[6], highly alkaline region, responsible for regulating the p53 DNA binding activity[1–4]. This intrinsically disordered region (IDP) is also highly targeted for post-translational modifications, which modulate its DNA binding activity[1,2,5,6,41]. The p53-CTD binds numerous receptors, adopting significantly different conformations when bound[6,8,9]. In this work we used extensive sampling *via* MD simulations to analyse the conformational propensity of a 22-residue peptide, bearing all the binding determinants of the p53-CTD[6,42–44], while unbound in solution. Our objective was to the dynamic nature of the p53-CTD peptide conformational ensemble at equilibrium and to search for structural distinctive elements, or Molecular Recognition Features (MoRFs), that could be specifically selected and bound by different receptors, initiating receptor-specific folding patters. As shown in **Figure 4.3**, we have identified a set of distinct structural MoRFs, significantly populated over 20 μs of cumulative sampling. These include β-sheet and β-bridges-containing asymmetric hairpins, conformations with 3_{10} and α single helical turns, as well as coils and turns-containing structures, which account for the disordered nature of the peptide. An analysis of the size of the peptide in terms of R_g values in relation to the relative populations of the different structural MoRFs is shown in **Figure 4.10**. The most compact, and highest populated, conformations correspond to asymmetric hairpin structures, examples of which are shown in **Figure 4.4**, while the most extended ones correspond to disordered coils.

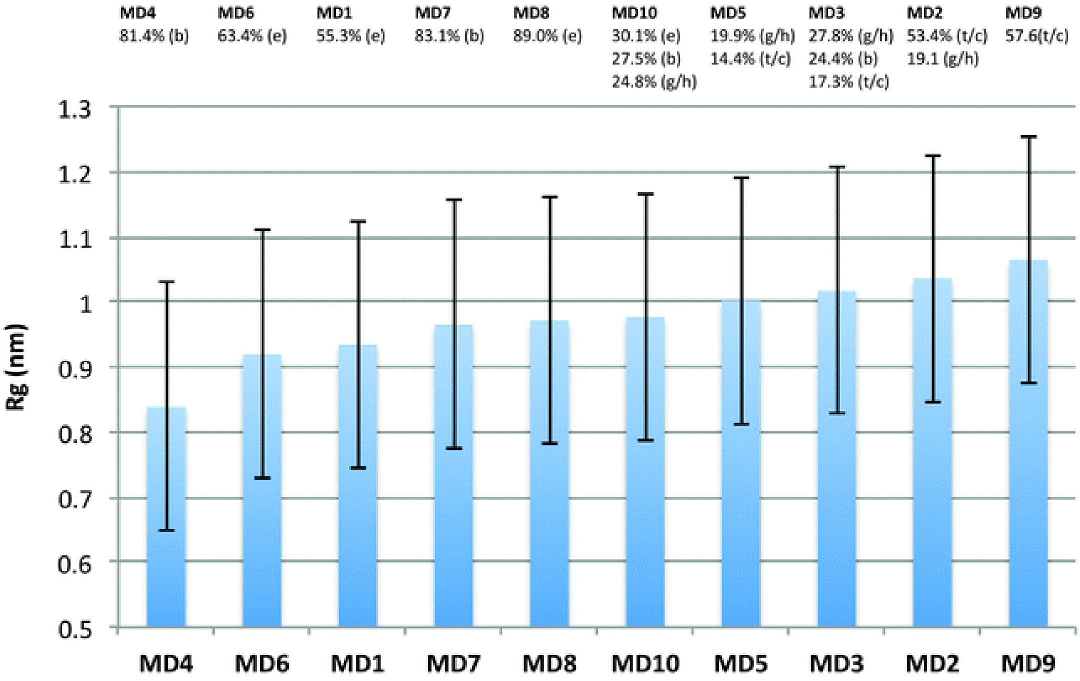


Figure 4.10: Average R_g values (nm), calculated over the 10, 2 μs trajectories, ordered from smaller to larger values. Above each bar are indicated the largest percentages of secondary structures identified during each trajectory, with (e) indicating β-sheet motifs, (b) β-bridges, (g h^{-1}) 3_{10}/α helical single turns, and (t/c) turns and coils.

A comparison between the secondary structure propensity per residue calculated over the 20 μs of MD simulation and the disorder predictions obtained with the s2D[27] and with the PONDR-VL-XT[28] tools is shown in **Figure 4.11**. Provided that the disorder scores obtained from the structure prediction tools are not numerically comparable to the secondary structure propensity values calculated from the MD simulations, or to each other, our analysis shows a good agreement with the s2D data, which predict a decrease in disorder (<70% disorder score) in the $_{377}$TSRHKKLMFKT$_{387}$ segment, where we also observe the highest propensity for secondary structure. The PONDR-VL-XT prediction is also in agreement with our data, showing a smoother decrease in disorder from N- to C-terminus, with a significant decrease for a slightly shorter sequence range relative to the s2D data, namely $_{380}$HKKLMFKTE$_{388}$. As shown in **Figure 4.11**, the MD data provide a rationalization for this decrease in disorder in the C-terminal half of the peptide, by showing a higher propensity for the formation of β-sheet asymmetric hairpins with a disordered N-terminal tail, and single helical turns. The propensity to form short helical motifs is also detected in the mostly disordered N-terminal half of the peptide, a trend mirrored by a slight decrease in disorder predicted by the s2D tool, see **Figure 4.11**, panel (a).

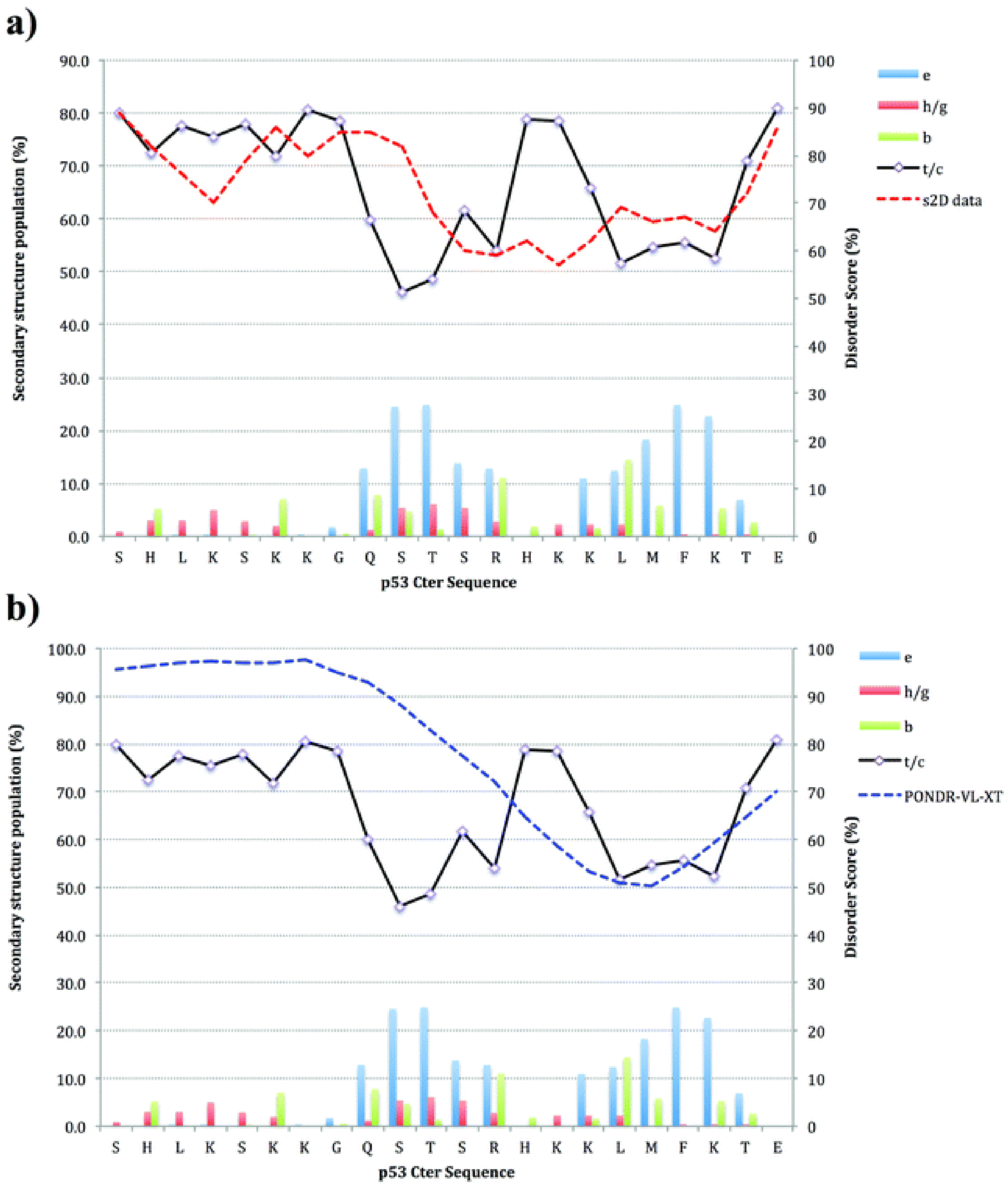


Figure 4.11: Comparison of the conformational propensity per residue calculated from the 20 μs MD simulations to the disorder prediction tools s2D on panel (a), and PONDR-VL-XT on panel (b). The legends on the right-hand side of the graphs indicate specific secondary structure motifs, namely β sheets (e), α (h) and 3_{10} helices (g), β bridges (b), turns (t) and coils (c). Secondary structure assignments have been done with STRIDE[45]

The identification of these minimal structural motifs, or MoRFs, provides a rationale that can explain the binding promiscuity of p53-CTD, or its specificity towards multiple receptors. Our working hypothesis is that the p53-CTD receptors have different binding affinities for the structural MoRFs accessible at equilibrium, because of their 'preformed' 3D spatial arrangements, which provides structural complementarity to different binding site. As the MoRFs constitute minimal structural motifs, the MoRF-receptor structural complementarity provides only on a few protein/peptide specific contacts, resulting in a relatively low binding enthalpy contribution. For instance, when in complex with the S100B(ββ) dimer, the p53-CTD peptide is in a helical conformation[6]. Within our the minimal MoRF-recognition framework, the S100B(ββ) dimer selects one (or more) of the single helical turn-containing MoRFs found in the p53-CTD conformational ensemble to form an initial recognition (or nucleation) complex. Completion of the folding, with a corresponding increment of the binding affinity, will proceed by induced fit[10]. The completion of the helical motif after binding is in agreement with the observation that in the absence of stabilizing tertiary interactions, α-helices rarely persist in isolation[39], and it is also supported by the instability of the S100B(ββ)-bound helical conformation of the p53-CTD peptide when unbound in solution, see **Table 4.1**.

As shown in **Figure 4.12**, an example of a potential recognition complex can be built by structural alignment of the backbone atoms of the $_{376}$STS$_{378}$ helical turn MoRF, the highest populated helical motif over the cumulative 20 μs, onto the bound peptide conformation. The structure of the S100B(ββ)/p53-CTD NMR complex[6] reveals sets of specific ligand-receptor interactions, which include a hydrogen bond between Ser 376 of p53-CTD and Glu 45 of S100B(ββ), a salt-bridge between Arg 379 of p53-CTD and Glu 49 of S100B(ββ), and the insertion of Leu 383 into the hydrophobic binding groove. As shown in **Figure 4.12**, these interactions are conserved in the putative recognition complex and retained during a 100 ns MD structure relaxation run. Notably, the disordered character of the unbound N-terminal tail of the peptide highlighted in the NMR structure[6] is also well reproduced. Another potential example of MoRF-based conformational selection can explain the binding of a p53-CTD peptide with sirtuin Sir2 (PDBid 2h2f)[9]. The only resolved residues of the p53-CTD peptide in the complex are the one directly in contact with the Sir2 binding site, namely $_{378}$SRKKLM$_{383}$. While most of the contact between Sir2 and its target peptide involve mostly backbone atoms[9], as shown in **Figure 4.13**, when in complex with the unmodified p53-CTD peptide, a few significant specific contacts can be highlighted. More specifically, Lys 381 of p53-CTD is in a salt bridge with the Gly 163 backbone carbonyl of Sir2, while the sidechain of Lys 382, targeted by acetylation, protrudes into the Sir2 binding site, with Phe 162 and His 116 flanking the aliphatic side chain,

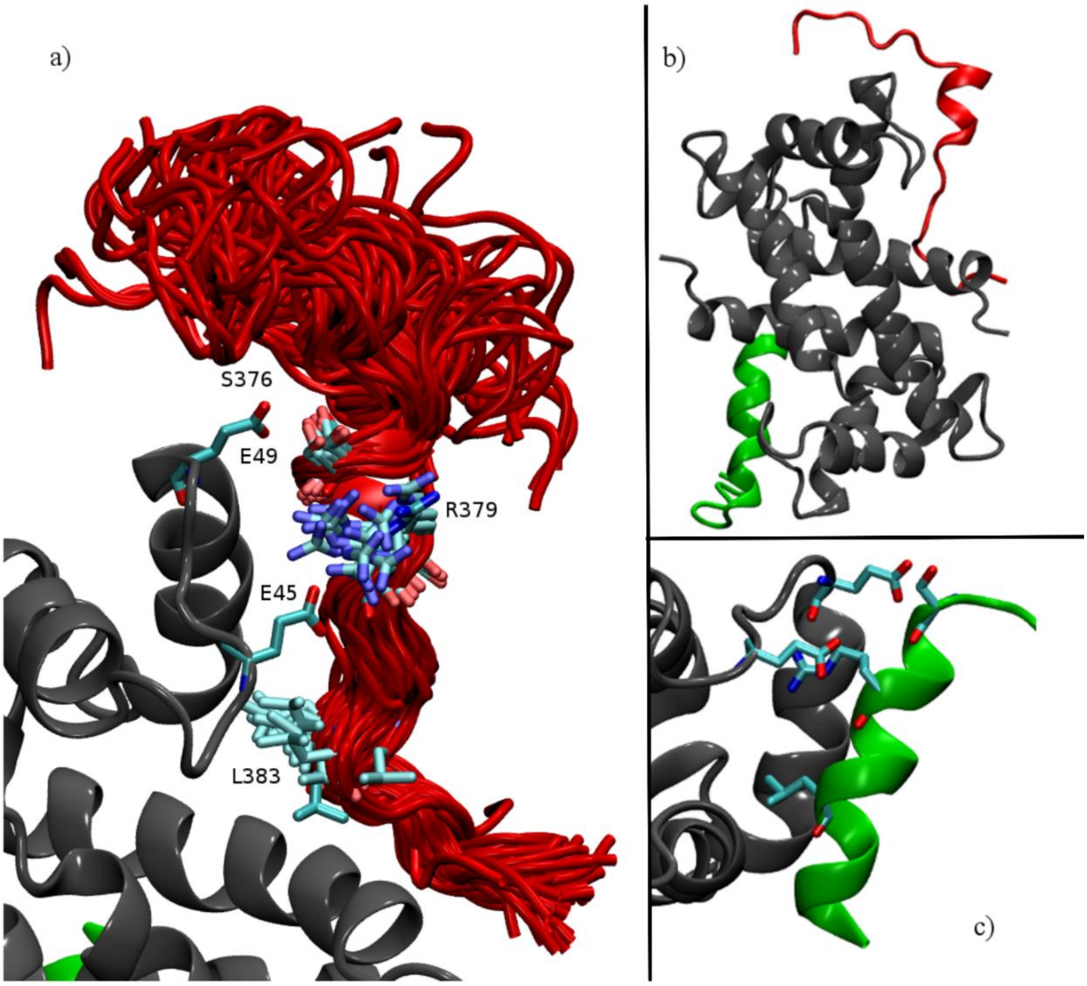


Figure 4.12: (Panel a) Close up on specific contacts the helical MoRF p53-CTD peptide (red) makes during the 100 ns MD when in one of the potential recognition complexes with S100B(ββ) (grey). Shown from top to bottom, the hydrogen bond interaction between Ser 376 of p53-CTD and Glu 49 of S100B(ββ), the salt bridge between Arg 379 of p53-CTD and Glu 45 of S100B(ββ), and the interaction of Leu 383 inserted in the hydrophobic binding groove of S100B(ββ). (Panel b) Structure of the recognition complex built by structural alignment of a peptide containing the STS helical (red) turn on to the p53-CTD bound conformation. (Panel c) Set of contacts highlighted in panel (a) shown for the bound conformation of the p53-CTD peptide (green) in complex with S100B(ββ), PDBid 1dt7.

and the amino group bound to a water molecule[9]. Based on the structural data available, the transition from conformational disorder to order upon binding appears to be less significant in this case relative to the S100B(ββ). Nevertheless, if we consider the interactions between the Sir2 and Lys 381 and 382 of p53-CTD as potential recognition contacts, we have 2 examples, shown in **Figure 4.13**, of potentially recognized MoRFs in coil/turn conformation, with an optimal orientation of the Lys sidechains for initial recognition.

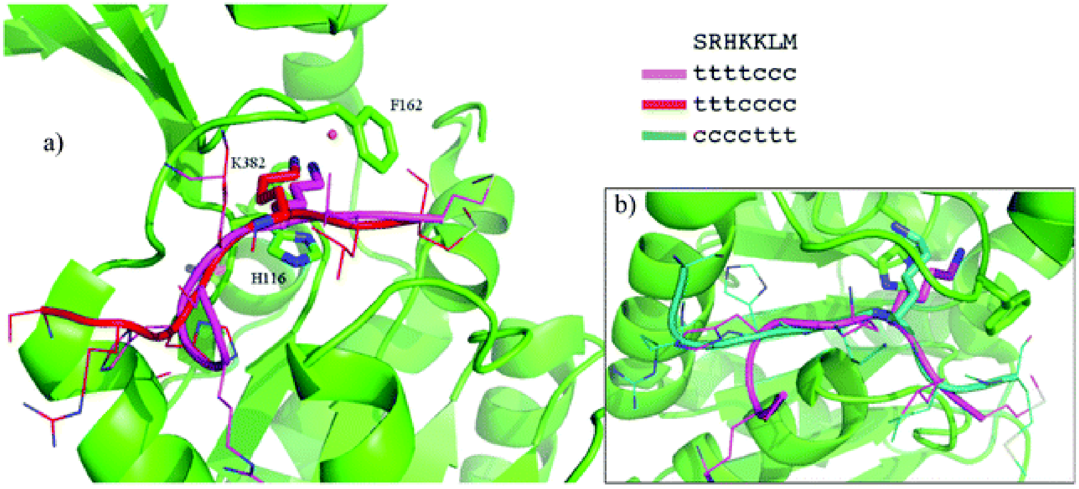


Figure 4.13: Potential recognition complexes between the p53-CTD peptide (in purple) and Sir2Tm (in green), PDBid 2h2f, obtained by structural alignment of MoRFs identified through clustering analysis of 20 μs MD trajectories. In panel (a) alignment of the middle structure from cluster 8 of MD5 (in red) with a RMSD based on 4 Ca atoms of 0.4 Å, in panel (b) alignment of the middle structure from cluster 8 of MD9 (in cyan) with a RMSD based on 4 Ca atoms of 0.3 Å. Peptide sequence and respective secondary structures assignments (STRIDE) are indicated in the legend.

To our knowledge there are no structures of complexes with the unmodified p53-CTD in an asymmetric β-sheet conformation; nevertheless, we are currently investigating the recognition of the highest populated p53-CTD MoRF, namely the β-sheet or β-bridge-containing asymmetric hairpin, by β-sheet structured binding sites, such as in the PCL1-PHD1 domain[46].

Because of their relatively low populations and frequent interconversion, the p53-CTD peptide structural MoRFs will be extremely difficult, if not impossible, to characterize experimentally. Indeed, existing NMR[6] and CD[40] data show that the p53-CTD section is highly disordered when unbound in solution. Nevertheless, based on the information discussed in this work, experimental support for the MoRF-driven molecular recognition mechanism could be obtained by biasing the conformational ensemble towards specific structural MoRFs by means, for example, of stapling the peptide[47]. More specifically, based on the potential recognition discussed earlier in this section of the $_{376}$STS$_{378}$ helical turn by S100B(ββ) as the initial, low affinity recognition complex, a suitably placed aliphatic chain staple would enhance the conformational propensity of the helical MoRF, thus its population, without affecting sequence integrity. Changes in relative conformers populations would affect binding kinetics. Furthermore, the role of MoRFs in p53-CTD recognition and binding could also be tested by introducing mutations that suppress specific structural motifs. Indeed, in the case of the small disordered protein PUMA, an Ala-Gly scanning scheme was put in place to suppress helicity without affecting residual structural propensity in solution[19]. Binding kinetics showed that the mutations do not affect k+, suggesting that folding occurs by an induced fit-driven mechanism[19]. As a proof of principle, based on the recognition mechanism proposed earlier for the $_{376}$STS$_{378}$ 3_{10} helical turn MoRF by S100B(ββ), mutations of Thr 377 and/or Ser 378 would affect conformational propensity without compromising specific contacts with the receptor. Work in this direction is currently underway.

4.5 Conclusions

In this work we have used extended conformational sampling through conventional MD simulations to determine the degree of residual secondary structure within the conformational disorder at equilibrium of a 22-residue peptide, corresponding to the 367–388 region of the p53 C-terminal domain (p53-CTD). This peptide contains all binding determinants of the p53-CTD within the active p53 tumour suppressor[6,9]. Clustering analysis of the MD trajectories, accounting for a cumulative time of over 20 μs, show the p53-CTD peptide has a high conformational flexibility, but also a distinct propensity for the formation of specific and short structural motifs that encompass 3 to 4 residues at most. Furthermore, a *per* residue analysis of the conformational propensity along the p53-CTD peptide shows that these structural motifs are localized along the sequence, involving specific groups of residues. Localization of the structural MoRFs makes the p53-CTD C-terminal half less disordered than the N-terminal half. This observation is also in agreement with disorder predictions obtained with the s2D[27] and PONDR-VL-XT[28] secondary structure and disorder prediction tools. We propose that the functional role of these minimal structural molecular recognition features (MoRFs) is to confer to the p53-CTD binding specificity towards different receptors, whereby each receptor would have a higher affinity for a specific MoRF due to 3D structural complementarity, based on a reduced number of contacts, relative to the final, bound conformation. Molecular recognition through selection of specific MoRFs, would lead to the completion of folding through induced fit. This mechanism is consistent with the molecular recognition proposed for other IDP systems[17], although it does not necessarily preclude access to other recognition and binding pathways[17,19], that could in principle coexist.

References

1. Ahn, J. & Prives, C. The C-terminus of p53: the more you learn the less you know. *Nature structural biology* **8**, 730–732 (2001).
2. Laptenko, O. *et al.* The p53 C terminus controls site-specific DNA binding and promotes structural changes within the central DNA binding domain. *Mol. Cell* **57**, 1034–1046 (2015).
3. Joerger, A. C. & Fersht, A. R. Structural biology of the tumor suppressor p53. *Annu. Rev. Biochem.* **77**, 557–582 (2008).
4. Hupp, T. R., Meek, D. W., Midgley, C. A. & Lane, D. P. Regulation of the specific DNA binding function of p53. *Cell* **71**, 875–886 (1992).
5. Feng, L., Lin, T., Uranishi, H., Gu, W. & Xu, Y. Functional analysis of the roles of posttranslational modifications at the p53 C terminus in regulating p53 stability and activity. *Mol. Cell. Biol.* **25**, 5389–5395 (2005).
6. Rust, R. R., Baldisseri, D. M. & Weber, D. J. Structure of the negative regulatory domain of p53 bound to S100B(ββ) . *Nat. Struct. Biol.* **7**, 570–574 (2000).
7. Shahar, O. D. *et al.* Acetylation of lysine 382 and phosphorylation of serine 392 in p53 modulate the interaction between p53 and MDC1 in vitro. *PLoS One* **8**, e78472 (2013).
8. Huart, A.-S. & Hupp, T. R. Evolution of Conformational Disorder & Diversity of the P53 Interactome. *Biodiscovery* **8**, e8952 (14AD).
9. Cosgrove, M. S. *et al.* The structural basis of sirtuin substrate affinity. *Biochemistry* **45**, 7511–7521 (2006).
10. Csermely, P., Palotai, R. & Nussinov, R. Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. *Trends Biochem. Sci.* **35**, 539–546 (2010).
11. Boehr, D. D., Nussinov, R. & Wright, P. E. The role of dynamic conformational ensembles in biomolecular recognition. *Nat. Chem. Biol.* **5**, 789–796 (2009).
12. Allen, W. J., Capelluto, D. G. S., Finkielstein, C. V & Bevan, D. R. Modeling the relationship between the p53 C-terminal domain and its binding partners using molecular dynamics. *J. Phys. Chem. B* **114**, 13201–13213 (2010).
13. Kannan, S., Lane, D. P. & Verma, C. S. Long range recognition and selection in IDPs: the interactions of the C-terminus of p53. *Sci. Rep.* **6**, 23750 (2016).
14. Keskin, O., Ma, B., Rogale, K., Gunasekaran, K. & Nussinov, R. Protein-protein interactions: organization, cooperativity and mapping in a bottom-up Systems Biology approach. *Phys. Biol.* **2**, S24-35 (2005).
15. Hayashi, T., Oshima, H., Yasuda, S. & Kinoshita, M. Mechanism of One-to-Many Molecular Recognition Accompanying Target-Dependent Structure Formation: For the Tumor Suppressor p53 Protein as an Example. *J. Phys. Chem. B* **119**, 14120–14129 (2015).
16. Toto, A. *et al.* Molecular Recognition by Templated Folding of an Intrinsically Disordered Protein. *Sci. Rep.* **6**, 21994 (2016).

17. Arai, M., Sugase, K., Dyson, H. J. & Wright, P. E. Conformational propensities of intrinsically disordered proteins influence the mechanism of binding and folding. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 9614–9619 (2015).

18. Metskas, L. A. & Rhoades, E. Order-Disorder Transitions in the Cardiac Troponin Complex. *J. Mol. Biol.* **428**, 2965–2977 (2016).

19. Rogers, J. M. *et al.* Interplay between partner and ligand facilitates the folding and binding of an intrinsically disordered protein. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 15420–15425 (2014).

20. Fuxreiter, M., Simon, I., Friedrich, P. & Tompa, P. Preformed structural elements feature in partner recognition by intrinsically unstructured proteins. *J. Mol. Biol.* **338**, 1015–1026 (2004).

21. Mittag, T., Kay, L. E. & Forman-Kay, J. D. Protein dynamics and conformational disorder in molecular recognition. *J. Mol. Recognit.* **23**, 105–116 (2010).

22. Dyson, H. J. & Wright, P. E. Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.* **6**, 197–208 (2005).

23. Tompa, P. The interplay between structure and function in intrinsically unstructured proteins. *FEBS Lett.* **579**, 3346–3354 (2005).

24. Babu, M. M., van der Lee, R., de Groot, N. S. & Gsponer, J. Intrinsically disordered proteins: regulation and disease. *Curr. Opin. Struct. Biol.* **21**, 432–440 (2011).

25. van der Lee, R. *et al.* Classification of Intrinsically Disordered Regions and Proteins. *Chem. Rev.* **114**, 6589–6631 (2014).

26. Tompa, P. The functional benefits of protein disorder. *J. Mol. Struct. THEOCHEM* **666–667**, 361–371 (2003).

27. Sormanni, P., Camilloni, C., Fariselli, P. & Vendruscolo, M. The s2D Method: Simultaneous Sequence-Based Prediction of the Statistical Populations of Ordered and Disordered Regions in Proteins. *J. Mol. Biol.* **427**, 982–996 (2015).

28. Romero, P. *et al.* Sequence complexity of disordered protein. *Proteins Struct. Funct. Bioinforma.* **42**, 38–48 (2001).

29. Krüger, A. *et al.* Interactions of p53 with poly(ADP-ribose) and DNA induce distinct changes in protein structure as revealed by ATR-FTIR spectroscopy. *Nucleic Acids Res.* **47**, 4843–4858 (2019).

30. Maestro, version 9.2. Schrödinger, LLC; New York, NY, U. 2011. .

31. Lindorff-Larsen, K. *et al.* Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* **78**, 1950–1958 (2010).

32. Horn, H. W. *et al.* Development of an improved four-site water model for biomolecular simulations: TIP4P-Ew. *J. Chem. Phys.* **120**, 9665–9678 (2004).

33. Galindo-Murillo, R., Roe, D. R. & Cheatham 3rd, T. E. Convergence and reproducibility in molecular dynamics simulations of the DNA duplex d(GCACGAACGAACGAACGC). *Biochim. Biophys. Acta* **1850**, 1041–1058 (2015).

34. Hess, B., Bekker, H., Berendsen, H. J. C. & Fraaije, J. G. E. M. LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem.* **18**, 1463–1472 (1997).

35. Abraham, M. J. *et al.* GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1–**

2, 19–25 (2015).

36. Daura, X. *et al.* Peptide Folding: When Simulation Meets Experiment. *Angew. Chemie Int. Ed.* **38**, 236–240 (1999).

37. Frishman, D. & Argos, P. Knowledge-based protein secondary structure assignment. *Proteins Struct. Funct. Bioinforma.* **23**, 566–579 (1995).

38. The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, L. .

39. Zagrovic, B., Jayachandran, G., Millett, I. S., Doniach, S. & Pande, V. S. How large is an alpha-helix? Studies of the radii of gyration of helical peptides by small-angle X-ray scattering and molecular dynamics. *J. Mol. Biol.* **353**, 232–241 (2005).

40. Iida, S. *et al.* Variation of free-energy landscape of the p53 C-terminal domain induced by acetylation: Enhanced conformational sampling. *J. Comput. Chem.* **37**, 2687–2700 (2016).

41. Xu, Y. Regulation of p53 responses by post-translational modifications. *Cell Death Differ.* **10**, 400–403 (2003).

42. Avalos, J. L. *et al.* Structure of a Sir2 enzyme bound to an acetylated p53 peptide. *Mol. Cell* **10**, 523–535 (2002).

43. Lowe, E. D. *et al.* Specificity determinants of recruitment peptides bound to phospho-CDK2/cyclin A. *Biochemistry* **41**, 15625–15634 (2002).

44. Mujtaba, S. *et al.* Structural mechanism of the bromodomain of the coactivator CBP in p53 transcriptional activation. *Mol. Cell* **13**, 251–263 (2004).

45. Heinig, M. & Frishman, D. STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. *Nucleic Acids Res.* **32**, W500-2 (2004).

46. Brien, G. L. *et al.* A chromatin-independent role of Polycomb-like 1 to stabilize p53 and promote cellular quiescence. *Genes Dev.* **29**, 2231–2243 (2015).

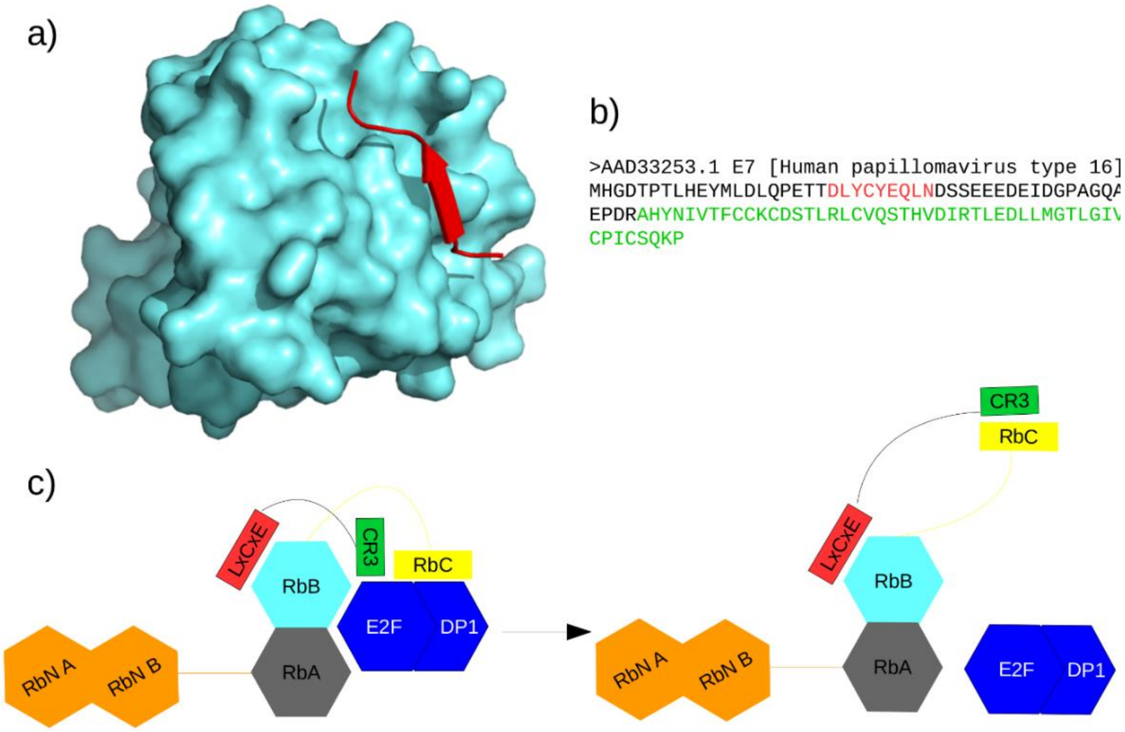
47. Walensky, L. D. & Bird, G. H. Hydrocarbon-stapled peptides: principles, practice, and progress. *J. Med. Chem.* **57**, 6275–6288 (2014).

Chapter 5: Effects on the conformational propensity of the variable residues on LxCxE short linear motifs and implications on binding affinity

5.1 Introduction

The Human Papilloma Virus (HPV) is responsible for almost all cases of cervical cancer, where the majority of these are caused by HPV 16/18 strain[1]. HPV has also been implicated in a number of other cancers[1]. After cell infection, HPV promotes cellular proliferation triggered by the interaction between the human retinoblastoma protein (Rb), a 928-residue tumour suppressor protein, and the HPV E7 oncoprotein[2]. The anti-tumour action of Rb derives from it arresting cell cycle at the G1 phase through binding the E2F family of proteins[3]. As shown in **Figure 5.1**, the Rb B-domain binds specifically a short linear motif (SLiM), namely LxCxE[2], which is found in the HPV16 protein E7 located in the intrinsically disordered CR2 region of E7[4,5]. The LxCxE motif is a high affinity binder for Rb, with a peptide with sequence corresponding to residues 16 to 40 of E7 ($E7_{16\text{-}40}$) having a Kd in the low nM range[6]. The binding of the LxCxE motif may promote the E7 CR3-region to bind the Rb C-domain[3] which displaces the E2F "marked box" region and destabilises the Rb-E2F complex[7–9], releasing E2F and causing the cell to progress from the G1 phase to S phase[10]. E2F itself does not bind Rb through an LxCxE motif[7,11]. LxCxE is a highly conserved motif found in many viral proteins and in cellular proteins, see **Figure 5.2**.

In this work, in collaboration with Prof Lucia Chemes at the Universidad de San Martin in Buenos Aires, we performed enhanced sampling simulations to gain an insight into the role of the variable residues, i.e. the x residues in LxCxE, on the motif's binding affinity and on its structural propensity. More specifically, our aim was to determine if the choice of variable residues would affect pre-structuring of the motif to a conformation complementary to the Rb binding site, thus affecting the entropy of binding, regulating its binding affinity. Within this framework, in addition to the variable residues in the motif, we also looked at the effect of the LxCxE flanking residues, namely the residues immediately before and after the motif with which the x residues could interact. In terms of the simulations set-up, we examined the role of salt



```
>AAD33253.1 E7 [Human papillomavirus type 16]
MHGDTPTLHEYMLDLQPETTDLYCYEQLNDSSEEEDEIDGPAGQA
EPDRAHYNIVTFCCKCDSTLRLCVQSTHVDIRTLEDLLMGTLGIV
CPICSQKP
```

Figure 5.1: Potential displacement pathway of E2F from Rb by E7. a) The pRb B domain (in cyan) is represented through its solvent accessible surface. The LxCxE binding motif shown in red (PDBid 1GUX) with the corresponding sequence highlighted in red in Panel b). Panel c) shows the pRb C-domain (in yellow) bound to the heterodimer E2F-DP1 (in blue) and binding of the E7 CR3 region (in green) which displaces the E2F from the Rb C region and causes the dissociation of E2F-DP1 from the RbAB interface.

```
HPV16 E7   ETTDLYCYEQLNDSSEEEDEIDG
HPV18 E7   IPVDLLCHEQLSDSEEENDEIDG
HPV11 E7   DPVGLHCYEQLEDSSEDEVDKV-
Sv40LT     NEENLFCSEEMPSSDDEATADS-
AdE1A      ETTDLYCYEQGFPLSDEEDHEKE
Cyclin D   MEHQLLCCEVETIRRAYPDANLL
BRG1       EVERLTCEEEEEKMFGRGSRHRK
BRCA1      NKQKLPCSENPRDTEDVPWITLN
```

Figure 5.2: Sequence alignment of LxCxE motifs from selected viral and cellular proteins, listed in bold on the left-hand side. The LxCxE motif is highlighted in red, while the region rich in acidic residues is highlighted in blue. The Ser residues in blue are known to be phosphorylated

concentration in the simulations by comparing simulations in bulk water to simulations in 200 mM NaCl, which corresponds more closely to the experimental conditions we are comparing our results to. Also, we looked at the effects of phosphorylation at Ser 31 and 32 (pSer), which are conserved almost to the extent as the LxCxE motif[6], corresponding to the naturally active form of the E7 protein[5,12]. Notably, the peptides tested in the Prof Chemes' lab for binding kinetics are not phosphorylated, while phosphorylation in such positions is known increase the binding ten-fold, the non-phosphorylated peptides still bind in the low nM range[6]. In addition to HPV16 E7 LxCxE variants we also studied the structure and dynamics of a few other known and potential LxCxE motifs, namely from Cyclin D, BRCA1, FOG2 and GATA1 proteins. From all these results we found no evidence of prestructuring, which for the E7 peptide variants is consistent with experimental results see **Table 5.5** in results. We also found that salt concentration has minimal effect on the structure and dynamics of the peptides. Finally, our results of the Rb-bound LxCxE-containing peptides are able to clearly explain the experimental kinetic data and the differences in affinity of the different mutants and variants.

5.2 Interaction between the HPV E7 LxCxE and Rb

In HPV 16 E7 the variable residues are both Tyr residues, i.e. the motif is LYCYE as shown in **Figure 5.2**, and the flanking residues we are concerned with are Asp 21 and Leu 28. The interactions between the variable and flanking residues play a role in modulating the binding affinity of the motif to Rb. The mutations proposed in this work, shown in **Table 5.1,** were designed to alter or eliminate one or more of these interactions at same the time. The main interactions found in the crystal structure of the HPV E7 peptide in complex with Rb (PDBid 1GUX) are a hydrogen bond between Tyr 25 and Asp 21, a π-stacking interaction between Tyr 23 and Tyr 25. Additionally, there is a stacking between Tyr 23 and Rb Asn 757, Asn 757 forms two hydrogen bonds with the backbone of the E7 LxCxE motif, the stacking between Tyr 23 and Rb Asn 757 may serve to exclude water from interacting with Asn 757 which would reduce the strength of the hydrogen bonds to E7. We also looked at the importance of the highly acidic region of E7 found after the LxCxE motif, namely residues 30-40, as there is a highly alkaline patch on Rb and the interaction between the two oppositely charged regions maybe be important for binding. Leu 22, Cys 24 and Leu 28 sit into hydrophobic pockets on Rb while Glu 26 forms a bidentate hydrogen bond with the backbone of Rb, see **Figure 5.3**.

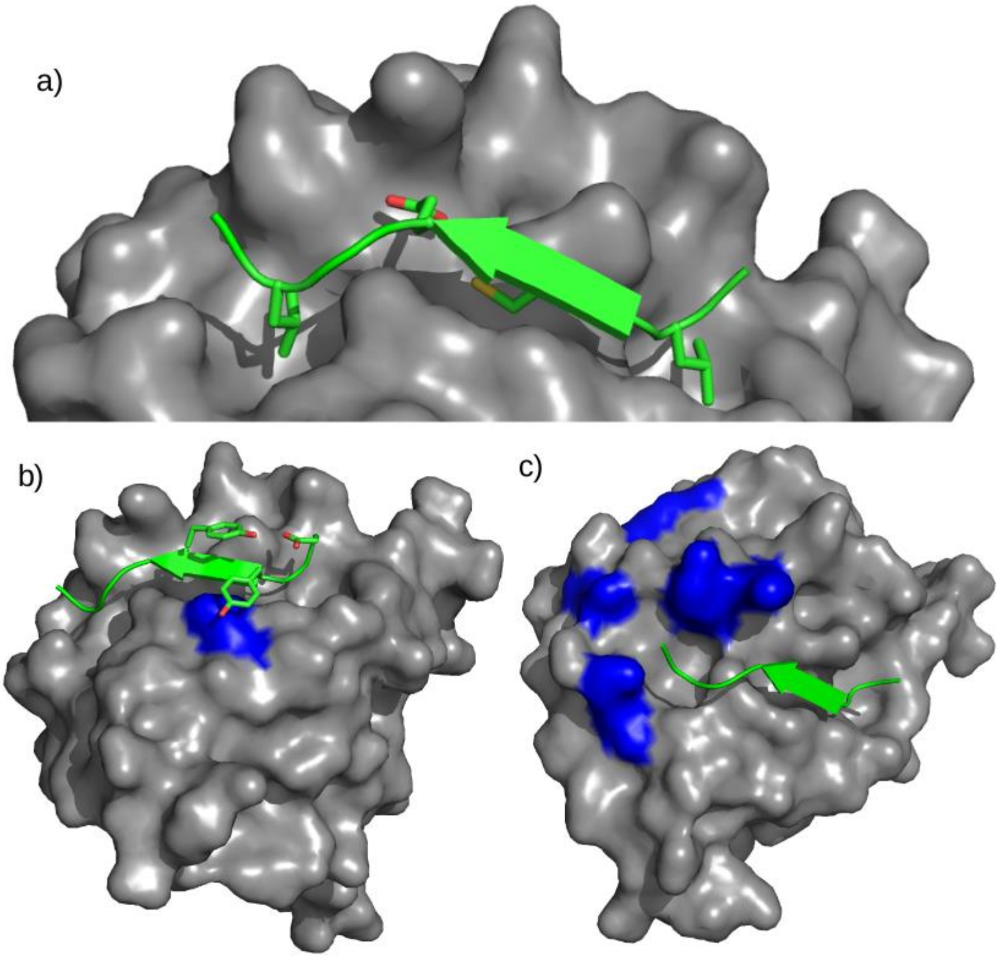


Figure 5.3: Crystal structure of E7 21-28 shown in green bound to Rb B-region shown in grey (PDBid 1GUX). A) From right to left, interactions between conserved E7 Leu 22, Cys 24, Glu 26 and Leu 28 and Rb. B) Variable residue interactions of E7, π-stacking between Tyr 23 and 25, Hydrogen bond between Asp21 and Tyr 25 and stacking between Asp 23 and Rb Asn 757 (shown in blue). C) Alkaline patch of Rb, Arg and Lys residues are shown in blue.

The Y23A mutation eliminates the π-stacking with Tyr 25 of the E7 and with Asn 757 of Rb, the Y25F mutation aims to determine the importance of the hydroxyl group of the Tyr in this specific position as the hydroxyl forms a hydrogen bond with Asp 21 in the crystal structure. The Y25A mutation represents the reversed case of the previous one, eliminating the π-stacking with Tyr 23 and the hydrogen bond to Asp 21. The D21A and Y25F mutations both eliminate the hydrogen bond between Ty r25 and Asp 21. Finally, the E26A mutation was proposed based on preliminary kinetic data to assess the role of the highly conserved Glu from the LxCxE motif, while a series of Ala “spacers” is designed to push Leu 28 out of the hydrophobic pocket it sits in, see **Figure 5.3b**. Also, the LxCxE motifs from GATA1 and FOG2 proteins, see **Table 5.1**, were introduced into the E7 peptide frame to investigate how different variable residues known to bind Rb affect binding of the E7 peptide.

5.3 Computational method

In this study we analysed the conformational dynamics of 24 residue peptides with sequences shown in **Table 5.1** by replica exchange MD (REMD) simulations[13] both free in solution and bound to Rb. The simulations were repeated in 200 mM NaCl and with peptides phosphorylated and non-phosphorylated at Ser 31 and 32. All REMD simulations were carried out with 87 replicas to cover a range of temperatures between 300 K to 500 K, chosen based on the "*Temperature Generator for REMD"* online tool (http://folding.bmc.uu.se/remd/)[14]. All simulations were run with CHARMM22*[15] to represent the protein atoms and counterions and with TIP4P-D as a water model[16]. This force field combination was chosen as at the time we started the project in 2016 it was the one for intrinsically disordered proteins (IDPs) that best agreed with experimental results[17]. All calculations were run with two GROMACS versions, namely 4.6.3 and 2018.3[18]. Computational resources were provided by the Irish Centre for High-End Computing (ICHEC). Energy minimizations were carried out on 2 nodes, i.e. 80 cores. NVT equilibrations and Replica Exchange MD (REMD) production simulations were run on 2 processors per replica for 87 replicas, i.e. 174 processors over 5 nodes. We estimated that the final cost for all MD simulations in this project reached approximately 6,000,000 CPU hours. We performed the same simulations as described above for a number of mutants of E7$_{17\text{-}40}$ as well as LxCxE containing peptides from BRCA, corresponding to residues 353-376, Cyclin D, corresponding to residues 1-23, FOG2, corresponding to residues 40-63 and GATA1, corresponds to residues 76-99. In the following text, all peptides are numbered based on the E7 numbering going forward with the LxCxE motif as residue 22 to 26. The N-termini of all peptides were capped with an acetyl group (ACE) and the C-termini were capped with an N-methyl (NME) group. All systems studied here, i.e. bound and unbound peptides, were placed in a rhombic dodecahedron simulation box with a minimum distance between the peptide and the edge of the box of 1.2 nm. Bond lengths to hydrogen atoms were constrained using the LINCS algorithm. Long range electrostatic interactions were treated with Particle Mesh Ewald (PME) with a switch from real space to reciprocal space at 1.2 nm. The van der Waals (vdW) interactions were cut-off at 1.2 nm.

Table 5.1: Peptides sequences studied in this work with the conditions under which they were simulated. All peptides were simulated free in solution and bound to Rb except E7PE26A, which was only simulated when bound to Rb. E7 mutations are highlighted in bold and phosphoserines are underlined. Checkmarks indicate whether the simulations were done in both bulk water and 200 nM NaCl.

Peptides	Sequence	bulk water	200 mM NaCl
E7	17 PETTDLYCYEQLNDSSEEEDEIDG 40	✓	✓
E7P	PETTDLYCYEQLNDSSEEEDEIDG	✓	✓
E7P$_{D21A}$	PETT**A**LYCYEQLNDSSEEEDEIDG	✓	✓
E7P$_{E26A}$	PETTDLYCY**A**QLNDSSEEEDEIDG	✓	✓
E7P$_{AA}$	PETTDL**A**C**A**EQLNDSSEEEDEIDG	✓	✓
E7$_{AY}$	PETTDL**A**CYEQLNDSSEEEDEIDG	✓	✓
E7P$_{AY}$	PETTDL**A**CYEQLNSSEEEDEIDG	✓	✓
E7$_{FF}$	PETTDL**F**C**F**EQLNDSSEEEDEIDG	✓	✓
E7P$_{FF}$	PETTDL**F**C**F**EQLNDSSEEEDEIDG	✓	✓
E7P$_{FY}$	PETTDL**F**CYEQLNDSSEEEDEIDG	✓	✓
E7P$_{YA}$	PETTDLYC**A**EQLNDSSEEEDEIDG	✓	✓
E7P$_{YF}$	PETTDLYC**F**EQLNDSSEEEDEIDG	✓	✓
FOG2	FGPENLSCEEVEYFCNKGDDEGIQ		✓
E7P$_{FOG}$	PETT**N**L**S**C**E**EQLNDSSEEEDEIDG		✓
GATA1	QVYPLLNCMEGIPGGSPYAGWAYG		✓
E7P$_{GATA}$	PETTLLNCMEQLNDSSEEEDEIDG		✓
E7$_{1A}$	PETTDLYCYEQ**A**LNDSSEEEDEIDG	✓	✓
E7$_{2A}$	PETTDLYCYEQ**AA**LNDSSEEEDEIDG	✓	✓
E7$_{3A}$	PETTDLYCYEQ**AAA**LNDSSEEEDEIDG	✓	✓
E7P$_{1A}$	PETTDLYCYEQ**A**LNDSSEEEDEIDG	✓	✓
E7P$_{2A}$	PETTDLYCYEQ**AA**LNDSSEEEDEIDG	✓	✓
E7P$_{3A}$	PETTDLYCYEQ**AAA**LNDSSEEEDEIDG	✓	✓
BRCA	WNKQKLPCSENPRDTEDVPWITLN		✓
Cyclin D	MEHQLLCCEVETIRRAYPDANLL		✓

The starting structure for all bound peptides was based on the crystal structure of the E7 LxCxE motif bound to Rb (PDBid 1GUX), with the peptide extended in both directions built with Schrodinger MAESTRO[19], minimising the contact with the Rb protein to prevent biasing the starting conformation of the complex. In the interest of computational resources Rb A-box domain was omitted from the simulations. This starting structure was modified using *pymol* [20] to obtain the different sequences shown in **Table 5.1**. For the starting conformation of the peptides free in solution, we build the 24 residue E7 peptide in a fully extended conformation also with the MAESTRO molecular builder. For the extended peptide a 500k steps of steepest descent minimization was carried out in implicit solvent, followed by a 100 ns NPT MD run and a random structure was selected as the starting conformation for REMD. The REMD simulation protocol for both free E7 and E7 bound to RB involved an initial energy minimization, namely 500k of steepest descent with a force-based convergence threshold of 100 kJmol^{-1}nm^{-1} restraining the protein heavy atoms. After the energy minimization, we carried out a 500 ps equilibration in the NVT ensemble with the same set of restraints, where each replica was equilibrated to its specific temperature. To ensure the system maintained a pressure of 1 bar, the density of the system at 300 K was monitored throughout the simulation and verified that its value matched that of water at 300 K and 1 bar, so no that further equilibration was required. To integrate the equations of motion we used a leap-frog stochastic dynamics (sd) integrator, with a friction coefficient corresponding to the inverse of tau-t equal to 0.1 ps, where tau-t is the time constant for coupling. Production stage followed with all atoms unrestrained in the free peptide simulations and with only Rb backbone atoms restricted in the bound peptide simulations. The system attempted to exchange replicas every 500 steps, i.e. every 1 ps, and the production was extended to 26 ns per replica, for a total cumulative simulation time of 2.3 μs for each peptide.

As an important note, based on Dr P. Robustelli's personal communication, the TIP4P-D water model[21] was shown to partially unfold folded proteins, therefore we kept the Rb backbone atoms restrained in all simulations discussed in the Results section. Nevertheless, we tested the effects of TIP4P-D in a fully unrestrained complex and indeed found that the water model contributes to unfold helical motifs on the structured receptor. The test results are shown in **Figure 5.4**

5.4 Results

HPV16 E7 peptide. As the variable residues in the LxCxE are not evolutionarily conserved, we studied the role of these variable residues in the structure and dynamics of the peptides. More specifically, we looked at the interactions present in the crystal structure of the Rb-HPV E7 complex (PDBid 1GUX) and how they change throughout the simulation, in terms of their stability and how they change as a function of the choice of x residues. We measured π-stacking between Tyr 23 and Tyr 25 as an average RMSD value of the sidechain heavy atoms truncated at Cζ relative to the crystal structure and the hydrogen bond between Asp 21 and Tyr 25 as the average distance between the Asp 21 sidechain carboxylate oxygens (Oε1 and Oε2) and Tyr 25 sidechain hydroxyl oxygen. The π-stacking RMSD calculation is truncated at Cζ atom, i.e. excluding the Tyr sidechain hydroxyl oxygen, since mutants containing Phe in place of Tyr do not feature this hydroxyl oxygen and the RMSD values for those mutants may be skewed. Results are shown in **Table 5.2** for both the bound and the free peptide in solution, in bulk water and with 200 mM NaCl.

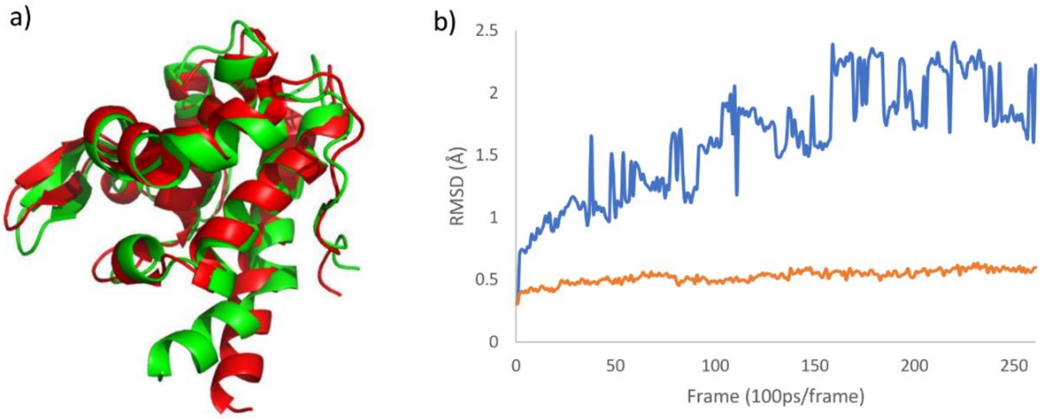


Figure 5.4: A) Rb structure from the final frame of the simulation of both restrained (red) and unrestrained (green) Rb-E7P complexes. B) RMSD values of Rb backbone atoms relative to the crystal structure with restrained backbone atoms shown in orange and unrestrained backbone atoms shown in blue

Table 5.2: Comparison of results of the wild-type LxCxE, in bulk water and with 200 mM NaCl, of interactions found in complex crystal structure, with standard deviation show in parentheses. Asp-Tyr Hydrogen bond is 3.81 Å and 4.18 Å respectively in the crystal structure

Bulk water	π-stacking (Å)	Asp-Tyr Hydrogen bond OE1 (Å)	Asp-Tyr Hydrogen bond OE2 (Å)	Salt bridge at 5 Å (count)
E7-free	2.15 (0.57)	11.71 (1.94)	11.54 (2.03)	
E7-bound	1.52 (0.11)	4.81 (1.76)	5.03 (1.68)	1.52 (0.80)
E7P free	1.93 (0.54)	10.57 (4.11)	10.89 (3.73)	
E7P bound	1.48 (0.07)	4.28 (1.45)	5.01 (1.65)	1.97 (1.17)
200 mM NaCl				
E7-free	2.31 (0.76)	10.53 (2.82)	11.08 (2.90)	
E7-bound	1.55 (0.12)	5.73 (2.02)	6.18 (2.12)	1.08 (1.04)
E7P free	2.10 (0.76)	7.48 (3.06)	7.99 (3.59)	
E7P bound	1.51 (0.09)	5.50 (1.67)	5.66 (2.05)	2.31 (1.23)

E7 peptide phosphorylated at Ser 31 and 32 (indicated here as E7P) is the naturally active, Rb binding form, thus we wanted to understand the role of phosphorylation on the peptide's structure and dynamics. As such we performed the same analysis of E7 as above on E7P. We also compared the number of salt bridges formed between the highly acidic C-terminal tail of the peptide and the alkaline surface patch on Rb, measured as the number of Lys or Arg residues within 5 Å of the carboxyl group oxygens of Asp Glu and the phosphate group oxygens of pSer between residues 30 to 40. In order to determine the conformational propensity of free and bound E7, we performed a per residue breakdown of the phi/psi backbone torsion angle values, see **Figure 5.5**. In E7 free we see that Glu 26 is helical in solution while it adopts a PPII conformation when bound to Rb. Phosphorylation at Ser 31 and 32 enhances the propensity to form PPII regions in the free peptide, but there are no conformational differences when the peptide is bound. As the binding assays are carried out in 200 mM salt, we sought to understand the effect of salt on the system, therefore we ran and analysed the simulations of all systems in of 200 mM NaCl. See **Table 5.2** and **Figure 5.5**.

E7-mutants and LxCxE-containing variants. In order to further understand the role of the variable residues on the structure and dynamics of the peptide, we analysed the E7 mutants and variant sequences indicated in **Table 5.1**. More specifically, we also looked at the average backbone RMSD values of residues 21 to 28, namely xLxCxExx, relative to the 1GUX crystal structure, the stacking between an aromatic residue in position 23 of E7 and Asn 757 residue on Rb, which forms a bidentate hydrogen bond with the E7 backbone, and finally the hydration of Asn 757, represented as the average number of water molecules within 3 Å of the Asn 757 sidechain heavy atoms. Results are shown in **Table 5.3** and **Table 5.4**. Where specific interactions are eliminated due to the mutation of one or both residues involved, no values are shown.

We also looked at the effect of a spacer after the LxCxE motif as there a hydrophobic pocket that Leu 28 sits into. We evaluated this via the number of salt bridges formed with the alkaline patch on the Rb surface to determine how well the acidic region of E7 overlaps with alkaline patch on Rb. And the average backbone RMSD values of the DLxCxEQA motif relative to the 1GUX crystal structure to determine how well bound to Rb the DLxCxEQA motif is. Results are shown in **Table 5.3 and Table 5.4.**

To understand the role of the LxCxE motif in other peptides we looked at peptides containing the motif from other known LxCxE containing proteins namely CyclinD, BRCA1, GATA1, FOG2 and E7 mutants based on GATA1 ($E7P_{gata}$) and FOG2 ($E7P_{fog}$). As few to none of the properties evaluated for the E7 mutants are not present in these sequences, different properties were measured, namely the Rg of the LxCxE-containing peptide when bound to Rb see **Figure 5.6**, and measuring the interactions between the variable residues.

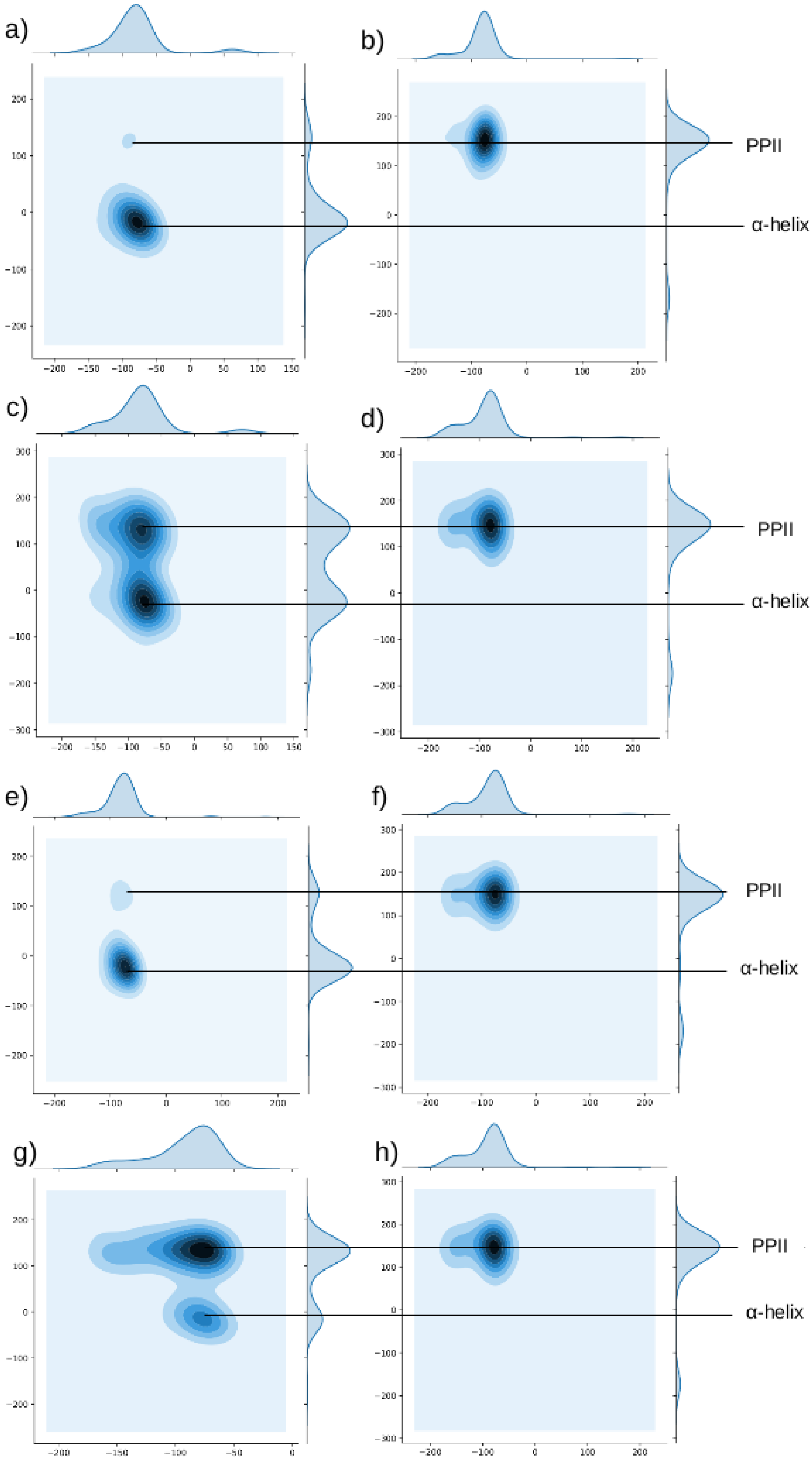


Figure 5.5: Ramachandran (phi/psi) plots showing dihedral angle values for Glu 26 in four different simulations with bulk water only, namely A) E7 unbound in solution B) E7 Rb-bound C) E7P unbound in solution D) E7P Rb-bound. And in 200 mM NaCl namely, E) E7 unbound in solution F) E7 Rb-bound, G) E7P unbound in solution, and H) E7P Rb-bound.

Table 5.3: Summary of results for E7 mutant peptides simulated in bulk water with error shown in parenthesis. Error is calculated as the standard deviation of the average value % π-stacking below a cuttoff of 1.7 Å and hydrogen bond below a cutoff of 5 Å are shown in brackets.

Bulk water **Sequence**	***π*-stacking (Å)**	**Asp-Tyr Hydrogen bond OE1 (Å)**	**Asp-Tyr Hydrogen bond OE2 (Å)**	**Aromatic–Asn 757 stacking**	**RMSD aa 21-28 (Å)**	**Hydration of Asn 757**	**Salt bridge at 5 Å (count)**
E7-free	2.15 (0.57) [10]	11.71 (1.94) [0]	11.54 (2.03)		3.11 (0.73)		
E7-bound	1.52 (0.11) [93]	4.81 (1.76) [73]	5.03 (1.68)	3.86 (0.45)	1.39 (0.36)	2.95 (1.30)	1.52 (0.80)
E7P free	1.93 (0.54) [15]	10.57 (4.11) [3]	10.89 (3.73)		2.66 (0.83)		
E7P bound	1.48 (0.07) [95]	4.28 (1.45) [74]	5.01 (1.65)	3.93 (0.35)	1.32 (0.34)	3.07 (1.10)	1.97 (1.17)
E7P$_{E26A}$ bound	1.51 (0.10) [92]	4.61 (0.86) [70]	4.61 (1.31)	4.33 (1.03)	1.60 (0.56)	3.20 (1.17)	2.12 (1.18)
E7P$_{D21A}$ free	2.53 (0.83) [0]				3.18 (0.71)		
E7P$_{D21A}$ bound	1.81 (0.15) [77]			5.89 (1.14)	1.72 (0.28)	3.79 (1.26)	2.44 (1.06)
E7P$_{AA}$ free					3.34 (0.59)		
E7P$_{AA}$ bound					1.69 (0.24)	4.34 (1.15)	2.06 (0.82)
E7$_{AY}$ free		13.10 (4.01) [0]	13.65 (4.12)		2.39 (0.88)		
E7$_{AY}$ bound		5.60 (1.84) [55]	5.83 (1.79)		1.56 (0.46)	4.26 (1.24)	1.61 (1.14)
E7P$_{AY}$ free		15.41 (3.74) [4]	15.24 (3.68)		2.42 (0.59)		
E7P$_{AY}$ bound		5.58 (1.61) [57]	5.75 (1.67)		1.61 (0.27)	4.23 (1.28)	1.68 (1.25)
E7$_{FF}$ free	1.90 (0.49) [17]				2.19 (0.65)		
E7$_{FF}$ bound	1.54 (0.10) [85]			5.09 (1.40)	1.60 (0.34)	3.44 (1.07)	2.20 (1.60)
E7P$_{FF}$ free	2.29 (0.56) [17]				2.36 (0.91)		
E7P$_{FF}$ bound	1.65 (0.30) [80]			5.06 (1.37)	1.66 (0.36)	3.68 (1.17)	2.15 (0.90)
E7P$_{FY}$ free	2.62 (0.19) [3]	11.09 (3.84) [2]	11.17 (3.83)		3.24 (0.67)		
E7P$_{FY}$ bound	1.79 (0.19) [75]	5.25 (1.69) [62]	5.79 (1.61)	5.66 (1.13)	1.67 (0.21)	3.01 (1.29)	1.67 (1.05)
E7P$_{YA}$ free					3.39 (0.70)		
E7P$_{YA}$ bound				5.68 (1.39)	1.73 (0.40)	3.90 (1.32)	1.92 (0.78)
E7P$_{YF}$ free	2.51 (1.02) [12]				2.93 (0.55)		
E7P$_{YF}$ bound	1.82 (0.23) [79]			5.54 (1.17)	1.65 (0.25)	3.46 (1.34)	2.05 (1.00)
E7$_{1A}$ free	2.36 (0.47) [9]	11.88 (3.39) [0]	11.84 (3.57)		2.61 (0.96)		
E7$_{1A}$ bound	1.48 (0.08) [95]	5.42 (1.86) [60]	5.62 (1.61)	3.95 (0.55)	2.18 (0.39)	2.87 (1.16)	1.27 (1.42)
E7$_{2A}$ free	2.14 (0.41) [6]	13.99 (3.82) [4]	13.99 (3.80)		2.37 (0.87)		
E7$_{2A}$ bound	1.51 (0.13) [87]	5.34 (1.86) [62]	5.95 (1.79)	4.11 (0.89)	1.68 (0.27)	3.05 (1.16)	2.05 (0.99)
E7$_{3A}$ free	2.17 (0.43) [7]	14.51 (4.48) [0]	14.12 (4.61)		2.93 (0.75)		
E7$_{3A}$ bound	1.53 (0.17) [91]	5 .75 (2.06) [64]	6.57 (2.85)	4.50 (1.13)	1.58 (0.27)	3.15 (1.11)	2.12 (1.04)
E7P$_{1A}$ free	2.43 (0.41) [6]	14.33 (4.31) [0]	14.74 (3.93)		2.28 (0.58)		
E7P$_{1A}$ bound	1.51 (0.07) [94]	5.48 (1.23) [53]	6.01 (1.88)	4.33 (1.06)	1.97 (0.48)	3.08 (1.20)	1.26 (0.91)
E7P$_{2A}$ free	2.31 (0.44) [6]	13.90 (3.16) [5]	14.26 (3.74)		2.63 (0.85)		
E7P$_{2A}$ bound	1.53 (0.15) [88]	5.31 (1.65) [58]	5.77 (2.03)	3.94 (0.74)	1.63 (0.25)	3.05 (1.15)	2.03 (1.12)
E7P$_{3A}$ free	2.20 (0.42) [95]	15.61 (3.62) [0]	15.44 (3.45)		2.54 (0.43)		
E7P$_{3A}$ bound	1.66 (0.21) [9]	5.71 (1.40) [60]	5.68 (1.68)	5.03 (1.34)	1.60 (0.30)	3.08 (1.14)	2.53 (1.12)

Table 5.4: Summary of results for E7 mutant peptides simulated with 200 mM NaCl. Error shown in parentheses. % π-stacking below a cuttoff of 1.7 Å and hydrogen bond below a cutoff of 5 Å are shown in brackets

200 mM NaCl Sequence	***π*-stacking (Å)**	**Asp-Tyr Hydrogen bond OE1 (Å)**	**Asp-Tyr Hydrogen bond OE2 (Å)**	**Aromatic–Asn stacking**	**RMSD of xLxCxExL (Å)**	**Hydration of Asn 757**	**Salt bridge at 5 Å (count)**
E7 free **E7 bound**	2.31 (0.76) [16] 1.55 (0.12) [92]	10.53 (2.82) [2] 5.73 (2.02) [51]	11.08 (2.90) 6.18 (2.12)	3.78 (0.27)	2.77 (0.55) 1.57 (0.41)	3.35 (1.20)	1.08 (1.04)
E7P free **E7P bound**	2.10 (0.76) [13] 1.51 (0.09) [95]	7.48 (3.06) [7] 5.50 (1.67) [58]	7.99 (3.59) 5.66 (2.05)	3.81 (033)	2.37 (0.92) 1.53 (0.39)	3.19 (1.13)	2.31 (1.23)
E7P$_{E26A}$ bound	1.57 (0.11) [85]	5.45 (1.86) [50]	5.64 (1.71)	4.45 (1.13)	1.60 (0.56)	3.40 (1.06)	2.15 (1.38)
E7P$_{D21A}$ free **E7P$_{D21A}$ bound**	2.35 (0.61) [0] 1.69 (0.20) [77]			5.04 (1.38)	2.76 (0.76) 1.66 (0.30)	3.84 (1.19)	2.37 (0.90)
E7P$_{AA}$ free **E7P$_{AA}$ bound**					3.00 (0.55) 1.69 (0.23)	3.86 (1.25)	1.71 (1.42)
E7$_{AY}$ free **E7$_{AY}$ bound**		13.37 (3.99) [2] 5.20 (2.07) [55]	13.54 (4.00) 5.22 (1.98)		2.56 (0.66) 1.57 (0.33)	4.26 (1.34)	2.23 (1.01)
E7P$_{AY}$ free **E7P$_{AY}$ bound**		15.20 (3.60) [52] 5.64 (1.52) [0]	16.36 (4.06) 6.10 (1.46)		2.47 (0.69) 1.60 (0.36)	4.06 (1.44)	1.72 (0.90)
E7$_{FF}$ free **E7$_{FF}$ bound**	2.42 (0.78) [8] 1.62 (0.24) [92]			5.62 (1.88)	2.61 (0.77) 1.67 (0.50)	3.55 (1.20)	2.09 (1.07)
E7P$_{FF}$ free **E7P$_{FF}$ bound**	1.87 (0.39) [17] 1.58 (0.11) [88]			5.35 (1.42)	2.69 (0.80) 1.75 (0.44)	3.66 (1.05)	1.96 (1.07)
E7P$_{FY}$ free **E7P$_{FY}$ bound**	2.17 (0.52) [3] 1.70 (0.22) [77]	9.48 (2.69) [0] 5.87 (1.57) [50]	10.15 (2.75) 5.92 (1.23)	4.86 (1.33)	3.10 (0.64) 1.71 (0.41)	3.46 (1.16)	2.08 (0.79)
E7P$_{YA}$ free **E7P$_{YA}$ bound**				5.33 (1.29)	3.36 (0.84) 1.70 (0.33)	4.18 (1.23)	1.67 (1.29)
E7P$_{YF}$ free **E7P$_{YF}$ free**	2.17 (0.48) [12] 1.63 (0.19) [79]			5.33 (1.26)	3.21 (0.74) 1.67 (0.40)	3.37 (1.27)	2.10 (1.09)
E7$_{1A}$ free **E7$_{1A}$ bound**	2.25 (0.36) [11] 1.54 (0.17) [85]	16.38 (3.64) [0] 5.47 (1.44) [55]	16.48 (3.79) 5.50 (1.55)	3.80 (1.14)	2.77 (0.64) 2.19 (0.54)	3.13 (1.11)	1.05 (0.98)
E7$_{2A}$ free **E7$_{2A}$ bound**	2.39 (0.52) [14] 1.51 (0.13) [89]	14.64 (2.87) [2] 5.72 (1.79) [49]	15.06 (2.81) 5.78 (1.47)	4.12 (0.88)	2.33 (1.23) 1.64 (0.24)	3.37 (1.12)	1.91 (0.98)
E7$_{3A}$ free **E7$_{3A}$ bound**	2.61 (0.58) [0] 1.58 (0.15) [84]	15.53 (4.38) [0] 5.60 (2.03) [52]	15.21 (4.41) 5.95 (1.70)	4.18 (0.89)	3.02 (0.71) 1.64 (0.42)	3.14 (1.48)	2.07 (1.54)
E7P$_{1A}$ free **E7P$_{1A}$ bound**	2.70 (0.52) [10] 1.56 (0.25) [87]	11.63 (2.42) [6] 5.50 (1.41) [49]	12.06 (2.20) 6.14 (1.51)	3.83 (0.41)	2.49 (0.75) 2.19 (0.40)	3.33 (1.17)	1.25 (0.75)
E7P$_{2A}$ free **E7P$_{2A}$ bound**	2.64 (0.75) [86] 1.52 (0.14) [6]	13.59 (3.17) [3] 5.53 (1.87) [54]	13.40 (3.40) 6.08 (1.86)	4.05 (0.92)	2.20 (0.68) 1.65 (0.35)	3.31 (1.14)	2.33 (0.97)
E7P$_{3A}$ free **E7P$_{3A}$ bound**	2.29 (0.31) [7] 1.66 (0.29) [88]	15.46 (4.38) [1] 5.81 (1.88) [53]	15.75 (3.99) 6.36 (1.84)	4.06 (1.13)	2.67 (0.62) 1.62 (0.30)	3.32 (1.19)	1.98 (0.98)

.

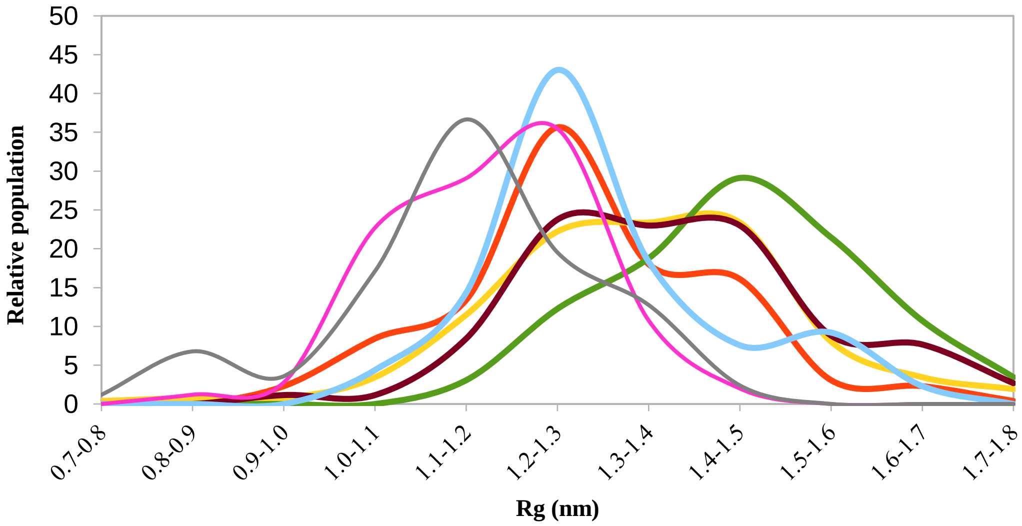


Figure 5.6: Histogram populations of peptide Rg for E7P (light blue) BRCA1 (magenta, $p < 0.01$), Cyclin D (grey $p < 0.01$), E7P$_{gata}$ (orange), GATA1 (yellow $p < 0.01$), E7P$_{fog}$ (green $p < 0.01$) and FOG2 (maroon $p < 0.01$). P-values calculated relative to E7P

The LxCxE found in Cyclin D is LLCCE with a Gln in position 21 in place of Asp found E7. The LxCxE motif is held in the bound conformation, when bound to Rb, by a hydrophobic interaction between these three residues, namely Gln 21 Leu 23 and Cys 25, for 100% of the simulation. Which we calculated as the amount of time in the simulation Leu 23 sidechain is within 4 Å of Cys 25 or Gln 23. As there is no acidic region in the Cyclin D peptide, it doesn't form any interaction with the basic patch of Rb and instead folds back upon itself. This is reflected in the fact that the Rg of the peptide is smaller than that of the E7 peptide bound. There is a small population with a Rg value closer to E7 which corresponds to residues 30-40 overlapping with the alkaline patch of Rb but these interactions are comprised hydrophobic interactions rather than electrostatic interactions. See **Figure 5.7**.

The LxCxE motif found in BRCA1 is LPCSE with Lys in position 21 instead of Asp. As Pro sidechain can't form interactions with other sidechains only Lys 21 and Ser 25 can interact to stabilise the motif in a bound conformation. However, there are no interactions between these two variable residues, with an average bond length of 7.34 ± 3.63 Å. The section containing residues 30 to 40 exhibits two conformations, one which overlaps the alkaline region of Rb, which is the same behaviour found in E7 and a second conformation in which this segment folds over onto itself which is the same as found in Cyclin D.

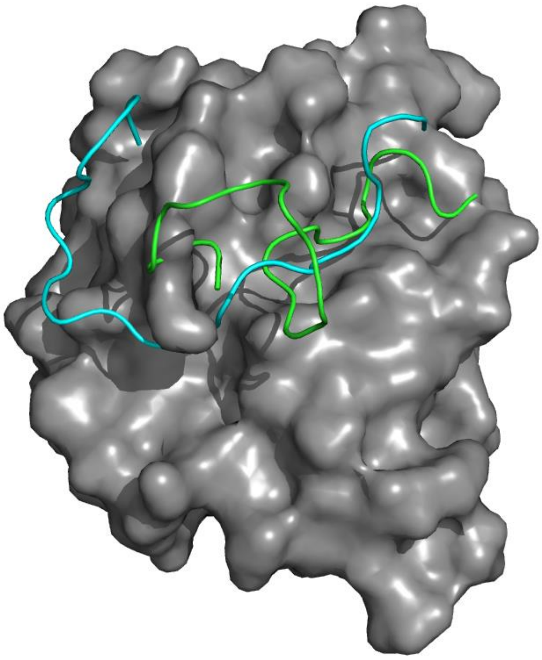


Figure 5.7: Rb B domain shown in grey with more compact Cyclin D peptide shown in green and less compact peptide shown in cyan.

The GATA LxCxE motif is LNCME with Leu in position 21 in place of Asp. The motif is held in a binding conformation when bound by hydrophobic interaction between Leu 21 and Met 25 which are within 4 Å of each other for 100% of the simulation. Two peaks are found in the Rg plot for GATA. The lower of these populations occurs when residues 30-40 are bound to Rb, not through ionic interactions with the alkaline patch but instead forming intermittent hydrophobic interactions with Rb Lys and Arg hydrocarbon sidechains. The higher population occurs when residues 30-40 do not interact at the with Rb alkaline patch, nor does the peptide fold back upon itself such as the case found in Cyclin D and BRCA1, see **Figure 5.8**

E7P$_{gata}$ LxCxE motif is held in a binding conformation when bound by hydrophobic interaction between Leu 23 and Met 25 which are within 4 Å of each other for 100% of the simulation as is the case for GATA1. Unlike GATA1, E7P$_{gata}$ has the highly acidic region of E7 and as such it overlaps with the alkaline region of Rb which is reflected in the Rg of E7P$_{gata}$ resembling that of E7. The FOG LxCxE motif is LSCEE with an Asn in position 21 instead of Asp. The LxCxE motif is held in a bound conformation when bound, by hydrogen bond between the Asn 21 and Glu 25 variable residues with average bond length values between Asn sidechain nitrogen and Glu oxygen OE1 and OE2 of 4.56 ± 2.01 and 4.88 ± 1.95 Å, respectively. Similar to GATA1, two Rg populations exist, one which is similar to the E7 peptide bound and another than is neither bound to Rb nor folds back upon itself. In this manner residues 30-40 behave similarly to the same residues in the GATA peptide.

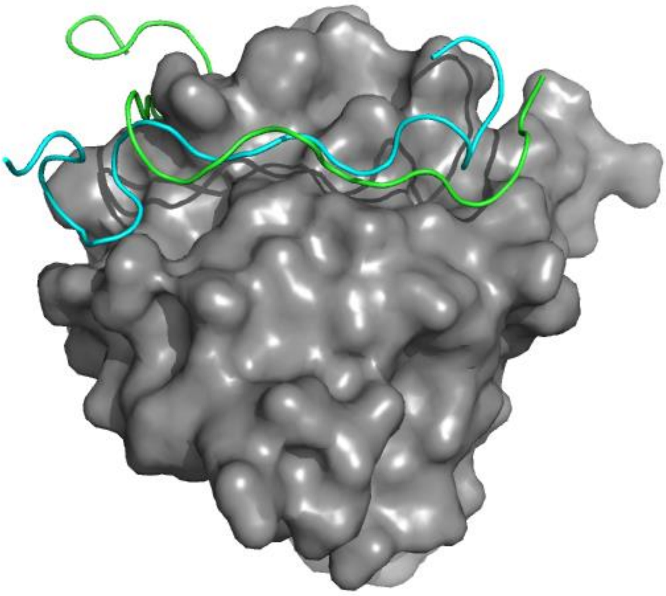


Figure 5.8: Rb B domain shown in grey with more compact GATA1 peptide shown in green and less compact peptide shown in cyan.

E7P$_{\mathrm{Fog}}$ LxCxE motif is held together by the same interactions as FOG but, unlike E7 and E7P$_{\mathrm{GATA,}}$ does not form any interactions between the acidic residues 30-40 and the alkaline patch of Rb this is reflected in the high Rg value for the peptide bound.

Experimental Results. In Prof Chemes' lab, the binding kinetics between Rb and HPV E7 was studied through fluorescence spectroscopy of FITC labelled peptides and results are shown in **Table 5.5**. These peptides correspond to E7 and the E7 derived mutants we simulated in this work. These peptides feature non-phosphorylated serines.

Table 5.5 Summary of experimental equilibrium and kinetic data from Prof Chemes' lab. k_{on} represents the association rate constant of the peptide and protein. k_{off} represents the dissociation rate constant of the complex

Peptide	**Sequence**	**K_D (nM)**	**k_{on} x 10^7 ($M^{-1}s^{-1}$)**	**k_{off} (s^{-1})**
E7	QPETTDLYCYEQLNDS	7	1.95	0.14
E7$_{D21A}$	QPETT**A**LYCYEQLNDS	80	1.71	1.37
E7$_{AY}$	QPETTDL**A**CYEQLNDS	60	1.82	1.10
E7$_{FY}$	QPETTDL**F**CYEQLNDS	10	2.24	0.22
E7$_{YA}$	QPETTDLYC**A**EQLNDS	139	1.90	2.64
E7$_{YF}$	QPETTDLYC**F**EQLNDS	51	1.69	0.86
E7$_{E26A}$	QPETTDLYCY**A**QLNDS	12477	0.30	37.76
E7$_{1A}$	QPETTDLYCYEQ**AL**ND	22	1.12	0.24
E7$_{2A}$	QPETTDLYCYEQ**AAL**N	120	0.45	0.53
E7$_{AA}$	QPETTDL**A**C**A**EQLNDS	860	-	17.71
E7$_{FF}$	QPETTDL**F**C**F**EQLNDS	57	3.70	2.10

5.5 Discussion

From the dynamics of the E7 peptide free in solution, and bound to Rb, we were looking for evidence of pre-structuring of the peptide in solution. By comparing the π-stacking of Tyr 23 and Tyr 25, the hydrogen bond between Asp 21 Tyr 25 and RMSD of residues 21 to 28 to the crystal structure we find no evidence of pre-structuring of the peptide. We can see this as the average hydrogen bond distance is significantly larger than a cut-off of 5 Å[22] and π-stacking RMSD is 0.63 Å higher for the free motif compared to the bound motif which represents a significant increase in the dynamics of both Tyr sidechains. Additionally, the solution RMSD is quite high, which shows the motif is quite flexible and not locked into the linear motif found in the crystal structure. This is consistent with experimental results that show changing the variable residues only affect the k_{off} and not the k_{on}, see **Experimental Results Table 5.5**, i.e. changing the variable residues doesn't affect the dynamics of the peptide in solution and only affects the peptide once it is bound to Rb.

As Ser 31 and Ser 32 are phosphorylated *in vivo*[12] and there is a higher population of PPII content in the peptide free in solution experimentally when phosphorylated[5,6], we calculated the same properties for E7P as for E7 and found there is little difference between E7 and E7P properties though the Asp 21 Tyr 25 Hydrogen bond is slightly longer for E7 free compared to E7P free but still significantly above 5 Å. We performed a per-residue breakdown of the dihedrals and found that there is a high degree of PPII content in the bound peptide of both E7 and E7P, most notably at Glu 26. Phosphorylation leads to an increase in PPII content of the free peptide, where Glu 26 goes from having α-helical dihedrals in E7 free to having PPII-like dihedrals in E7P free. Phosphorylation makes the solution peptide dihedrals more like that of the bound peptide, leading to pre-structuring of the E7P peptide not seen in the E7 peptide. We also expected that the phosphoserines would increase the strength of the interaction between the tail and the basic patch however this is not the case and both E7 and E7P form the same number of salt bridges. E7P bound has a higher average number of salt bridges than E7 the difference is within the error. Experimental results of the E7P peptide shows that only the k_{on} is affected by the phosphorylation of Ser 31 and 32. This increase in association rate agrees with our results that results that show phosphorylation leads to a prestructuring of the peptide.

As the experiments are performed in 200 mM salt, we investigated what effect if any, the increased ionic strength has on the system. What we find is that the increased ionic strength weakens the hydrogen bond between Asp 21 and Tyr 25 of E7 and E7P. This is caused by the screening of the negative charge of the Asp by the Na ions. The results for all the mutants agree closely with results for mutants simulated with bulk water only. Perhaps most surprisingly, the higher ionic strength has no effect on the formation of salt bridges between the acidic tail and the alkaline patch of Rb. What we expected to happen with the introduction of ions is that the association of the C-terminal tail and the basic patch on Rb would be disrupted due to the ionic screening effect. However, this is not the case and the same number of salt bridges are formed on average both with bulk water and with the 200 mM NaCl.

The results from E7 and E7P, bound and free, show that the variable residues do not play a role in pre-structuring the unbound peptide. Our results show that Y23A mutation causes an increase in the hydration of the Rb Asn 757. More specifically, Tyr 23 plays a role in protecting the hydrogen bonding with Asn 757 of Rb to the backbone of E7 from water by forming a stacking interaction with Asn 757, strengthening it. In the Y23F mutant the average number of water molecules around Asn 757 is the same as E7, although the distance between Phe 23 and Asn 757 is larger as shown in **Table 5.3** and **Table 5.4**. The reason for this is that the sidechain of Rb Met 761, which constitutes part of the hydrophobic pocket occupied by Leu 22, forms an intermittent hydrophobic interaction with the Phe sidechain in an orientation that does not allow the Phe sidechain to stack with Asn 757. While this interaction is possible in the wildtype sequence, it is not seen. The reason for this may be due to the higher hydrophilicity of the Tyr sidechain over the Phe sidechain disfavouring the hydrophobic interaction with Met 761. In the Y25A and Y23A Y25A double mutant we see the hydration of Asn 757 is increased, similar to Y23A mutant, which suggests the stacking interaction between Tyr 23 and Tyr 25 functions to stabilise the hydrogen bond between Asn 757 and E7. Both the Y25F mutant and Y23F Y25F double mutant show the same increase in distance between Phe 23 and Asn 757 but do not affect the hydration of Asn 757. Additionally, we expected the D21A mutant, which eliminates the Asp 21 Tyr 25 hydrogen bond while maintaining the aromaticity of residue 25 to behave like the Y25F mutant. As expected, our results show that the D21A and Y25F mutants behave the same. The simulation results of changing the “x” variable residues in the LxCxE motif are able to somewhat capture the experimental results, we can see the E7P$_{AA}$ double mutant is less stable due to the complete elimination of an aromatic residues. However, it is difficult to rank which mutant is less stable and which is more stable between E7$_{AY}$ and E7$_{YA}$. Similarly, for the E7P$_{FF}$ E7$_{FF}$ E7P$_{YF}$ E7P$_{FY}$ and E7P$_{D21A}$ mutants, the simulations are able to capture the destabilising effects of mutating the variable residues, however ranking the destabilising effects of the mutants is not possible.

The results of adding an Ala spacer between Gln 27 and Leu 28, shown in **Table 5.3** and **Table 5.4**, show that the variable residue interactions of the LxCxE motif are unaffected by the spacer, which we might have expected. We see that the introduction of a single alanine causes a dramatic increase in the RMSD of the motif backbone when bound to Rb which indicates that the peptide is more loosely bound than all other bound peptides simulated in this work. This is due to Leu 28 no longer being able to fit into the hydrophobic pocket and thus not being as tightly bound and as a result residues 30 to 40 have less optimal overlap with the alkaline patch resulting in lower salt bridge formation. Introducing a second and third Ala allows the two Ala residues to take the place of the leucine, with their Cβ atoms facing into the hydrophobic pocket. This can be seen by the RMSD values of E7$_{2A}$ and E7$_{3A}$ more closely matching that of the E7 bound value, and an increase in the number of salt bridges being formed by residues 30 to 40 as these residues can now properly overlap with the alkaline patch. Experimentally the introduction of a single Ala residue, namely in $E7_{1A}$ has a small effect on the energetics of binding of the peptide, however introduction of a second Ala residue in $E7_{2A}$ shows a much greater destabilising effect. The experimental results suggest that the binding groove is broad enough to accommodate a hydrophobic residue in position 28 or 29. A possible explanation for the disparity between experimental and simulation results is that the restraints on the Rb backbone atoms, which we showed are required to prevent Rb unfolding, could prevent this hydrophobic pocket which Leu 28 occupies in the crystal structure from being mobile enough to accommodate the hydrophobic residue when found in position 29.

Since Glu 26 is conserved in the sequence, we looked at the effects of mutating this residue to Ala in the $E7P_{E26A}$ mutant. We found that the Asp 21 Tyr 25 hydrogen bond and Tyr 23 Tyr 25 π-stacking are unaffected by the mutation. However, we found that the Rb Asn 757 is not as well protected from the solvent which can be seen in the longer Rb Asn 757 Tyr 23 distance value and increased hydration of Asn 757. We also see that the RMSD of the DLxCxAQL motif is higher than that of the E7 bound and E7P bound which indicated that the peptide is more loosely bound than that of either wildtype peptide. Additionally, this residue could play a role in recognition of the peptide by Rb. This mutant incurs a enthalpic penalties over E7 and E7P peptides where the loss of the bidentate hydrogen bond between Glu 26 and Rb leads to a loss of enthalpy of binding. Experimentally we see that the $E7P_{E26A}$ mutant has a destabilising effect on both the k_{on} and k_{off} values. The results of the simulation agree with these experimental results and the possibility of Glu 26 playing a role in recognition between Rb and E7 could account for the lower k_{on} value of the $E7P_{E26A}$ mutant.

From the R_g plots of the other LxCxE-containing sequences we can see that BRCA1 and Cyclin D have more compact conformations than the E7 mutants due to residues 30 to 40 not interacting strongly with the alkaline patch and instead folding back on itself. Conversely, the GATA1 and FOG2 peptides show partial overlap with the alkaline patch of Rb but also show a more expanded conformation when not interacting with Rb. The $E7P_{GATA}$ and $E7P_{FOG}$ residues 30 to 40 were expected to behave more like E7P than the GATA1 and FOG2 peptides respectively. We see that this is the case for E7P$_{\text{GATA}}$ as it has one major peak corresponding to the tail forming salt bridges with Rb which matches up with E7P. There is also one peak for E7P$_{\text{FOG}}$ however it is centred at ~2 Å higher than both E7P and E7Pgata and is close in value to the higher peaks of the GATA and FOG proteins. A reason that E7P$_{\text{FOG}}$ doesn't overlap with the basic patch is that the variable residue interaction introduced by from the FOG peptide is a tight hydrogen bond, namely between residues Asn 23 and Glu 25, and this pulls the LxCxE motif tighter together and prevents the tail from overlapping with the basic patch.

A possible explanation for why the tail of E7/E7P, and the mutants derived from them, overlap with the Rb basic patch while the other peptides do not is that, as previously mentioned, it is known that E7 binds to Rb in two binding pockets namely the LxCxE SLiM binding pocket found in the B domain of Rb and another found in the A-B domain interface. Having this acidic region on E7 may be required to guide the second binding region in CR3 region of E7 to its binding site at the Rb A-B region interface. This functionality may not be required for the other proteins to perform their roles as it possible they only bind Rb in the B-domain.

5.6 Conclusion

In this work we looked at the conformational propensities of a 24-residue peptide from E7, which contains the LxCxE motif, and mutants of this peptide in which the variable residues, flanking residues, or Glu 26 were modified. We also looked at the effects of phosphorylation of Ser 31 and 32 on the system, as well as the effects of an NaCl concentration of 200 mM. We also looked at other LxCxE containing peptides from BRCA1, Cyclin D, FOG2 and GATA1. We found that the flanking residues and variable residues are important for the binding of the E7 peptide, but only affect the bound peptide and have no influence on the peptide free in solution. The possible role of residues 30 to 40 of E7 is to guide the CR3 region of E7 to the C domain of Rb and thus lead to displacement of E2F from Rb, the guiding of the CR3 region is accomplished through the formation of salt bridges by residues 30 to 40 with the Rb alkaline patch. This interaction may not be necessary in the other LxCxE containing peptides as it is possible they only bind Rb in the B-domain, as a result their equivalent residues 30 to 40 are more alkaline or hydrophobic. Phosphorylation can be seen to pre-structure the peptide by changing the secondary structure propensity of some residues, when the peptide is free in solution, to more closely resemble the higher PPII content found in the bound peptide. We found that 200 mM NaCl has little effect on the system and only slightly weakens a single interaction found in the wildtype peptide, namely the Asp 21 Tyr 25 hydrogen bond. Our simulation results agreed quite well with the experimental results however, ranking the peptides, which contain similar mutations, in terms of their destabilising effects wasn't possible. The simulations also reproduced the destabilising effects of Ala mutation of Glu 26. However, the simulations failed to reproduce the properties of the E7$_{1A}$ and E7$_{2A}$ mutants, which is possibly due to the requirement of restraining the Rb backbone which prevents Rb from accommodating a hydrophobic residue in position 29, instead of position 28 as found in the crystal structure.

References

1. Hausen, H. zur. Papillomavirus infections — a major cause of human cancers. *Biochim. Biophys. Acta - Rev. Cancer* **1288**, F55–F78 (1996).
2. Munger, K. *et al.* Complex formation of human papillomavirus E7 proteins with the retinoblastoma tumor suppressor gene product. *EMBO J.* **8**, 4099–4105 (1989).
3. Liu, X., Clements, A., Zhao, K. & Marmorstein, R. Structure of the human Papillomavirus E7 oncoprotein and its mechanism for inactivation of the retinoblastoma tumor suppressor. *J. Biol. Chem.* **281**, 578–586 (2006).
4. Ohlenschläger, O. *et al.* Solution structure of the partially folded high-risk human papilloma virus 45 oncoprotein E7. *Oncogene* **25**, 5953–5959 (2006).
5. García-Alai, M. M., Alonso, L. G. & de Prat-Gay, G. The N-Terminal Module of HPV16 E7 Is an Intrinsically Disordered Domain That Confers Conformational and Recognition Plasticity to the Oncoprotein. *Biochemistry* **46**, 10405–10412 (2007).
6. Chemes, L. B., Sánchez, I. E., Smal, C. & de Prat-Gay, G. Targeting mechanism of the retinoblastoma tumor suppressor by a prototypical viral oncoprotein. *FEBS J.* **277**, 973–988 (2010).
7. Xiao, B. *et al.* Crystal structure of the retinoblastoma tumor suppressor protein bound to E2F and the molecular basis of its regulation. *Proc. Natl. Acad. Sci.* **100**, 2363–2368 (2003).
8. Rubin, S. M., Gall, A.-L., Zheng, N. & Pavletich, N. P. Structure of the Rb C-Terminal Domain Bound to E2F1-DP1: A Mechanism for Phosphorylation-Induced E2F Release. *Cell* **123**, 1093–1106 (2005).
9. Patrick, D. R., Oliff, A. & Heimbrook, D. C. Identification of a novel retinoblastoma gene product binding site on human papillomavirus type 16 E7 protein. *J. Biol. Chem.* **269**, 6842–6850 (1994).
10. Shan, B., Durfee, T. & Lee, W. H. Disruption of RB/E2F-1 interaction by single

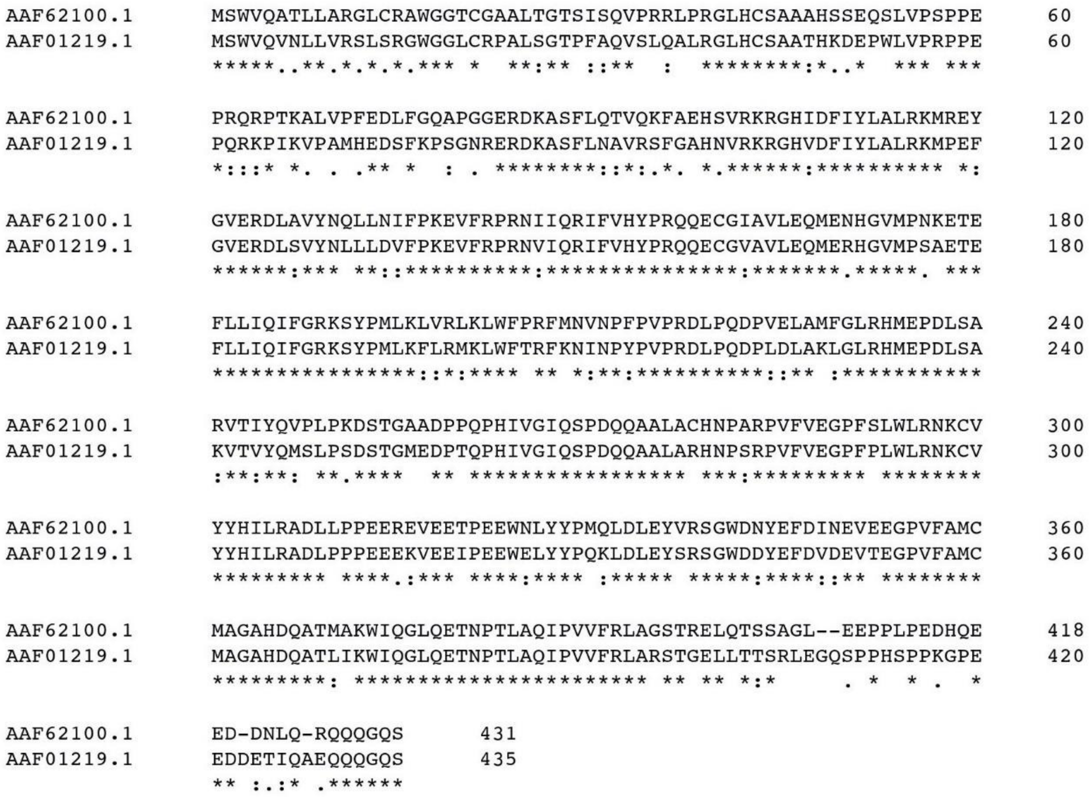
point mutations in E2F-1 enhances S-phase entry and apoptosis. *Proc. Natl. Acad. Sci.* **93**, 679–684 (1996).

11. Helin, K. *et al.* A cDNA encoding a pRB-binding protein with properties of the transcription factor E2F. *Cell* **70**, 337–350 (1992).
12. Barbosa, M. S. *et al.* The region of the HPV E7 oncoprotein homologous to adenovirus E1a and Sv40 large T antigen contains separate domains for Rb binding and casein kinase II phosphorylation. *EMBO J.* **9**, 153–160 (1990).
13. Sugita, Y. & Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **314**, 141–151 (1999).
14. Patriksson, A. & van der Spoel, D. A temperature predictor for parallel tempering simulations. *Phys. Chem. Chem. Phys.* **10**, 2073–2077 (2008).
15. Piana, S., Lindorff-Larsen, K. & Shaw, D. E. How robust are protein folding simulations with respect to force field parameterization? *Biophys. J.* **100**, L47–L49 (2011).
16. Piana, S., Donchev, A. G., Robustelli, P. & Shaw, D. E. Water Dispersion Interactions Strongly Influence Simulated Structural Properties of Disordered Protein States. *J. Phys. Chem. B* **119**, 5113–5123 (2015).
17. Rauscher, S. *et al.* Structural Ensembles of Intrinsically Disordered Proteins Depend Strongly on Force Field: A Comparison to Experiment. *J. Chem. Theory Comput.* **11**, 5513–5524 (2015).
18. Abraham, M. J. *et al.* GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1–2**, 19–25 (2015).
19. Maestro, version 9.2. Schrödinger, LLC; New York, NY, U. 2011. .
20. The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, L. .
21. Robustelli, P., Piana, S. & Shaw, D. E. Developing a molecular dynamics force field for both folded and disordered protein states. *Proc. Natl. Acad. Sci.* **115**, E4758–E4766 (2018).
22. Ahmed, M. C., Papaleo, E. & Lindorff-Larsen, K. How well do force fields capture the strength of salt bridges in proteins? *PeerJ* **6**, e4967–e4967 (2018).

Chapter 6: Conformational analysis of the intrinsic disorder in human and murine ECSIT C-terminal tails

6.1 Introduction

ECSIT isoform I is a 431-residue, 50 kDa protein involved in many different biochemical and metabolic pathways, including immune system activation and homeostasis[1]. The human and murine ECSIT sequences are shown in an alignment in **Figure 6.1**.



```
AAF62100.1      MSWVQATLLARGLCRAWGGTCGAALTGTSISQVPRRLPRGLHCSAAAHSSEQSLVPSPPE     60
AAF01219.1      MSWVQVNLLVRSLSRGWGGLCRPALSGTPFAQVSLQALRGLHCSAATHKDEPWLVPRPPE     60
                *****..**.*.*.*.*** *  **:** ::**  :  ********:*..*  *** ***

AAF62100.1      PRQRPTKALVPFEDLFGQAPGGERDKASFLQTVQKFAEHSVRKRGHIDFIYLALRKMREY     120
AAF01219.1      PQRKPIKVPAMHEDSFKPSGNRERDKASFLNAVRSFGAHNVRKRGHVDFIYLALRKMPEF     120
                *:::* *. . .** *  : . ********::*:.*. *.******:********** *:

AAF62100.1      GVERDLAVYNQLLNIFPKEVFRPRNIIQRIFVHYPRQQECGIAVLEQMENHGVMPNKETE     180
AAF01219.1      GVERDLSVYNLLLDVFPKEVFRPRNVIQRIFVHYPRQQECGVAVLEQMERHGVMPSAETE     180
                ******:*** **::**********:***************:*******.*****. ***

AAF62100.1      FLLIQIFGRKSYPMLKLVRLKLWFPRFMNVNPFPVPRDLPQDPVELAMFGLRHMEPDLSA     240
AAF01219.1      FLLIQIFGRKSYPMLKFLRMKLWFTRFKNINPYPVPRDLPQDPLDLAKLGLRHMEPDLSA     240
                ****************::*:**** ** *:**:**********::** :***********

AAF62100.1      RVTIYQVPLPKDSTGAADPPQPHIVGIQSPDQQAALACHNPARPVFVEGPFSLWLRNKCV     300
AAF01219.1      KVTVYQMSLPSDSTGMEDPTQPHIVGIQSPDQQAALARHNPSRPVFVEGPFPLWLRNKCV     300
                :**:**: **.****  ** ***************** ***:********* ********

AAF62100.1      YYHILRADLLPPEEREVEETPEEWNLYYPMQLDLEYVRSGWDNYEFDINEVEEGPVFAMC     360
AAF01219.1      YYHILRADLPPPEEEKVEEIPEEWELYYPQKLDLEYSRSGWDDYEFDVDEVTEGPVFAMC     360
                ********* ****.:*** ****:**** :***** *****:****::** ********

AAF62100.1      MAGAHDQATMAKWIQGLQETNPTLAQIPVVFRLAGSTRELQTSSAGL--EEPPLPEDHQE     418
AAF01219.1      MAGAHDQATLIKWIQGLQETNPTLAQIPVVFRLARSTGELLTTSRLEGQSPPHSPPKGPE     420
                *********: *********************** ** ** *:*      . *  * .  *

AAF62100.1      ED-DNLQ-RQQQGQS      431
AAF01219.1      EDDETIQAEQQQGQS      435
                ** :.:* .******
```

Figure 6.1 Clustal omega sequence alignment of human ECSIT and murine ECSIT from M. musculus

As one of its fundamental roles, ECSIT is a key player in the assembly of the mitochondrial respiratory Complex I through interactions with NDUFAF1 and ACAD9 Complex I assembly proteins[2,3]. Knockout of ECSIT leads to lethality in the embryo stage of mice[4], but knockout of ECSIT in mature macrophages is consistent with dysfunction of complex I[5]. No structural information is available on the ECSIT protein and very few details are known on the specific roles that its different domains play in the assembly of Complex I. Nevertheless, it has been determined that the N-terminal region, and more specifically the segment between residues 1 and 48, targets ECSIT for mitochondrial localisation[3]. This region is then cleaved off after mitochondrial localisation to produce 45 kDa mitochondrial ECSIT[3].

As part of a collaboration with Prof Paul N. Moynagh in the Department of Biology at Maynooth University, we have embarked in a project aimed at determining the key role of the ECSIT protein in the assembly of the respiratory complex I, and more specifically in characterising the domain responsible for this function and how mutations of its sequence affect it. Within this framework, experimental studies based on expression levels conducted in Prof Moynagh's lab, show that human ECSIT (hECSIT) is significantly less stable than murine ECSIT (mECSIT). Interestingly, a swap of the terminal 34 amino acid residues between human and murine ECSIT, regions we will name here hECSIT$_{398-431}$ and mECSIT$_{402-435}$, respectively determines an inversion of the relative stabilities of the two proteins, resulting in a mutated mECSIT which is much less stable than the hECSIT see **Figure 6.2**[6].

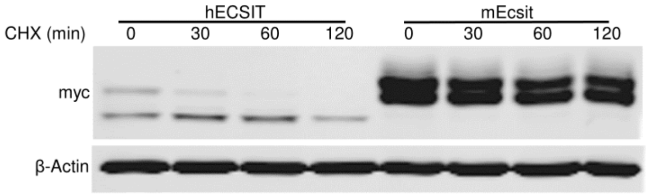


Figure 6.2: Cell lysates of HEK293T were generated at indicated times showing relative stabilities of hECSIT and mECSIT through western blot.

Furthermore, *in vivo* experiments in mice expressing humanised ECSIT show the development of a rather enlarged heart, a clear result of heart congestion due to a less efficient ATP production see **Figure 6.3**[6].

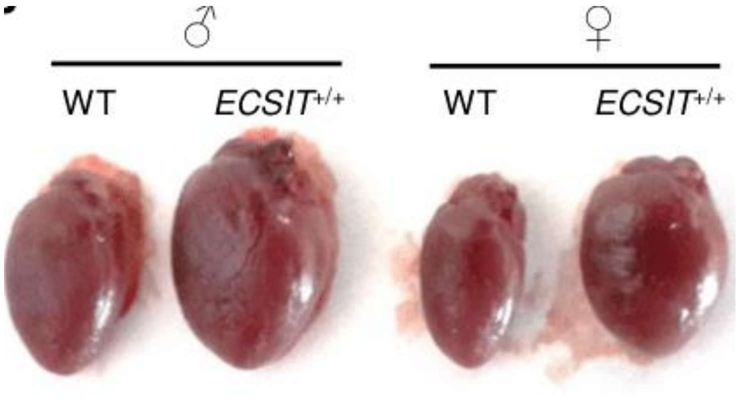


Figure 6.3: Representative images of male and female hearts of mice at 7-months of age showing enlargement of hearts in mice which express mECSIT with the C-terminal tail from hECSIT denoted as ECSIT$^{+/+}$.

At first glance the sequence of the extreme C-terminal regions of both human and murine ECSITs seem poorly structured, if not completely disordered. Sequence analysis done with the disorder prediction software disEMBL[7] indeed shows that the C-terminal tails of both hECSIT and mECSIT are highlighted to be potential IDP regions, see **Figure 6.5**.

Table 6.1: Wild-type human (hECSIT) and murine ECSIT (mECSIT) peptides and hybrid peptides (hECSITm and mECSITh) with the 12 residue regions that was swapped highlighted in red. Residue numbers for the residues within the whole ECSIT protein are shown on the left- and right-hand side of the sequence, with the re-numbering scheme that we used to identify the hybrid peptides indicated in parenthesis.

hECSIT:	398(1)	RELQTTSAGLEEPPLPEDHQEEDDNLQRQQQGQS	431(34)
mECSIT:	402	TTSRLEGQSPPHSPPKGPEEDDETIQAEQQQGQS	435
hECSITm:		RELQTTSAGLEEPPLPEDHQEEDDNLQRQQQGQS	
mECSITh:		TTSRLEGQSPPHSPPKGPEEDDETIQAEQQQGQS	

Here we used enhanced sampling molecular simulations via temperature REMD to gain insight into the structure and dynamics of the ECSIT C-terminal tails to determine the relative degrees of disorder and potential residual secondary structure in 32 residues peptides derived from the human and murine ECSIT, see **Error! Reference source not found.**. The simulation results support the bioinformatics prediction that the two tails are intrinsically disordered. Furthermore, our simulations provide the additional insight that hECSIT and mECSIT have quite different secondary structure propensities. Based on this information we proposed very specific mutations that were later on proven experimentally to invert the relative stability of mECSIT and hECSIT, very much like the whole C-terminal region swap. In the following sections we will discuss the simulations result within the context of their role in the ECSIT protein stability and function.

6.2 Computational method

DisEMBL 1.5[7] was used to predict the structural disorder of the hECSIT and mECSIT proteins with default parameters and the sequence for human and mouse as found on the NCBI website. DisEMBL uses three definitions of disorder, namely *loops/coils*, *hot loops*, *Remark-465*. *Loops/coils* correspond to any secondary structure that is not an α-helix, 3_{10} helix or β-strand according to DSSP[8]. The *hot loops* method refines the *loops/coils* method to consider the mobility of the residue determined by B-factor of the α-carbon[9]. The *Remark-465* definition is based on coordinates missing in the PDB[10]. PSIPRED 4.0[11,12] was used to predict the structure of the 34-residue hECSIT and mECSIT peptides.

All REMD simulations were carried out with 77 replicas to span temperatures in the range 300 K to 500 K, based on the *Temperature Generator for REMD* online tool (http://folding.bmc.uu.se/remd/)[13], with two different forcefield setups, namely a99SB-disp[14] with the TIP4P-D[15] water model and CHARMM36m with the TIP3P for CHARMM water model[16]. All calculations were run with GROMACS v. 2018.3 s[17]. Computational resources were provided by the Irish Centre for High-End Computing (ICHEC). All calculations were run on the ICHEC supercomputer "*Kay*" on 2 x 20 2.4 GHz Intel Xeon Gold 6148 (Skylake) processors, 192 GiB of RAM, a 400 GiB local SSD for scratch space and a 100 Gbit OmniPath network adaptor. Energy minimizations were carried out on 2 nodes, i.e. 80 cores. NVT equilibrations and REMD production simulations were run on 2 processors per replica for 77 replicas, i.e. 154 processors over 4 nodes. We estimated that the final cost for all MD simulations in this project reached approximately 175,000 CPU hours.

The human ECSIT (hECSIT) and murine ECSIT (mECSIT) sequences correspond to the 34 terminal amino acids of their respective protein C-terminal regions. Both mECSIT and hECSIT were built in the fully extended conformation using the academic version of Schrödinger MAESTRO software[18]. The two hybrid peptides were generated via point mutations with *pymol*[19]. The N-termini of all peptides were capped with an acetyl group (ACE) and the C-termini were left uncapped because they are the protein's termini. A 500k steps of steepest descent minimization was carried out in implicit solvent, followed by a 100 ns NPT MD run and a random structure was selected as the starting conformation for REMD. This structure was placed in a rhombic dodecahedron simulation box with a minimum distance between the peptide and the edge of the box of 1.2 nm. Counterions were added to neutralise the system and to bring the ionic strength up to 200 mM. Bond lengths were constrained using LINCS and the long-range electrostatic interactions were treated with Particle Mesh Ewald (PME) with a switch from real space to reciprocal space at 1.2 nm. The van der Waals interactions were cut off at 1.2 nm.

The REMD simulation protocol involved an initial energy minimization, namely 500k of steepest descent with a force-based convergence threshold of 100 kJmol^{-1}nm^{-1}, of the ions, water and hydrogen atoms positions with the protein heavy atoms restrained. After the energy minimization, we carried out a 500 ps equilibration in the NVT ensemble with the same set of restraints, where each replica was equilibrated to its specific temperature. To ensure that the system maintained a pressure of 1 bar, the density of the system at 300 K was monitored throughout the simulation and ensured that it matched that of water at 300 K and 1 bar, so that no further equilibration was required. To integrate the equations of motion we used a leap-frog stochastic dynamics (sd) integrator, with a friction coefficient corresponding to the inverse of tau-t equal to 0.1 ps, where tau-t is the time constant for coupling. Production stage followed with all atoms unrestrained. The system attempted to exchange replicas every 500 steps, namely every ps, and the production was extended to 26 ns per replica, for a total cumulative simulation time of 2.0 μs for each peptide, for each force field set-up. The trajectory corresponding to the replica at 300K was analysed in terms of their radius of gyration and secondary structure propensity. The clustering analysis was based on the *Gromos* method[20] with an RMSD cut-off of 0.75 nm, chosen within a range between 0.4 nm and 0.75 nm, as it allowed to obtain the highest number of clusters while avoiding redundancy. The cluster's secondary structure was determined using STRIDE[21]. STRIDE requires two consecutive hydrogen bonds between residue pairs [k,k+4] and [k+1,k+5] and residues k+1 to k+4 are labelled as helical[21]. Since we used a large cut-off in the clustering procedure, the helicity might not be shown in the structure selected to represent the cluster. As such when STRIDE recognised a residue as a turn (T) a visual inspection and phi/psi angle calculation was performed to confirm the presence or absence of helicity. The data convergence was assessed by means of average RMSD values correlation function[22]. Based on the plots shown in **Figure 6.4,** the simulations can be considered as converged after approximately 15 ns per replica.

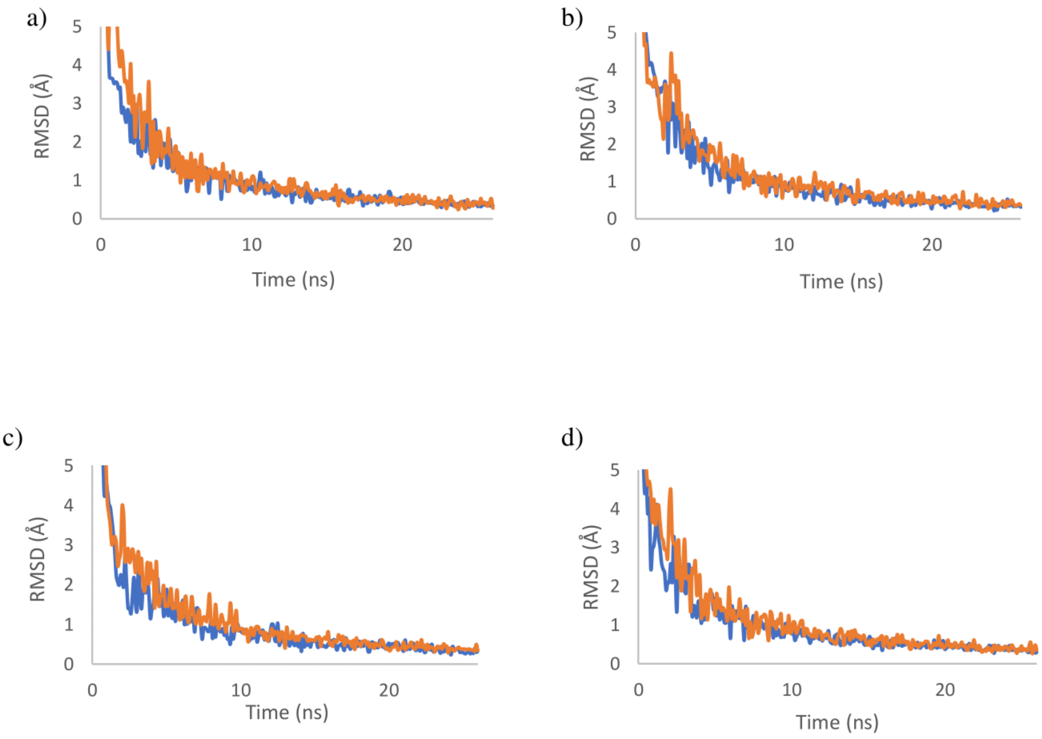
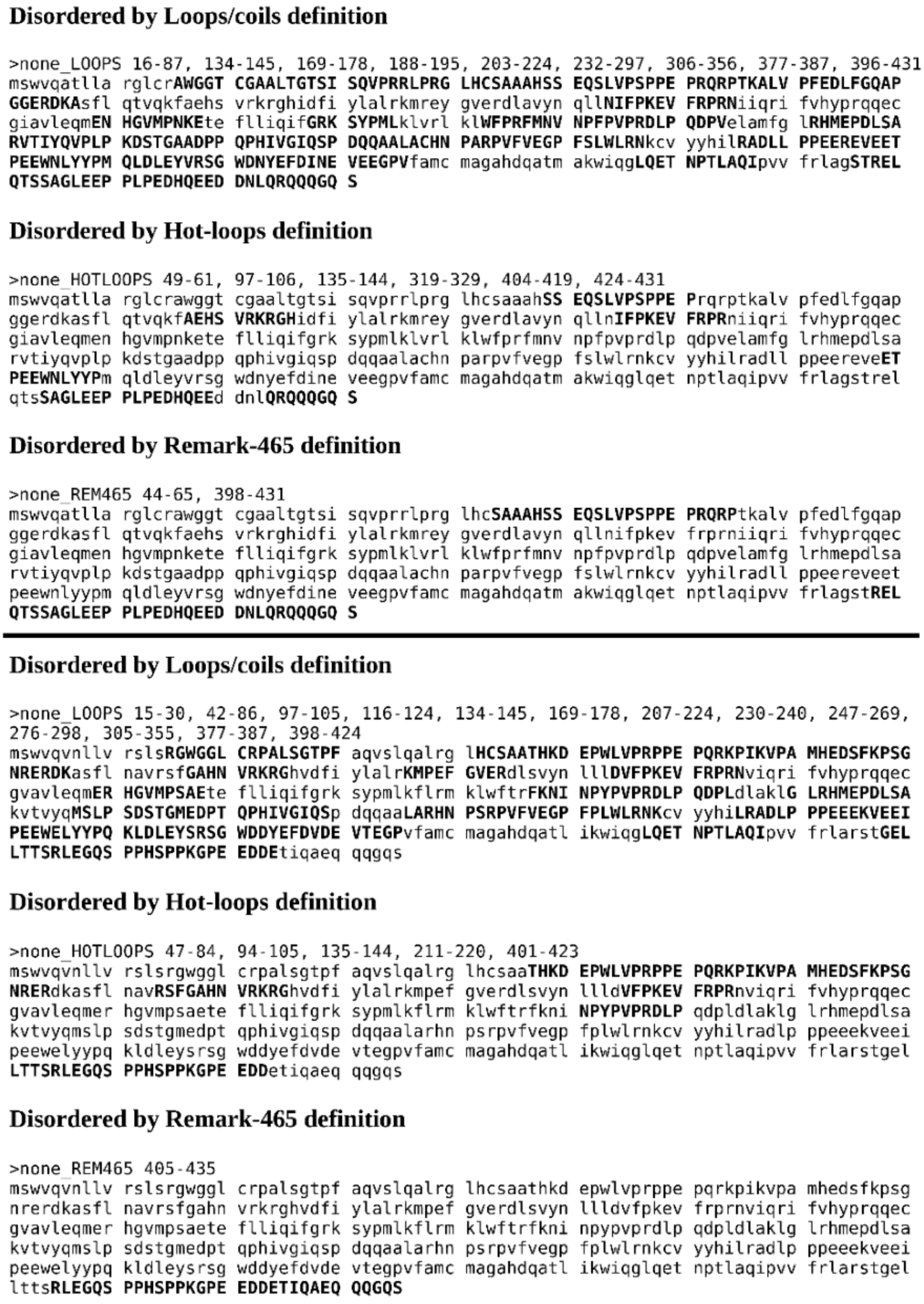


Figure 6.4: Data convergence assessed by average RMSD values correlation functions. The plots show data convergence for the simulations of a) hECSIT, b) mECSIT c) hECSITm and d) mECSITh. Data relative to the simulations with a99SB-disp/TIP4P-D are shown in blue and with CHARMM36m/TIP3P are shown in orange.

6.3 Results

The ECSIT sequence was analysed with DisEMBL[7] to assess the intrinsic propensity for conformational disorder. As shown in see **Figure 6.5**, the software identified a region in the ECSIT extreme C-terminal domain highly likely to be structurally disordered. More specifically, results obtained with the *hot-loops* and *Remark-465* definitions suggest that the last 34 and 31 residues of the human and the murine ECSIT C-terminal tail, respectively, are disordered. The hECSIT and mECSIT peptides in this study were built based on these on this information, with sequences corresponding to the to the last 34 residues of the human and murine ECSIT C-terminal tails.

In order to get further insight, we also analysed the mECSIT and hECSIT peptides sequence through PSIPRED[11] (http://bioinf.cs.ucl.ac.uk/psipred/) to check for any secondary structure propensity. The results shown in **Figure 6.6**, indicate that the peptides C-terminal region is expected to have a propensity to form helical motifs. As shown in **Figure 6.7,** the REMD results indicate that the two 34-residue hECSIT and mECSIT peptides have similar dimensions in terms of radius of gyration (Rg), with Rg values differences between the a99SB-disp and CHARMM36m force fields of 0.1 nm, a value within one standard deviation. Namely, hECSIT has average Rg



Disordered by Loops/coils definition

```
>none_LOOPS 16-87, 134-145, 169-178, 188-195, 203-224, 232-297, 306-356, 377-387, 396-431
mswvqatlla rglcrAWGGT CGAALTGTSI SQVPRRLPRG LHCSAAAHSS EQSLVPSPPE PRQRPTKALV PFEDLFGQAP
GGERDKAsfl qtvqkfaehs vrkrghidfi ylalrkmrey gverdlavyn qllNIFPKEV FRPRNiiqri fvhyprqqec
giavleqmEN HGVMPNKEte flliqifGRK SYPMLklvrl klWFPRFMNV NPFPVPRDLP QDPVelamfg lRHMEPDLSA
RVTIYQVPLP KDSTGAADPP QPHIVGIQSP DQQAALACHN PARPVFVEGP FSLWLRNkcv yyhilRADLL PPEEREVEET
PEEWNLYYPM QLDLEYVRSG WDNYEFDINE VEEGPVfamc magahdqatm akwiqgLQET NPTLAQIpvv frlagSTREL
QTSSAGLEEP PLPEDHQEED DNLQRQQQGQ S
```

Disordered by Hot-loops definition

```
>none_HOTLOOPS 49-61, 97-106, 135-144, 319-329, 404-419, 424-431
mswvqatlla rglcrawggt cgaaltgtsi sqvprrlprg lhcsaaahSS EQSLVPSPPE Prqrptkalv pfedlfgqap
ggerdkasfl qtvqkfAEHS VRKRGHidfi ylalrkmrey gverdlavyn qllnIFPKEV FRPRniiqri fvhyprqqec
giavleqmen hgvmpnkete flliqifgrk sypmlklvrl klwfprfmnv npfpvprdlp qdpvelamfg lrhmepdlsa
rvtiyqvplp kdstgaadpp qphivgiqsp dqqaalachn parpvfvegp fslwlrnkcv yyhilradll ppeereveET
PEEWNLYYPm qldleyvrsg wdnyefdine veegpvfamc magahdqatm akwiqglqet nptlaqipvv frlagstrel
qtsSAGLEEP PLPEDHQEEd dnlQRQQQGQ S
```

Disordered by Remark-465 definition

```
>none_REM465 44-65, 398-431
mswvqatlla rglcrawggt cgaaltgtsi sqvprrlprg lhcSAAAHSS EQSLVPSPPE PRQRPtkalv pfedlfgqap
ggerdkasfl qtvqkfaehs vrkrghidfi ylalrkmrey gverdlavyn qllnifpkev frprniiqri fvhyprqqec
giavleqmen hgvmpnkete flliqifgrk sypmlklvrl klwfprfmnv npfpvprdlp qdpvelamfg lrhmepdlsa
rvtiyqvplp kdstgaadpp qphivgiqsp dqqaalachn parpvfvegp fslwlrnkcv yyhilradll ppeereveet
peewnlyypm qldleyvrsg wdnyefdine veegpvfamc magahdqatm akwiqglqet nptlaqipvv frlagstREL
QTSSAGLEEP PLPEDHQEED DNLQRQQQGQ S
```

Disordered by Loops/coils definition

```
>none_LOOPS 15-30, 42-86, 97-105, 116-124, 134-145, 169-178, 207-224, 230-240, 247-269,
276-298, 305-355, 377-387, 398-424
mswvqvnllv rslsRGWGGL CRPALSGTPF aqvslqalrg lHCSAATHKD EPWLVPRPPE PQRKPIKVPA MHEDSFKPSG
NRERDKasfl navrsfGAHN VRKRGhvdfi ylalrKMPEF GVERdlsvyn lllDVFPKEV FRPRNviqri fvhyprqqec
gvavleqmER HGVMPSAEte flliqifgrk sypmlkflrm klwftrFKNI NPYPVPRDLP QDPLdlaklG LRHMEPDLSA
kvtvyqMSLP SDSTGMEDPT QPHIVGIQSp dqqaaLARHN PSRPVFVEGP FPLWLRNKcv yyhiLRADLP PPEEEKVEEI
PEEWELYYPQ KLDLEYSRSG WDDYEFDVDE VTEGPvfamc magahdqatl ikwiqgLQET NPTLAQIpvv frlarstGEL
LTTSRLEGQS PPHSPPKGPE EDDEtiqaeq qqgqs
```

Disordered by Hot-loops definition

```
>none_HOTLOOPS 47-84, 94-105, 135-144, 211-220, 401-423
mswvqvnllv rslsrgwggl crpalsgtpf aqvslqalrg lhcsaaTHKD EPWLVPRPPE PQRKPIKVPA MHEDSFKPSG
NRERdkasfl navRSFGAHN VRKRGhvdfi ylalrkmpef gverdlsvyn llldVFPKEV FRPRnviqri fvhyprqqec
gvavleqmer hgvmpsaete flliqifgrk sypmlkflrm klwftrfkni NPYPVPRDLP qdpldlaklg lrhmepdlsa
kvtvyqmslp sdstgmedpt qphivgiqsp dqqaalarhn psrpvfvegp fplwlrnkcv yyhilradlp ppeeekveei
peewelyypq kldleysrsg wddyefdvde vtegpvfamc magahdqatl ikwiqglqet nptlaqipvv frlarstgel
LTTSRLEGQS PPHSPPKGPE EDDetiqaeq qqgqs
```

Disordered by Remark-465 definition

```
>none_REM465 405-435
mswvqvnllv rslsrgwggl crpalsgtpf aqvslqalrg lhcsaathkd epwlvprppe pqrkpikvpa mhedsfkpsg
nrerdkasfl navrsfgahn vrkrghvdfi ylalrkmpef gverdlsvyn llldvfpkev frprnviqri fvhyprqqec
gvavleqmer hgvmpsaete flliqifgrk sypmlkflrm klwftrfkni npypvprdlp qdpldlaklg lrhmepdlsa
kvtvyqmslp sdstgmedpt qphivgiqsp dqqaalarhn psrpvfvegp fplwlrnkcv yyhilradlp ppeeekveei
peewelyypq kldleysrsg wddyefdvde vtegpvfamc magahdqatl ikwiqglqet nptlaqipvv frlarstgel
lttsRLEGQS PPHSPPKGPE EDDETIQAEQ QQGQS
```

Figure 6.5: DisEMBL disorder prediction for hECSIT (top) and mECSIT (bottom) with disordered regions shown in bolded capital letters.

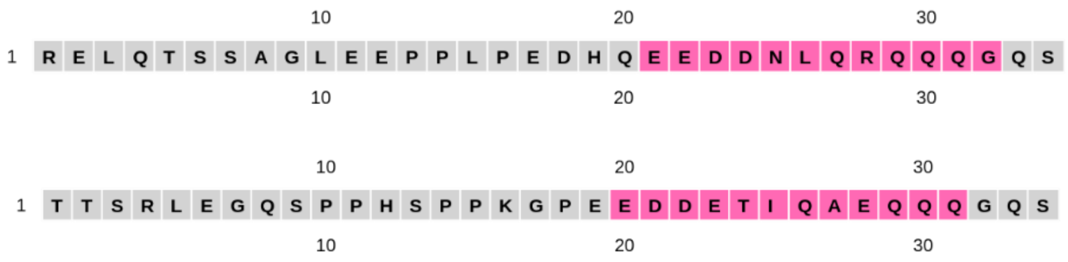

10
20
30
1 R E L Q T S S A G L E E P P L P E D H Q E E D D N L Q R Q Q Q G Q S
10
20
30
10
20
30
1 T T S R L E G Q S P P H S P P K G P E E D D E T I Q A E Q Q Q G Q S
10
20
30

Figure 6.6: PSIPRED sequence analysis results for the hECSIT (top) and mECSIT (bottom) peptides. Coils are shown in grey and helices in magenta.

values of 14 (3) Å with a99SB-disp and of 15 (3) Å with CHARMM36m, while mECSIT has average Rg values of 14 (3) Å with a99sb-disp and of 15 (3) Å with CHARMM36m.

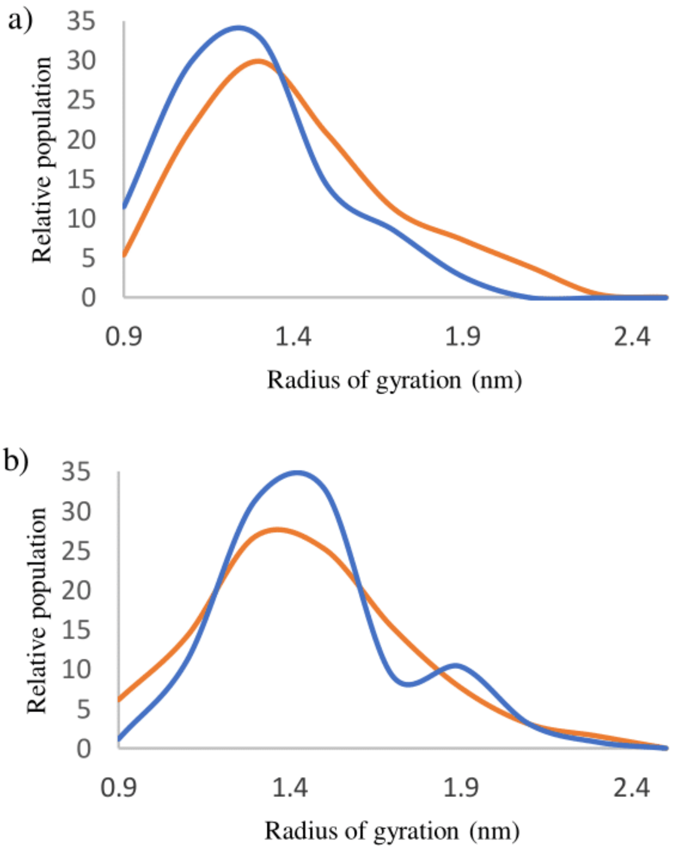


Figure 6.7: Histogram analysis of Rg values obtained for hECSIT (left) and mECSIT (right) with a99SB-disp (blue) and CHARMM36m (orange).

In terms of secondary structure propensity, the hECSIT peptide’s REMD data resulted in 10 clusters for both, the a99SB-disp and CHARMM36m force fields. Meanwhile, the mECSIT peptide REMD data were clustered in 13 and 17 clusters for a99SB-disp and CHARMM36m forcefields, respectively. The STRIDE[21] secondary structure analysis of the clusters is shown in **Table 6.2**. There is 78% and 70% helicity in hECSIT for a99SB-disp and CHARMM36m respectively, with cluster 2 6 and 8 being reclassified as helical for CHARMM36m. While mECSIT has helicity it is represented much lower at 10% for both a99SB-disp CHARMM36m. The structures of these helical motifs are shown in **Figure 6.8**.

An overall analysis of the peptides secondary structure propensity throughout the simulations shows that the sequences can be divided in two regions, one with different degrees of helicity, high for the hECSIT and low for the mECSIT, and one with a high degree of PPII content for both sequences, see **Figure 6.9**. Based on this information, we proposed to generate two mutants, namely hECSITm and mECSITh, were in hECSITm the residues shown in red in **Figure 1.7 panel c)** are swapped with the corresponding residues from mECSIT shown in red in **Figure 1.7 panel d)**. The same criterium was used to create the mECSITh peptide

Table 6.2: STRIDE secondary structure of the 10 highest populated clusters determined for the hECSIT (top) and mECSIT peptides (bottom) with both forcefields. The letters represent the following: T – turn, G – 3.10 helix, H – helix, B β-bridge.

CHARMM36m/TIP3P	***a99SB-disp/TIP4P-D***
RELQTSSAGLEEPPLPEDHQEEDDNLQRQQQGQS(%Pop)	RELQTSSAGLEEPPLPEDHQEEDDNLQRQQQGQS(%Pop)
TTTBTBTT TTTT HHHHTT (33%)	TTTTTTTTT TTTT B GGGG TTTT (60%)
B HHHHH TTTT TTTT (16%)	TTTTTTTTTT TTTT HHHHHHHHHBTT (13%)
TTTT TTTTTTTTTTHHHH BTTTT (14%)	TTTT TTTTTT TTTT TTTTTTT (13%)
TTTT B TTTTTTTBTTT (10%)	HHHHH TTTT TTTTT TTTT (3%)
TTTT (8%)	TTTTT BTTTT (3%)
TTTTTTTB TTT TTTTTTT (4%)	TTTTB TTTTT HHHHTTTTT (3%)
TTTT B TTTT (4%)	HHHHH (2%)
TTTT TTTT TTTTTT TBTT (3%)	B (2%)
TTTTT TTTT (2%)	TTTTTT TTTT TTTTT (1%)
TTTT TBTT TTTTTTT (2%)	TTTTTTTT B (1%)

CHARMM36m/TIP3P	***a99SB-disp/TIP4P-D***
TTSRLEGQSPPHSPPKGPEEDDETIQAEQQQGQS(%Pop)	TTSRLEGQSPPHSPPKGPEEDDETIQAEQQQGQS(%Pop)
TTTTbTTT BBTTTTTTTBT B TTTB (29%)	TTTTTTTT BBTTTTTTBT TTTT (51%)
B TTTT (15%)	TTTT B TTTT (18%)
TBTT TTTTTTTB HHHHH TTTGGG (10%)	TTTT HHHH BTTTTT(8%)
TTTT TTTT B TTTT (8%)	TTTTT BTTTTTTTT (7%)
b BTTTTTTTTT (7%)	TTTT BTTTTTTB B (5%)
TTTT (5%)	TTTTTT BTTTT TTTTT B (2%)
TTTT TTTT TTTT B (4%)	TTTTTTTTTTTTTT TTTTTTTTTTTTTT TTT (2%)
TTTTTTTT B B (4%)	TTTTTTTTTTTTT TTTT B GGG (2%)
TTTTT B (3%)	TTTT TTTT (1%)
TTTTTTTTTT BTTT B (3%)	TTTTTTTTTBTT (1%)

.

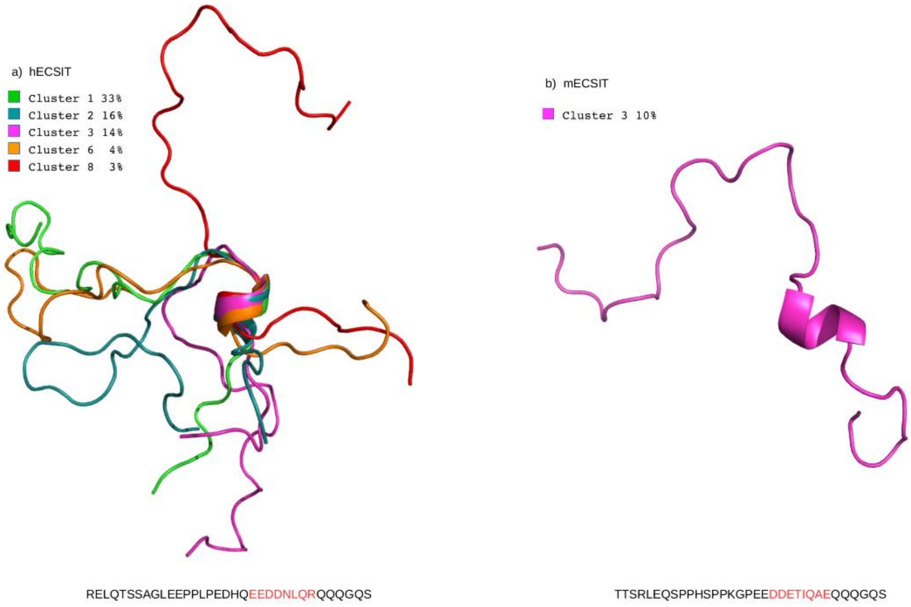


Figure 6.8: Representative structures of the clusters from hECSIT and mECSIT, on the left- and right-hand side, respectively, with propensity to form helical motifs. Results correspond to the simulations run with CHARMM36m, see also **Table 6.2**.

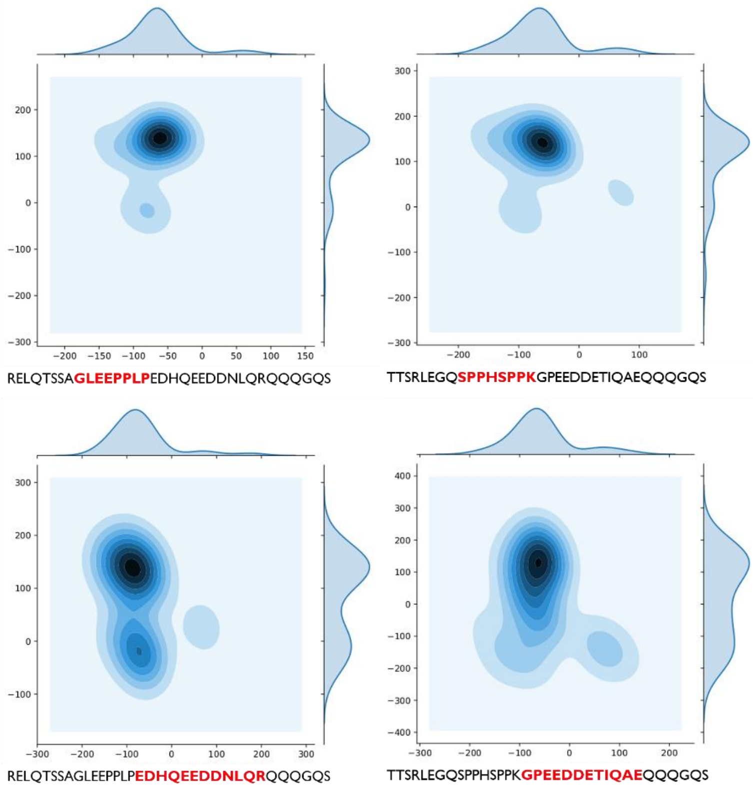


Figure 6.9: Ramachandran (phi/psi) plots of the peptide sequences regions (highlighted in red) with high degree of polyproline II (PPII) for hECSIT in panel a) and mECSIT in panel b) and with higher degree of helicity for for hECSIT in panel c) and for mECSIT in panel d).

The REMD results show that the hybrid peptides, hECSITm and mECSITh, also have very similar dimensions in terms of Rg values, see **Figure 6.10**. As for the simulation of the wild type peptides, the differences in Rg values between the two force fields, is 0.1 nm.

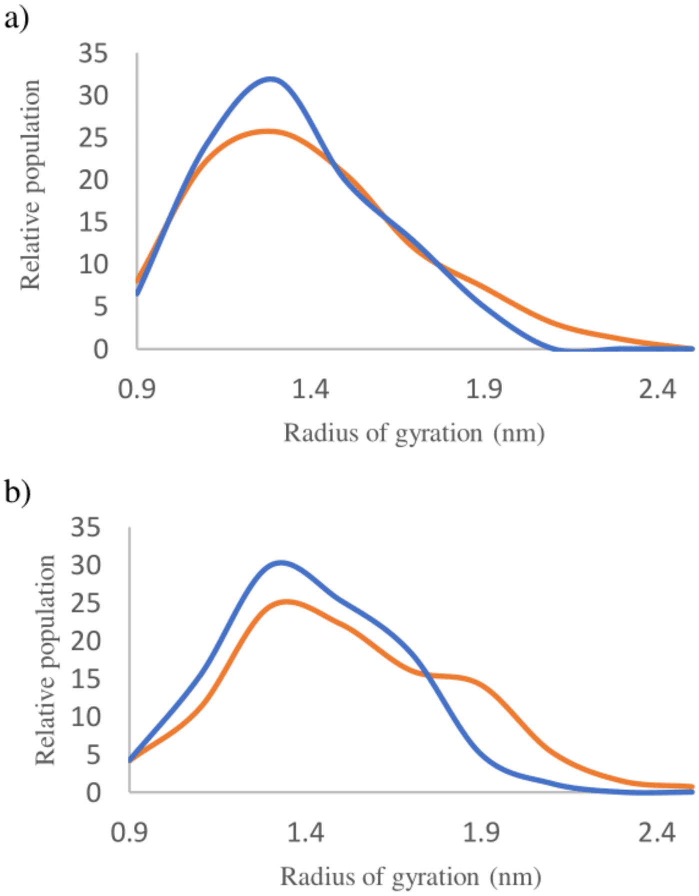


Figure 6.10: Histogram analysis of Rg values obtained for hECSITm (left) and mECSITh (right) with the a99SB-disp (blue) and CHARMM36m (orange) force fields.

Indeed, the hECSITm average Rg value is 15 (3) Å with CHARMM36m and 14 (3) Å with a99sb-disp, while the mECSITh Rg values are 1.4 (0.2) with CHARMM36m and 1.5 (0.3) with a99sb-disp. Using the same clustering parameters determined as optimal for the wild type peptides, the clustering analysis classifies the hECSITm peptide into 12 clusters with a99SB-disp (16 with CHARMM36m) and the mECSITh peptide into 13 clusters with a99SB-disp (17 with CHARMM36m). The STRIDE secondary structure analysis of these clusters is shown in **Table 6.3**. The hECSITm peptide shows a 65% helicity with a99SB-disp and 61% helicity with CHARMM36m. Meanwhile, the mECSITh peptide has a 27% helicity with a99SB-disp and 21% helicity with CHARMM36m with cluster 3 being reclassified as helical for CHARMM36m.

Table 6.3: STRIDE secondary structure of the 10 highest populated clusters determined for the hECSIT (top) and mECSIT peptides (bottom) with both forcefields. The letters represent the following: T – turn, G – 3.10 helix, H – helix, B β-bridge.

CHARMM36m/TIP3P	***a99SB-disp/TIP4P-D***
`RELQTSSAGLEEPPLPGPEEDDETIQAEQQQGQS(%Pop)`	`RELQTSSAGLEEPPLPGPEEDDETIQAEQQQGQS(%Pop)`
`  TTTTTTTTTTT          GGGG TTTT  (32%)`	`  TTTTTTGGGB       TTTHHHH   BTTB (51%)`
`  TTTT          B     GGG         (21%)`	`  TTTTT TTTT B TTTTTTTTTTT        (14%)`
`    TTTT            B             (11%)`	`HHHHHH           TTTTGGGGG        (14%)`
`      TTTT       B TTBTB          (11%)`	`TTTTTTTTBb     TTTT          TTTT ( 8%)`
`  TTTT  TTTT B   B    TTTTTTGGG   ( 8%)`	`TTTTTTT TTTT     B   TTTTTTT      ( 5%)`
`         B            B           ( 3%)`	`TTTTTTTTT       TTTT TTTTT        ( 3%)`
`  TTTT TTTTTT      B      TTTTTT  ( 3%)`	`TTTT             TTTTTTTT  TTTT   ( 2%)`
`TTTTTT  TTTT          GGGGTTTT    ( 2%)`	`                 TTTTTTTTTTTTTT   ( 1%)`
`   TTTTTTT       TTTTTTTTT        ( 2%)`	`     b              TTTTTTTTTTTTT ( 1%)`
`TTTTB                BTTTTTT      ( 1%)`	`TTTTT TTTTTT         TTTTTT       ( 1%)`

CHARMM36m/TIP3P	***a99SB-disp/TIP4P-D***
`TTSRLEGQSPPHSPPKEDHQEEDDNLQRQQQGQS(%Pop)`	`TTSRLEGQSPPHSPPKEDHQEEDDNLQRQQQGQS(%Pop)`
`TTTTTTTTT       BTTTTTTTBT   TTTT (31%)`	`TTTTTTTT      BB TTTTTTTTTTTTTTTT (43%)`
`     B    TTTT  TTTT  TTTT        (17%)`	`TTTTB    TTTT  TTTT  HHHHH   TTTT (15%)`
`TTTTTTTTT       TTTTTTTTB         (11%)`	`  BTBTTB       TTTTTTTTTGGG       (12%)`
`                     HHHHB        ( 8%)`	`TTTTTTTT TTTTT TTTTTBBTTTTTTTTTT  ( 7%)`
`      B  TTTT                     ( 7%)`	`    TTTTT TTTT BTTTTTTTTT         ( 6%)`
`         TTTT  TTTT  TTTTTT       ( 5%)`	`  BTTTTTT  B    TTTTTTTTTT TTTTBTT( 4%)`
`         TTTT              TTTTT  ( 3%)`	`TT        TTTT TTTT     TTTTTTTT  ( 4%)`
`      TTTT   TTTTTTT GGG TTTT     ( 3%)`	`     BTTT TTTT  TTTTTTTTT TTT     ( 4%)`
`   BTTTT TTTT               TTTT  ( 2%)`	`TTTTTT   TTTT        TTTTTTTTTTT  ( 2%)`
`    B                TTTTT        ( 2%)`	`            B  TTTT  TTTTTTTT     ( 2%)`

6.4 Discussion

The REMD simulations results provides important information on the structure and disorder of the two ‘wild-type’ peptides, i.e. hECSIT and mECSIT, corresponding to the C-terminal regions of the human and murine ECSIT proteins, and of the two mutant peptides, were specific sets of residues were swapped between hECSIT and mECSIT. In terms of average dimensions/compactness of the ensemble, hECSIT and mECSIT, and also hECSITm and mECSITh are very similar, if not identical. Also, the Rg average values obtained with the two different force fields, i.e. a99SB-disp/TIP4P-D and CHARMM36m/TIP3P, for all peptides are within one standard deviation, with the ensemble obtained with the a99SB-disp/TIP4P-D force field slightly more compact. Compactness has been shown to be highly dependent on forcefield[23]. However, this is an usual finding as the TIP4P-D was parameterised to alleviate the over compactness found with the TIP3P water model[15]. Overall, the difference in compactness between a99SB-disp/TIP4P-D and CHARMM36m/TIP3P is on the scale of approximately 1 Å which is less than the standard * of the values.

The structural disorder of the peptides was gauged based on the number of clusters determined through *Gromos* clustering analysis[20] followed by a secondary structure analysis based on the STRIDE classification[21]. The number of clusters (10) does not change between a99SB-disp and CHARMM36m for hECSIT, meanwhile the two force fields provide slightly different results in the case of mECSIT, predicting 17 clusters with CHARMM36m and 13 clusters with a99SB-disp. Based on this classification mECSIT appears to be slightly more disordered than hECSIT, although only the first 10 clusters are significantly populated, accounting for 99% and 90% of the total simulation for a99SB-disp and CHARMM36m, respectively.

The secondary structure analysis provides important clues about the origin of this slight difference in the conformational propensity between hECSIT and mECSIT. Indeed, the two peptides can be divided two regions in terms of sequence, one with a high degree of PPII structure in both hECSIT and mECSIT towards the N-terminus and another region that has a high degree of helicity in hECSIT, but a significantly lower degree of helicity in mECSIT at the C-terminus, see **Figure 1.7**. Based on this information we swapped the sequence of residues that have the characteristic high helicity in hECSIT with the residues with low helicity in mECSIT to test if this specific section was determinant for the observed protein stability differences. Experimental results have shown that the swap of the 12-residue sequence we identified between reversed the relative stabilities of the whole ECSIT proteins see **Figure 6.11**[6]. More specifically human ECSIT became stable, while murine ECSIT became unstable, to the same extent as obtained by swapping the whole C-terminal tail[6].

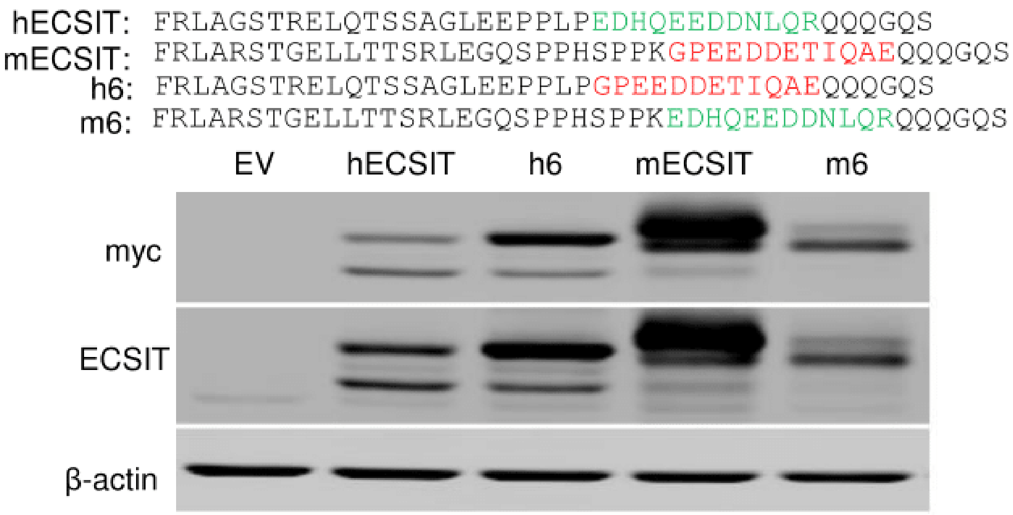

hECSIT: FRLAGSTRELQTSSAGLEEPPLPEDHQEEDDNLQRQQQGQS
mECSIT: FRLARSTGELLTTSRLEGQSPPHSPPKGPEEDDETIQAEQQQGQS
h6: FRLAGSTRELQTSSAGLEEPPLPGPEEDDETIQAEQQQGQS
m6: FRLARSTGELLTTSRLEGQSPPHSPPKEDHQEEDDNLQRQQQGQS
EV
hECSIT
h6
mECSIT
m6
myc
ECSIT
β-actin

Figure 6.11: Western blot analysis of cell lysates from HEK293T cells transfected with empty vector (EV) or expression constructs encoding myc-tagged hECSIT, myc-tagged mEcsit mutants or indicated hybrid mutants of mECSIT and hECSIT.

Simulations of the mutant sequences, namely mECSITh and hECSITm, provide further insight into the structural features that may be responsible for the observed behaviour. The mECSITh peptide showed an increased helicity relative to mECSIT, with 27% and 21% helicity of the swapped sequence obtained with a99SB-disp and CHARMM36m, respectively. In case of the hECSITm peptide, based on the behaviour of the 12 residues sequence from mECSIT we introduced, we were expecting a decreased helicity, meanwhile we found that the helicity remains approximately the same as observed for the wild-type, namely 65% and 60% for a99SB-disp and CHARMM36m, respectively. The link between helicity and lack of stability we were pursuing, considering that being the only difference in the conformational propensity of the two peptides, derives from the role that residual secondary structure plays in the molecular recognition of motifs within disordered regions[24,25]. These residual structural features contain the molecular determinants that facilitate specific recognition and binding. The partially structured motifs function as nucleation sites for folding which occurs upon binding[24,25]. There are a few examples where partial helical motifs within intrinsically disordered regions are specifically recognized by receptors are fold into full alpha helices upon binding[26,27]. Within this framework, the helical motif within the hECSIT peptide could be the docking point for the interaction with a specific receptor that can lead to the protein degradation, meanwhile this interaction cannot occur in mECSIT. When we swap the sequences that may carry a degree of intrinsic propensity for helicity, we may have introduced in mECSITh the ability of binding to this not-yet identified receptor, meanwhile we lost it in hECSITm, not because the motive has lost its helicity but because the specific residues that form that helix are actually different, as they correspond to the mECSIT sequence. Indeed, the helical turn that may function as nucleation site to bind the receptor involves 4 residues, namely EDDNL in the case of hECSIT and ETIQA in the case of hECSITm. These specific mutations could negate the initial specific contacts with the receptor binding site and prevent recognition, thus binding.

6.5 Conclusions

In this work we determined that the extreme C-terminal domain of the protein ECSIT is intrinsically disordered in both human and murine proteins. From an initial scan of the sequence with disorder and structure prediction tools, namely disEMBL and PSIPRED, we selected a 10-residue region of the human and murine ECSIT C-terminal tails to build peptides and carry out a thorough conformational analysis by means of REMD simulations. Our results indicate a significantly different propensity to form helical motifs between the two tails, located in a specific 12 residues region, with hECSIT having a much higher propensity than mECSIT. Based on this information we proposed to swap these regions to create mutant peptides that we have further investigated with REMD, and that our collaborator have done within the context of the whole protein. Experimental data strongly indicate that these 12 residues regions are indeed the key to understand the very different stabilities of human and murine ECSIT and simulation results indicate that the residual helicity in hECSIT and mECSITh could determinant for the recognition and binding to a not-yet identified receptor. This interaction could be key to understand the role of ECSIT in the mitochondrial respiratory pathway.

References

1. Soler-López, M., Badiola, N., Zanzoni, A. & Aloy, P. Towards Alzheimer’s root cause: ECSIT as an integrating hub between oxidative stress, inflammation and mitochondrial dysfunction. *BioEssays* **34**, 532–541 (2012).
2. Nouws, J. *et al.* Acyl-CoA dehydrogenase 9 is required for the biogenesis of oxidative phosphorylation complex I. *Cell Metab.* (2010).

doi:10.1016/j.cmet.2010.08.002

3. Vogel, R. O. *et al.* Cytosolic signaling protein Ecsit also localizes to mitochondria where it interacts with chaperone NDUFAF1 and functions in complex I assembly. *Genes Dev.* **21**, 615–624 (2007).

4. Xiao, C. *et al.* Ecsit is required for Bmp signaling and mesoderm formation during mouse embryogenesis. *Genes Dev.* **17**, 2933–2949 (2003).

5. Carneiro, F. R. G., Lepelley, A., Seeley, J. J., Hayden, M. S. & Ghosh, S. An Essential Role for ECSIT in Mitochondrial Complex I Assembly and Mitophagy in Macrophages. *Cell Rep.* (2018). doi:10.1016/j.celrep.2018.02.051

6. Xu, L. *et al. Submitted,* Human ECSIT protein is critical for assembly of mitochrondrial complex I in the heart: Key limiting factor for cardiac function.

7. Linding, R. *et al.* Protein disorder prediction: implications for structural proteomics. *Structure* **11**, 1453–1459 (2003).

8. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).

9. Brooks, B. & Karplus, M. Normal modes for specific motions of macromolecules: application to the hinge-bending mode of lysozyme. *Proc. Natl. Acad. Sci. U. S. A.* **82**, 4995–4999 (1985).

10. Li, X., Obradovic, Z., Brown, C. J., Garner, E. C. & Dunker, A. K. Comparing predictors of disordered protein. *Genome Inform. Ser. Workshop Genome Inform.* **11**, 172–184 (2000).

11. Buchan, D. W. A. & Jones, D. T. The PSIPRED Protein Analysis Workbench: 20 years on. *Nucleic Acids Res.* (2019). doi:10.1093/nar/gkz297

12. Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292**, 195–202 (1999).

13. Patriksson, A. & van der Spoel, D. A temperature predictor for parallel tempering simulations. *Phys. Chem. Chem. Phys.* **10**, 2073–2077 (2008).

14. Robustelli, P., Piana, S. & Shaw, D. E. Developing a molecular dynamics force field for both folded and disordered protein states. *Proc. Natl. Acad. Sci.* **115**, E4758–E4766 (2018).

15. Piana, S., Donchev, A. G., Robustelli, P. & Shaw, D. E. Water Dispersion Interactions Strongly Influence Simulated Structural Properties of Disordered Protein States. *J. Phys. Chem. B* **119**, 5113–5123 (2015).

16. Huang, J. *et al.* CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat. Methods* **14**, 71 (2016).

17. Abraham, M. J. *et al.* GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1–2**, 19–25 (2015).

18. Maestro, version 9.2. Schrödinger, LLC; New York, NY, U. 2011. .

19. The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, L. .

20. Daura, X. *et al.* Peptide Folding: When Simulation Meets Experiment. *Angew. Chemie Int. Ed.* **38**, 236–240 (1999).

21. Frishman, D. & Argos, P. Knowledge-based protein secondary structure assignment. *Proteins Struct. Funct. Bioinforma.* **23**, 566–579 (1995).

22. Galindo-Murillo, R., Roe, D. R. & Cheatham 3rd, T. E. Convergence and

reproducibility in molecular dynamics simulations of the DNA duplex d(GCACGAACGAACGAACGC). *Biochim. Biophys. Acta* **1850**, 1041–1058 (2015).

23. Rauscher, S. *et al.* Structural Ensembles of Intrinsically Disordered Proteins Depend Strongly on Force Field: A Comparison to Experiment. *J. Chem. Theory Comput.* **11**, 5513–5524 (2015).
24. Wright, P. E. & Dyson, H. J. Intrinsically disordered proteins in cellular signalling and regulation. *Nat. Rev. Mol. Cell Biol.* **16**, 18 (2014).
25. Fadda, E. & Nixon, M. G. The transient manifold structure of the p53 extreme C-terminal domain: insight into disorder, recognition, and binding promiscuity by molecular dynamics simulations. *Phys. Chem. Chem. Phys.* **19**, 21287–21296 (2017).
26. Rust, R. R., Baldisseri, D. M. & Weber, D. J. Structure of the negative regulatory domain of p53 bound to S100B(ββ) . *Nat. Struct. Biol.* **7**, 570–574 (2000).
27. Mer, G. *et al.* Structural Basis for the Recognition of DNA Repair Proteins UNG2, XPA, and RAD52 by Replication Factor RPA. *Cell* **103**, 449–456 (2000).

Conclusions

As outlined in the introduction, the overarching goal of my thesis work was to investigate the relative structural propensity of IDPs in function of the specific sequence and how this structure level is integral to the specific IDP/R function. Also, another important goal was to show how computer simulation methods, such as conventional and enhanced sampling MD simulations, can be instrumental for the identification of structural motifs that cannot be detected on the experimental timescale and thus that it can be used as a primary discovery tool and not only as a support for experimental data. Indeed, while some studies in my thesis work are only computational and the links between structure level and function of the IDP systems studied are made based on earlier experimental studies, the work on LxCxE SLiM-containing peptides, see **Chapter 5**, and the human *vs.* murine ECSIT C-terminal tail, see **Chapter 6**, are both computational and experimental, where computational results are an essential support for the understanding of the molecular/atomistic basis of the IDP function.

As an overall summary, we obtained many interesting results from our simulations that all support the key role of residual secondary structure in the molecular recognition and function of IDPs. In the case of the XPA 14-residue peptide, discussed in **Chapter 3**, we found that there is a transiently stable β-hairpin motif that is potentially recognised by ERCC1 as a MoRF. Because of the high structural similarity of this MoRF with the ERCC1-bound conformation of the XPA 14-residue peptide, we suggest that the peptide follows a coupled folding and binding mechanism, through a recognition mechanism that sits in between the conformational selection and induced-fit schemes. Crucially, our results also show that the β-hairpin propensity is highly sequence-dependent. For example, in the case of the *C. lanigera* XPA$_{67\text{-}80}$, a single Gly-to-Glu mutation halves the population of a binding conformer. Within this context we have also examined how restricting the peptide conformational flexibility to a MoRF-like structure through macrocyclization is a potentially viable tool to reduce the entropic penalty of binding. Computing here can be instrumental in ranking potential candidates for synthesis both in terms of conformational flexibility and optimal fit into the receptor binding site, but also in terms of binding affinity, via end-point methods, as it was done in this work.

In case of the p53-CTD our aim was to investigate the molecular basis for the very interesting binding promiscuity of the peptide. Our simulations results, discussed in **Chapter 4**, show that the ability of this peptide to bind specifically different number of receptors may rest in the complexity of its substructure, namely in its ability to adopt different stable secondary structure motifs that can act as multiple different MoRFs within the conformational ensemble. These MoRFs comprise helices, namely α-helical and 3_{10} helical turns, and β structures, i.e. β-hairpins and β-strands. Notably recent experimental data support our findings[1].

In **Chapter 5** we discuss the role of the variable “x” residues in regulating the binding affinity of a large set of LxCxE SLiM-containing peptides to the Rb protein. This work has been run hand-in-hand with peptide synthesis and binding kinetic experiments. We found that modification of the “x” variable residues and of the residues before and after LxCxE motif on the HPV E7 peptide does not affect the structural propensity of the free peptide, which is not pre-structured, or that does not form any stable residual secondary structure. Through lengthily simulations of the bound conformation run in many different conditions, we found that the modulation of the binding affinity as a function of the variable “x” residues results from changes in the direct contacts between the motif and the Rb binding cleft that are triggered by the nature of the “x” residues. Interestingly and very excitingly all our simulation results support very well the narrative obtained from the experimental results. Additionally, we found that the peptide’s phosphorylation promotes a PPII structure, a conformation more similar to that of the bound peptide.

Finally, in **Chapter 6** I discussed the results we obtained in the investigation of the human *vs* murine ECSIT C-terminal tail and its potential links to the very different stability of the two proteins verified experimentally by our collaborators. Though the exact function of the C-terminal tail of ECSIT is not known, we found a striking difference in the structural propensities of the hECSIT and mECSIT C-terminal peptides. Indeed, hECSIT shows a high degree of helicity relative to mECSIT, which has a much lower helicity and no particular or specific residual secondary structure. Through extensive sampling by REMD we were able to identify the specific regions in the ECSIT C-terminus that carried these helical motifs and proposed to our collaborator to swap of these regions between human and murine and to test the relative stability of the two mutants. Experiments proved that swapping this region inverts the whole protein stability to the same degree as swapping the whole tails and thus suggesting that this specific helical motif could function as a MoRF, facilitating binding of the hECSIT to a degrading protein or to a protein that leads hECSIT to a degrading pathway.

In this work I have looked at four different IDP systems with different sequences, level of intrinsic disorders and functions. Indeed, XPA is scaffolding protein involved in macromolecular assembly along the NER pathway, p53 is a tumour suppressor, LxCxE-containing peptides are contained in an oncoprotein, while the ECSIT C-terminal tail function is yet to be established. They all show an over representation of key residues, namely the charged residues, Arg, Glu, and Lys. As well the hydrophilic residues Asn, Gln, and Ser. They also feature high levels of Pro and Gly which are all considered to be disorder promoting residues, however as shown in **Figure 7.1** there is no significant difference in composition of amino acids between disordered and ordered peptides. As IDPs, they all show a lack of structure at the experimental time scale, with the exception of ECSIT which structure has not been solved and for which the C-terminal tail is only predicted to be disordered, a prediction supported by our results discussed in **Chapter 6**. Our simulations results have shown that prestructuring plays an important role in all of these peptides. In the case of XPA there is a high degree of prestructuring towards a highly populated hairpin that can be potentially recognised by ERCC1. Prestructuring of p53 is a bit different, as it involves different structural motifs. Indeed, the p53 C-ter conformational ensemble includes a number of low populated α-helix and β-turn structures that could be recognised by multiple different binding partners, which can explain its binding promiscuity. The LxCxE motif shows little to no prestructuring, which is not surprising considering the size of this motif. However, we have shown that upon phosphorylation it adopts an increased PPII content, consistently with experimental results. Finally, we have determined that the ECSIT C-terminal tail shows a propensity to adopt an α-helix motif in both human and mouse sequences, with significantly different relative populations, which could be affecting the stability of the whole protein.

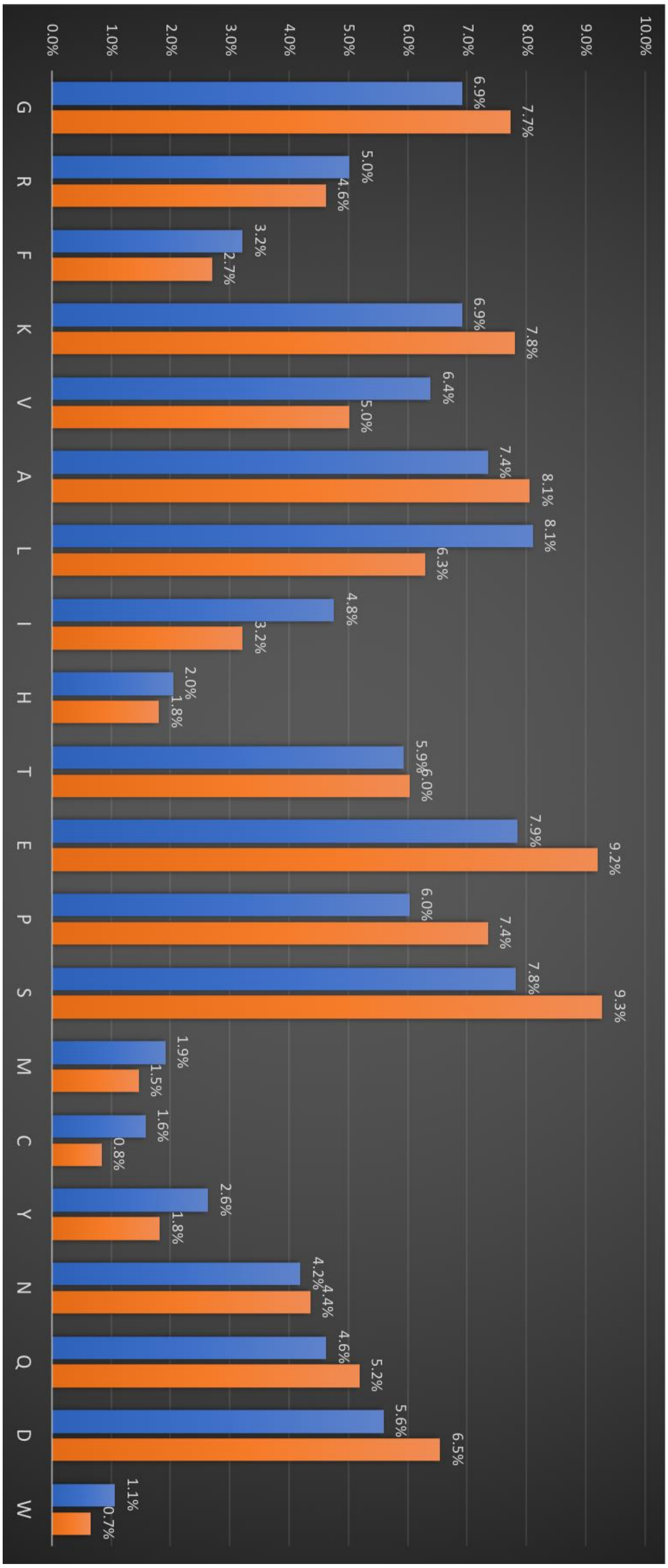


Figure 7.1: Comparison of amino acid composition of ordered proteins (blue) and IDPs (orange)

The main difference between the LxCxE-containing peptide and the three other systems is that it is much less sensitive to the effects of point mutations. In the case of XPA a single Phe to Trp mutation is sufficient to significantly destabilise the prestructuring. Since the LxCxE containing peptide is only prestructured by the phosphorylation of two serine residues, a double Tyr to Ala mutation or Glu to Ala mutation have no effect on the prestructuring as seen in our results and also confirmed by lack of change in K_{on} experimentally.

Through all my thesis work presented here I hope to have contributed in showing how conformational disorder can be functionally important and diverse in different proteins and protein regions. My work as a whole supports the view that the term intrinsic disorder does not indicate that a sequence is devoid of secondary structure, but only that its structural propensity cannot be probed at the experimental scale and that its rich and informative complexity can be studied successfully through computer simulation methods.

References

1. Krüger, A. *et al.* Interactions of p53 with poly(ADP-ribose) and DNA induce distinct changes in protein structure as revealed by ATR-FTIR spectroscopy. *Nucleic Acids Res.* **47**, 4843–4858 (2019).