A COMBINED COMPUTATIONAL STUDY OF THE STRUCTURE AND BINDING OF THE HISTONE H3 N-TERMINAL DOMAIN IN THE NUCLEOSOME

by

Louis Lategan Du Preez

Submitted in accordance with the requirements for the degree

Magister Scientiae

In the Department of Microbial, Biochemical and Food Biotechnology
Faculty of Natural and Agricultural Sciences
University of the Free State
Bloemfontein
South Africa

FEBUARY 2012

Supervisor: Prof. Hugh-George Patterton

Co-supervisor: Prof. Matie Hoffman

ACKNOWLEDGEMENTS

I would like to thank and acknowledge the following:

My parents, for providing food, shelter, laundry service and encouragement even though they still think I attend classes at university

Prof. Hugh-George Patterton, for providing me with support, insight, guidance and most importantly the freedom I need to develop as a scientist

Mr. Albert van Eck, for helping me with the UFS High Performance Cluster

My friends, for distracting me from work whenever I need it most

Dr. Gabrè Kemp, for making Monday lab meetings worth attending

Ms. Charlene van der Vyver, for doing the administration and making hot beverages

Dr. E Patterton, for giving advice and gossip

Departmental lab mates, for offering encouraging words in a dire situation

The NRF and UFS Strategic Academic clusters, for financial support

Lulu and Tienkie, for being pugs and making oversleep.

Creator/God, because I sure as hell didn’t possess the mental fortitude or strength to finish this thesis and resist The Elder Scrolls V – Skyrim for almost 2 months

INDEX

TABLE OF CONTENTS iv

LIST OF NON – SI ABBREVIATIONS xii

TABLE OF CONTENTS

CHAPTER 1

Literature Review

1.1 INTRODUCTION 1

1.1.1 The need for DNA packaging 1

1.1.2 Histone tails: more than just fashionable 2

1.1.3 Histone tails: beyond the histone code 3

1.2 CHROMATIN STRUCTURE 5

1.2.1 A model for the 30 nm fibre 5

1.2.2 Position of the H2A and H2B tails 6

1.2.3 Position of the H3 tail 6

1.2.4 Position of the H4 tail 8

1.3 HISTONE TAIL ASSOCIATIONS 9

1.3.1 N-H2A, H2A-C and N-H2B 9

1.3.2 N-H3 10

1.3.3 N-H4 11

1.4 HISTONE TAIL STRUCTURE 11

1.4.1 Secondary structure prediction 11

1.4.2 Molecular Dynamics 13

1.4.3 Biophysical methods 15

1.5 THE NUCLEOSOME SURFACE – POKER FACE OF CHROMATIN REGULATION? 16

1.5.1 The acidic patch 17

1.5.2 Molecular Docking 17

1.6 HISTONE TAILS AND HUMAN HEALTH 19

1.7 CONCLUSIONS AND INTRODUCTION TO CURRENT STUDY 20

1.8 REFERENCES 22

CHAPTER 2

Development of tools for the analysis of Molecular Dynamics Trajectories in YASARA

2.1 INTRODUCTION 30

2.2 AN OBJECT ORIENTATED MODEL FOR CONDUCTING MD TRAJECTORY ANALYSIS USING PYTHON AND YASARA 31

2.2.1 MD Trajectories in YASARA 31

2.2.2 The trajectory object: Traj() 31

2.2.3 The snapshot object: Snapshot() 32

2.2.4 The analysis object: Analysis() 32

2.2.5 Implementation of the model: md_analyze.py 32

2.3 SIMDB : A RELATIONAL DATABASE FOR THE STORAGE OF MD TRAJECTORY ANALYSIS RESULTS 34

2.3.1 Database schema and implementation 34

2.3.1.1 Experiment 36

2.3.1.2 Author 37

2.3.1.3 Parameter 38

2.3.1.4 Molecular system 39

2.3.1.5 Atom 40

2.3.1.6 Residue 40

2.3.1.7 Residue info *41*

2.3.1.8 Snapshot *42*

2.3.1.9 Sec_struct *42*

2.3.1.10 H_bond *44*

2.3.2 Methods (software used) 44

2.4 DISCUSSION AND CONCLUSION **45**

2.5 REFERENCES **47**

CHAPTER 3

Validation of the Docking Method

3.1 INTRODUCTION **48**

3.2 METHODS **50**

3.2.1 Structure Preparation: 50

3.2.2 Parameters tested *52*

3.2.2.1 Ligand Flexibility *53*

3.2.2.2 Docking Search Area Size *53*

3.2.2.3 Ligand Orientation and Position *53*

3.2.3 Docking Algorithm Parameters 54

3.2.4 Grid – based Docking 54

3.2.5 Docking Analysis 55

3.3 RESULTS **57**

3.3.1 Non-grid – based docking experiments 57

3.3.1.1 Rigid Docking *57*

3.3.1.2 Flexible Docking *59*

3.3.2 Grid – based Docking 62

3.3.2.1 Rigid Grid – based Docking *62*

3.3.2.2 Flexible Grid – based Docking *63*

3.4 DISCUSSION 68

3.5 CONCLUSION 70

3.6 REFERENCES 71

CHAPTER 4

A Molecular Dynamics analysis of the role of epigenetic modifications on the structure of the histone H3 N-terminal tail

4.1 INTRODUCTION 73

4.2 METHODS 74

4.2.1 Structure preparation 74

4.2.2 H3 N-terminal peptides 75

4.2.3 Modeling of PTMs 75

4.2.4 Control peptides 77

4.2.5 Experimental systems 78

4.2.6 Molecular Dynamics (MD) simulation setup 78

4.2.7 Energy minimization 79

4.2.8 MD production run 79

4.2.9 Secondary structure prediction using internet-based tools 80

4.2.10 Secondary Structure and Hydrogen bonding analysis 81

4.2.11 Clustering analysis 81

4.3 RESULTS 82

4.3.1 Secondary structure prediction 82

4.3.2 Secondary structure during 500 ns explicit MD simulations 83

4.3.2.1 WT tail 83

4.3.2.2 ACTIVE tail 84

4.3.2.3 INACTIVE tail 87

4.3.2.4 HYPER – ALY tail 87

4.3.2.5 K9ME1_S10PHO tail 91

4.3.2.6 K9ME2_S10PHO tail 91

4.3.2.7 K9ME3_S10PHO tail 94

4.3.2.8 K9ME1 tail 96

4.3.2.9 K9ME2 tail 98

4.3.2.10 K9ME3 tail 100

4.3.2.11 K9ACE_S10PHO tail 102

4.3.2.12 ALA_POS_CTRL tail 102

4.3.2.13 GLY_NEG_CTRL tail 102

4.3.3 Hydrogen Bonding Analysis 107

4.3.3.1 ALA_POS_CTRL tail 107

4.3.3.2 GLY_NEG_CTRL tail 109

4.3.3.3 WT tail 109

4.3.3.4 ACTIVE tail 112

4.3.3.5 Inactive tail 114

4.3.3.6 HYPER-ALY tail 114

4.3.3.7 K9ME1_S10PHO tail 117

4.3.3.8 K9ME2_S10PHO tail 117

4.3.3.9 K9ME3_S10PHO tail 120

4.3.3.10 K9ME1 tail *120*

4.3.3.11 K9ME2 tail *123*

4.3.3.12 K9ME3 tail *123*

4.3.3.13 K9ACE_S10PHO tail *126*

4.3.4 Clustering of 500 ns explicit MD trajectories 128

4.3.4.1 WT tail *128*

4.3.4.2 ACTIVE tail *128*

4.3.4.3 INACTIVE tail *129*

4.3.4.4 HYPER – ALY tail *129*

4.3.4.5 K9ME1_S10PHO tail *129*

4.3.4.6 K9ME2_S10PHO tail *129*

4.3.4.7 K9ME3_S10PHO tail *130*

4.3.4.8 K9ME1 tail *130*

4.3.4.9 K9ME2 tail *131*

4.3.4.10 K9ME3 tail *131*

4.3.4.11 K9ACE_S10PHO tail *131*

4.3.4.12 GLY_NEG_CTRL tail *132*

4.3.4.13 ALA_POS_CTRL tail *132*

4.4 DISCUSSION **138**

4.4.1 The unmodified tail shows the formation of two distinct a – helical regions 138

4.4.2 Active tail versus Inactive tail – a structural difference 138

4.4.3 Hyper acetylation induces to a loss of secondary structure in the H3 tail 144

4.4.4 Methylation of K9 – Invisible hand guiding tail folding? 144

4.4.5 Serine 10 phophorylation – Epigenetic Mercenary? 147

4.5 CONCLUSION 149

4.6 REFERENCES 150

CHAPTER 5

Docking of the H3 N-terminal tail to the nucleosome core

5.1 INTRODUCTION 157

5.2 METHODS 157

5.2.1 Clustering of the N-terminal tip structures from MD trajectories 157

5.2.2 Selection of structures for molecular docking 158

5.2.3 Molecular docking of N-terminal tip structures 158

5.2.4 Analysis of docking results 159

5.3 RESULTS 159

5.3.1 Clustering of the N-terminal tip structures from MD trajectories 159

5.3.2 Molecular docking of the N-terminal tip structures to the NCP 164

5.3.3 Contact of the histone H3 N-terminal tail with the nucleosome 170

5.4 DISCUSSION 175

5.4.1 Grid – based docking – a new idea using existing techniques 175

5.4.2 The H3 tail binds between the DNA and the side of the octamer 175

5.4.3 The effect of PTMs on the binding of the H3 tail to the nucleosome 176

5.5 CONCLUSION 177

5.6 REFERENCES 178

CHAPTER 6

GENERAL DISCUSSION AND CONCLUSION

6.1 DISCUSSION 180

6.2 CONCLUSION 183

6.3 REFERENCES 184

SUMMARY 186

OPSOMMING 190

LIST OF NON – SI ABBREVIATIONS

ATP	Adenosine triphosphate
CD	Circular Dichroism
CPU	Central processing unit
CSV	Comma-separated values
DNA	Deoxyribonucleic acid
GUI	Graphical user interface
HP1	Heterochromatin Protein I
KSHV	Kaposi’s Sarcoma-associated herpes virus
LANA	Latency - associated nuclear antigen
LGA	Lamarckian Genetic Algorithm
MD	Molecular Dynamics
NCP	Nucleosome core particle
NMR	Nuclear magnetic resonance
PDB	Protein databank
RMSD	Root – mean – square deviation
SQL	Structured Query Language
TFE	Tetrafluoroethylene

CHAPTER 1

Literature Review

1.1 INTRODUCTION

1.1.1 The need for DNA packaging

The total length of the DNA in a single diploid human cell is approximately 2m, a length that must fit into a cell nucleus that is roughly 10μm in diameter. To accomplish this, the DNA is packaged into arrays of nucleosomes. Each nucleosome is formed by spooling about 168 bp of DNA in two negative superhelical turns onto a histone octamer, which is composed of two copies of each of the core histones H2A, H2B, H3 and H4. A fifth histone, linker histone H1, binds to the outside of the structure, close to the point of DNA entry and exit. The H1 causes partial charge neutralisation of the linker DNA, which connects adjacent nucleosomes in the array[1].

This array of nucleosomes is further condensed into a 30 nm fibre, which is composed of a helical arrangement of nucleosomes. No definitive structural detail is currently available on the 30 nm fibre. The 30 nm fibre undergoes additional levels of folding to form higher order structures, culminating in the condensed structures observed electron microscopically in the metaphase chromosome[2].

Although the packaging of DNA into chromatin solves the problem of fitting an extended, poly-anionic, linear polymer into the confined space of a eukaryotic nucleus, a significant problem is introduced in that the DNA molecule also becomes masked from most of the proteins and enzymes that must interact with it as part of its genetic function. Thus, to allow access to the DNA molecule, eukaryotic cells have evolved intricate mechanisms whereby the chromatin is locally and reversibly decondensed. Mechanisms involved in this local decondensation include the structural perturbation of chromatin structures by ATP-dependent chromatin remodelling enzymes, the deposition of different histone isotypes, and the reversible chemical modification of the "histone tails".

1.1.2 Histone tails: more than just fashionable

The histone tails are seemingly unstructured extensions of the core histones beyond the central histone fold domains [3, 4], and contribute approximately 38% of the core histone mass to the histone octamer (Figure 1.1). The chemical modification of the histone was first observed by Phillips[5, 6] and by Murray [7]. The Mirsky group subsequently showed that acetylation of histones facilitated synthesis of RNA in cell-free extracts [8]. These initial observations defined the beginning of a field that has become known as *Epigenetics* and its high-throughput application, *Epigenomics*. A substantial scientific literature has since developed describing an extensive range of modifications (Figure 1.1) and the function of these modifications [see Kouzarides [9] and Kundu [10] for reviews].

Early indications were that modified residues served as molecular beacons for the recruitment of specific proteins to such flagged areas of the genome [11]. For instance, it was shown that the heterochromatin-associated protein HP1 was recruited to regions marked for transcriptional silencing by tri-methylated K9 of histone H3 [12]. Regulatory effects of one modification on another modification in the same tail or in the tail of a different histone were also discovered, termed *cis*-tail and *trans*-tail pathways, respectively. For instance, phosphorylation of S10 of H3 was shown to inhibit demethylation of the mono- and di-methylated K4, thus maintaining an "active" epigenetic signal [13]. Ubiquitination of K123 of histone H2B required a sequence motif in the H2A tail, which is in close proximity to H2B K123 in the nucleosome, and may be involved in the recruitment of the ubiquitination machinery [14]. Ubiquitination of H2B K123, in turn, was required for recruitment of the methyltransferase complexes for the subsequent methylation of H3 K4 by Set1 [15] and of H3 K79 by Dot1 [16], marks associated with transcriptional activation. Finally, it was shown that DNA methylation disrupted the recruitment of Fbxl11/KDM2A, a demethylase complex targeting methylated lysines [17]. This system of molecular flags and interdependencies is known as the histone code, proposing that histone modification (or DNA modifications) represent a template for direct "read-out" by other proteins that then perform specific chromatin-associated functions [11].

1.1.3 Histone tails: beyond the histone code

Although there are many instances where this histone code model is an accurate description of biochemical functions *in vivo*, instances were also observed where histone tail modifications represented more than simple molecular beacons. The most striking observation involved K16 of histone H4. *In vitro* data showed that deacetylation of K16 was required for full compaction of chromatin into a condensed fibre in the presence of a linker histone [18]. In the absence of H1, acetylation of H4 K16 was also shown to inhibit formation of a condensed structure in a reconstituted nucleosome array, although the relationship of this array to the canonical 30 nm fibre was not determined [19, 20]. This represented an example where a histone tail modification had a significant effect on chromatin structure, but was not involved in the recruitment of any protein to accomplish the structural effect in an *in vitro* system composed of purified and defined components. A genome-wide gene expression analysis also demonstrated a redundant, cumulative effect for mutations of K5, K8 and K12 of histone H4 to arginine, designed to mimic the unacetylated state of lysine. The H4 K16R mutation, on the other hand, had a transcriptional effect that was independent of the state of K5, K8 and/or K12, suggesting that acetylation played a fundamentally different functional role in these two groups of residues [21]. Also, unlike K5 and K12 of H4, which showed a strong correlation between acetylation state and gene expression, there was little correspondence between the acetylation state of H4 K16 in nucleosomes adjacent to the transcription start site, and the average transcriptional activity of genes [22]

Biochemical studies have shown that some chemical modifications of amino acid residues in peptides caused significant changes in the secondary structure of the peptides [23]. However, very little attention has been given to the possible effect of epigenetic modifications on the secondary structures of the histone tails, and the impact this may have on the association of the tails in chromatin.

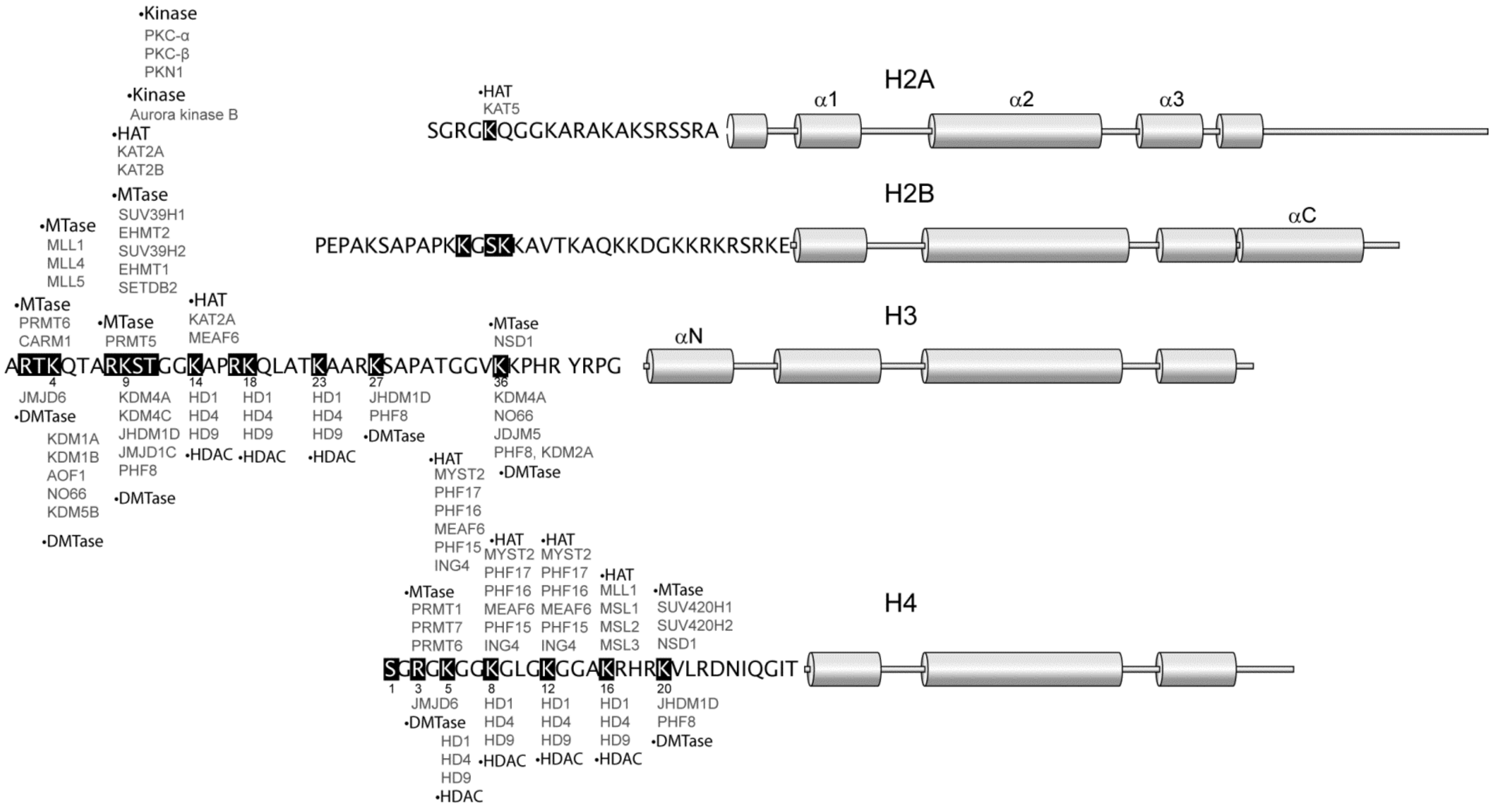


Figure 1.1 The sites of epigenetic modification in the core histone tails. The core histone tail sequences are shown, as well as the central histone folds with the additional α-N and α-C helices of H3 and H2B. The individual residues that are sites of epigenetic modification are indicated. Human enzymes in the UniProt database [24] that were annotated with gene ontology terms indicating specificity towards each of these different residues are identified (http://www.uniprot.org; accessed January 2011).

.

The fact that the histone tails appeared unstructured in X-ray crystallographic studies, most likely due to the dissociation of the tails from binding sites under conditions of moderate salt [4, 25], may have contributed to an impression that they were structurally unimportant. However, the finding that deacetylation of H4 K16 was required for full compaction of the chromatin fibre [18-20], renewed interest in a direct structural role of the histone tails in chromatin. In this Chapter we review the literature on the effect of amino acid residue modifications on the secondary structures of peptides including histone tails, citing biophysical, biochemical and *in silico* computational studies. We finally discuss how this may impact on chromatin structure and the epigenetic basis of human disease.

1.2 CHROMATIN STRUCTURE

1.2.1 A model for the 30 nm fibre

A series of images recorded of chromatin in the presence of a linker histone at increasing salt concentrations showed the systematic compaction of the chromatin through successively more condensed structures, reaching a fibre of approximately 30 nm diameter as a compaction limit [26]. This most condensed state of packaging of the nucleosomes relative to each other was termed the "30 nm fibre", which may undergo addition levels of folding into higher-order structures and helices [2].

Despite significant effort spanning many decades, there is still no agreement on the exact structural arrangement of nucleosomes in the 30 nm fibre. One model proposed a continuation of the folding of a linear array of nucleosomes into a contact helix or solenoid, where each neighbour in the solenoid was also adjacent in the linear array [27]. An alternative model suggested that the fibre was assembled in a manner that placed neighbouring nucleosomes consecutively on opposite sides of the fibre axis, to form a two-start helix, with the linker DNA running through the fibre centre [28-30]. This latter model has received strong experimental support from cross-linking [31] and X-ray crystallographic studies [32]. Many variations of these two central proposals exist, mainly based on the connectivity between nucleosomes in the fibre [33, 34].

1.2.2 Position of the H2A and H2B tails

Irrespective of differences in connectivity, all models place the site of DNA entry and exit of the nucleosome pointing inwards, towards the fibre axis [2]. This places the base of the tails at specific spatial positions within the fibre, and imposes a constraint on the possible sites of interaction of the histone tails, both within and between nucleosomes of the same fibre, as well as to different fibres. Figure 1.2 shows the maximal reach of the N-terminal tails of the core histones with the tail either fully extended, or with the full length in an α-helical conformation in a hypothetical, idealised fibre. This provides the maximal and minimal reach of the tails, respectively. It is clear that both the N-terminal tails of histones H2A and H2B have limited or no access to an adjacent nucleosome in the idealised fibre structure, but may be involved in fibre-fibre contacts, and should be accessible to *trans*-acting proteins, even in the condensed 30 nm fibre. It is therefore interesting that the sterically accessible H2B K123 is ubiquitinated as a prelude to methylation of the K4 in the histone H3 tail[14].

1.2.3 Position of the H3 tail

The histone H3 and H4 N-terminal tails appear to be able to contact distal positions within the same nucleosome as well as neighbouring nucleosomes (see Figure 1.2). The histone H3 tail is the most extensive, and exits the nucleosome between the two DNA superhelical gyres close to the pseudo-dyad axis [4]. If the H3 tail continued on its exit trajectory, it would point towards the 30 nm fibre axis, and may approach nucleosomes on the other side of the fibre (see Figure 1.2).

There exists substantial evidence that the lysine-rich tail of linker histone H1 is associated with the inter-nucleosomal linker DNA in the fibre centre [reviewed in [35]]. Since fibre compaction required histone H1 [18, 26] as well as the N-terminal tail of histone H4 [18, 36], but not the H3 tail, it appears unlikely that the H3 tail contributed to any significant partial charge neutralisation of the linker DNA in the fibre centre, or acted as a nucleosome-nucleosome stabilisation scaffold, such as the H4 tail. Thus, the possibility of the strong binding of the extended histone H3 tail to the DNA in the fibre centre appears remote. In fact, many studies suggested that the H3 tail was readily accessible in chromatin, including in an H1-containing, condensed fibre. In native H1-containing chromatin, the H3 tail remained the most susceptible to trypsin cleavage [37]. It was also shown that recombinant PCAF, which preferentially acetylated K14 of H3 [38], could still acetylate the H3 tail in condensed chromatin lacking H1 [39]. Furthermore, HP1 was specifically bound to the H3 tail tri-methylated at K9 in condensed heterochromatin [12]. In a silenced *MAT***a**-specific gene in

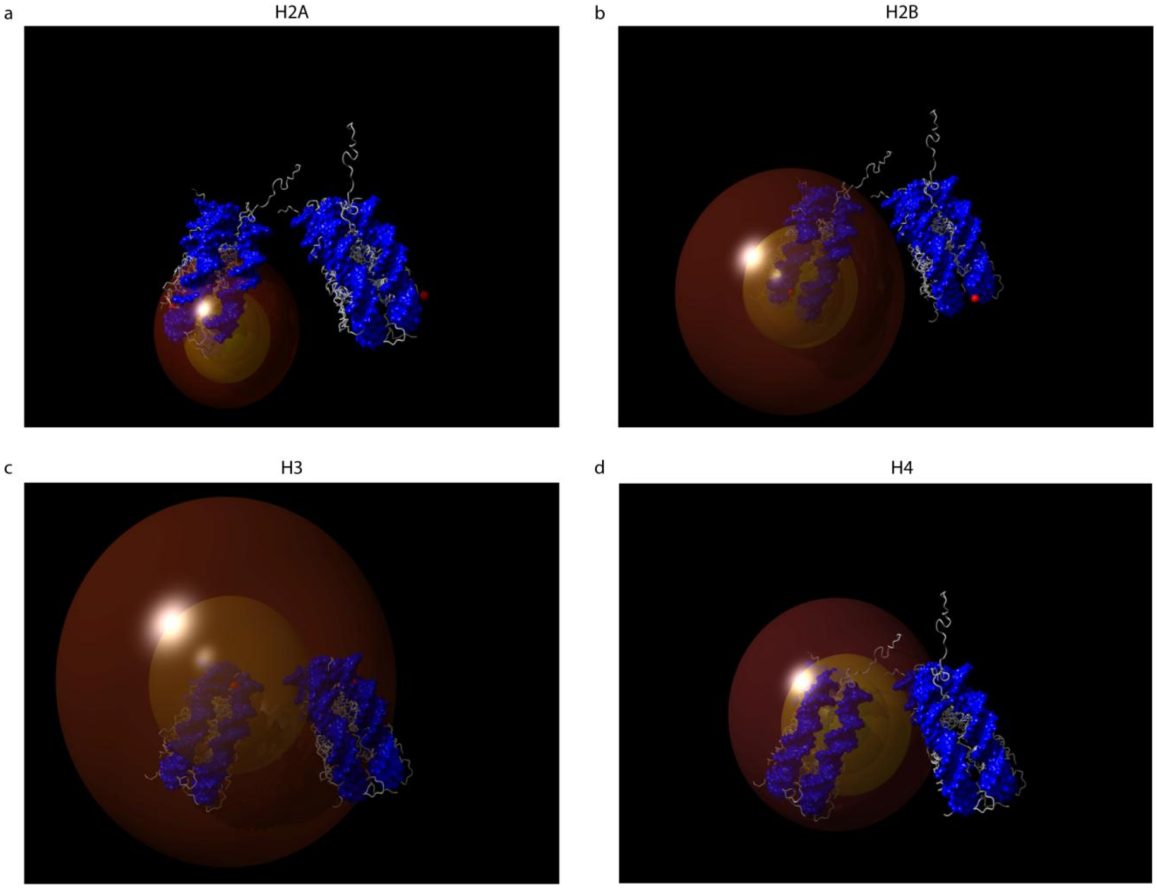


Figure 1.2 Reach of the N-terminal core histone tails in chromatin. The reach of each of the N-terminal tails of the core histones (a) H2A, (b) H2B, (c) H3 and (d) H4 is shown. The volume that can be swept out by each tail is represented by a sphere centred on the defined start of each tail [4], with the tail maximally extended (3.3 Å per residue) or with the full length of the tail in an α-helical conformation (1.5 Å per residue), represented by the outer (red) and inner (yellow) sphere in each panel, respectively. An idealised 30 nm fibre, independent of any connectivity model, is shown with the two nucleosomes rotated by 60° on the fiber axis, and with an internucleosomal rise of 20 Å. Note that the radii of the spheres assume free and unhindered rotation of the tails, which is not always physically possible. The H3 tail, for instance, would have to bend back over the nucleosomal DNA to approach the anterior side of the nucleosome on the "outside" of the fibre, a geometric path that would significantly decrease its reach in that direction.

Saccharomyces cerevisiae, Tup1, which bound at a density of two Tup1 molecules per nucleosome [40], was associated with the H3 tail in the repressive chromatin structure [41]. Also, a substituted cysteine residue, close to the tip of the H3 tail, could be cross-linked from one oligonucleosome array to another array [42].

All these studies are consistent with a histone H3 tail that is exposed for binding by proteins. Thus, the H3 tail may either continue on its exit trajectory or appear on the side of the central, crossed-linker stack, between the two nucleosome helices in the two-start helix model. Alternatively, it could follow a curved path over the nucleosomal DNA gyre, protruding into the space between two neighbouring nucleosomes in the 30 nm fiber. In either of these two possibilities, the H3 tail could be bound by sequence specific proteins, or the tail could bind to the originating or to an adjacent nucleosome.

1.2.4 Position of the H4 tail

The location of the H4 N-terminal tail on the lateral surface of the nucleosome places it in a position where it can easily be extended to contact the lateral surface of the adjacent nucleosome in the chromatin fibre (see Figure 1.2). Such a contact was, in fact, observed in the crystal structure of *Xenopus* histones reconstituted onto human α-satellite DNA repeats [4]. Clear contacts were observed to an acidic patch on the nucleosome surface, constituted by H2A E56, E61, E64, D90, E91 and E92 as well as H2B E110. The importance of this observed contact was shown by the absolute requirement for an intact H4 tail by a reconstituted fibre to condense fully in the absence of histone H1 [36]. None of the other core histone tails were required for full compaction [36, 43]. Also, nucleosome arrays reconstituted with the human histone variant H2A-Bbd [44], which lacks three glutamic acid residues that forms part of the acidic patch of H2A, did not condense to the same degree as nucleosome arrays reconstituted with H2A[45].

Interestingly, the contact of the H4 tail to the lateral surface of a nucleosome does not appear to require a single docking surface, such as the acidic patch. This was shown by a peptide comprised of residues 1-23 of the Kaposi's sarcoma-associated herpes virus latency-associated nuclear antigen (LANA). When LANA was bound to the acidic patch of a nucleosome [46], this association did not abolish the histone H4-dependent compaction of a nucleosome array [47], suggesting that the H4 tail could still bind to the adjacent nucleosome in the presence of bound LANA. This degeneracy in H4 binding was also demonstrated by the non-saturable nature of association of an H4 peptide with the nucleosome surface, suggesting that many binding sites existed for the H4 tail on the lateral nucleosome surface [47]. The binding of the LANA peptide, in contrast, was found to be saturable [47]. The H4 tail interaction was, nevertheless, sensitive to chemical modification. It was shown that 30% acetylation of the histone H4 N-terminal tail resulted in the inability of a 61-mer nucleosome array, containing linker histone H1, to fully condense *in vitro* [18]. Taken together, these studies provide a very strong argument that the N-terminal tail of histone H4 was bound to an adjacent nucleosome in the chromatin fibre, and that this interaction, which could be disrupted by acetylation of H4 K16, was essential to fully condense the chromatin into a 30 nm fibre structure.

1.3 HISTONE TAIL ASSOCIATIONS

1.3.1 N-H2A, H2A-C and N-H2B

Numerous studies have made use of chemical cross-linking to identify DNA and protein sites that can be contacted by the histone tails in the nucleosome and in a condensed fibre by using reagents that are either freely diffusible [48, 49] or immobilised [50]. The Hayes group have developed a technique where a photo-activatable azidophenacylbromide (ACP) is linked to a uniquely engineered cysteine residue [50]. The conjugated ACP group forms a reactive nitrene upon UV irradiation, cross-linking to spatially proximal DNA or protein molecules, and allowing the mapping of the contact positions of the region containing the substituted cysteine [50]. Using this technique, it was shown that the conjugated A12C of H2A cross-linked to approximately symmetrical positions 4 helical turns removed from the pseudo-dyad in reconstituted and purified nucleosome cores [50]. This is expected from the close proximity of H2A A12 to the DNA in the crystal structure [51]. In a reconstituted di-nucleosome, cross-linking of residue 12 of the H2A tail was almost exclusively within the same nucleosome [52]. The more distal portion of the H2A tail, mapped with a G2C substitution, was found to cross-link to two sites approximately 5 bp to either side of the A12C cross-linking position, in agreement with a less constrained motion of the tail further removed from the relatively immobile tail base [50]. This larger freedom of movement of the tail tip was also consistent with the cross-linking of almost 20% of H2A residue 2 to the neighbouring nucleosome in a di-nucleosome template [52]. Using a zero-length cross-linker, Bradbury and colleagues showed that the H2A C-terminal tail could be cross-linked to the DNA at the pseudo-dyad axis [53] in agreement with the exit location of this tail from the nucleosome [4]. Residue 2 of H2B was shown to participate in inter-nucleosomal contacts [52].

1.3.2 N-H3

The N-terminal tails of histone H3 made predominantly intra-nucleosomal contacts in a reconstituted di-nucleosome [52]. In a 13-mer nucleosome array, it was also found that the H3 tail was exclusively cross-linked intra-nucleosomally at 0 mM Mg^{2+}, but at higher concentrations of Mg^{2+}, where the 13-mer array became more condensed, an increase in inter-nucleosomal cross-links were observed [54]. A large proportion of the H3 tail, spanning residues 6-24, could be inter-nucleosomally cross-linked, but not the region of residue 35, close to the base of the tail, which is in agreement with the probable reach of these regions in the H3 tail [54]. Looking at the ability of the H3 tail to make contacts between different nucleosome array molecules, it was found that the entire region spanning residue 6 to 35 could be efficiently cross-linked within the same reconstituted 12-mer oligonucleosome. Long-range inter-array cross-linking was only detected at higher Mg^{2+} concentrations, ionic conditions suggested to promote self-association of the individual arrays [42]. This inter-array cross-linking efficiency was increased by the presence of H1, the binding of which may have limited unproductive associations of the H3 tail [42]. As expected, distal parts of the H3 tail could be cross-linked more efficiently to neighbouring nucleosome arrays compared to regions close to the tail base [42]. Acetylation of the H3 tail, studied in K->Q substitution mutants, required at least 4 modified residues to display a reduced inter-array cross-linking efficiency, an effect that disappeared at elevated Mg^{2+} concentrations [42]. The intra-array cross-linking did not appear sensitive to the K->Q substitutions.

1.3.3 N-H4

The Mirzabekov group showed that H18 of H4 could be cross-linked to the DNA approximately 15 bp from the pseudo-dyad in a nucleosome core particle [55]. More recently, using reconstituted nucleosome arrays, it was shown that at 0mM Mg^{2+} the H4 tail cross-linked exclusively within the originating array [56]. An increase in inter-array cross-links was observed at elevated Mg^{2+} concentrations [56]. Although an H2A-H4 cross-link, expected from binding of the H4 tail to the H2A-H2B acidic patch, was demonstrated by the simultaneous appearance of fluorescently labelled H2A and tritiated H4 in a higher mobility electrophoretic band, this band was also present in cross-linked mononucleosomes, suggesting that this interaction also occurred intra-nucleosomally [56]. This cross-link was severely diminished by the presence of the LANA peptide, previously shown to bind in the H2A-H2B acidic pocket [46, 47]. Interestingly, although tetra-acetylation of H4 reduced fibre self-association, no acetylation dependent difference in inter-fibre cross-linking efficiency was detectable [56]. In the presence of H1, however, inter-fibre cross-linking was enhanced, and a clear decrease was detected with tetra-acetylated H4 tail [56].

1.4 HISTONE TAIL STRUCTURE

Many different techniques have been used to study the structure of the histone N-terminal tails. These include the biophysical methods of circular dichroism (CD), nuclear magnetic resonance (NMR) and other forms of spectrometry, and computational methods including secondary structure prediction and molecular dynamics (MD).

1.4.1 Secondary structure prediction

Secondary structure predictions are often used to obtain insight into the secondary structures of proteins of unknown structure based solely on sequence, and have predictive accuracies in excess of 75% [57] that are continually being improved by algorithmic advances. The secondary structure predictions for the unmodified, major human core histones using PSIPRED [58] are shown in Figure 1.3.

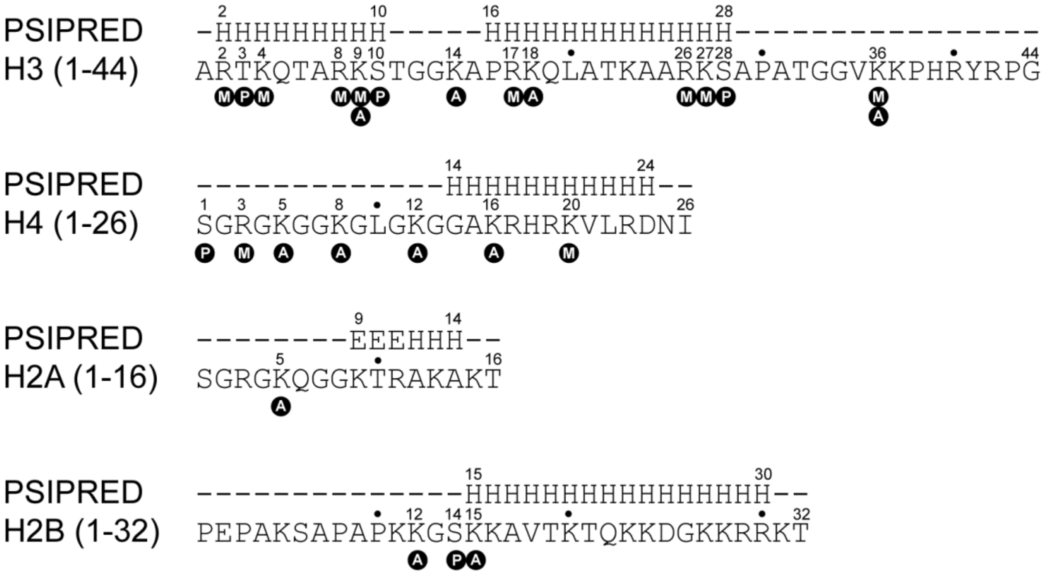

PSIPRED
H3 (1-44)
2 10 16 28
-HHHHHHHHH-----HHHHHHHHHHHHH----------------
2 3 4 8 9 10 14 17 18 • 26 27 28 • 36 • 44
ARTKQTARKSTGGKAPRKQLATKAARKSAPATGGVKKPHRYRPG
M P M M M P A A M A M M P M A
PSIPRED
H4 (1-26)
14 24
-------------HHHHHHHHHHH--
1 3 5 8 • 12 16 20 26
SGRGKGGKGLGKGGAKRHRKVLRDNI
P M A A A A M
PSIPRED
H2A (1-16)
9 14
--------EEEHHH--
5 • 16
SGRGKQGGKTRAKAKT
A
PSIPRED
H2B (1-32)
15 30
--------------HHHHHHHHHHHHHHHH--
• 12 14 15 • • 32
PEPAKSAPAPKKGSKKAVTKTQKKDGKKRRKT
A P A

Figure 1.3 Predicted secondary structures of the core histone tails. The secondary structures predicted by PSIPRED [57] are shown above the sequence of each of the four core histone tails with α-helix, β-strand, and random coil regions represented by the symbols "H", "E" and "-", respectively. Sites of epigenetic modification as well as the types of modification are indicated.

Two α-helical segments are predicted for H3 spanning 9 residues from R2 to S10, and 13 residues from P16 to S28, respectively. Interestingly, known post-translational modifications (PTMs) appear to be clustered at the predicted α-helix termini, and in both bases serine, which can be phosphorylated [59, 60], are present at the C-terminal end of the predicted α-helices.

In the case of H4 a single 11-residue α-helical segment is predicted spanning from G14 to D24. This segment contains K16, known to be required in a de-acetylated state to allow condensation of the 30 nm fibre *in vitro* [18].

A 3-residue β-strand segment from K9 to R11 followed by a 3-residue α-helical segment spanning A12 to A14 is predicted for the H2A tail, and a single 16 residue α-helical segment is predicted, stretching from K15 to R30, in the case of H2B.

It is therefore clear, based on the propensity of amino acid residues to assume defined secondary structures that the N-terminal tails of the core histones are likely to be highly structured.

1.4.2 Molecular Dynamics

MD is a molecular mechanics technique that involves the modelling of molecular systems using potential energy functions, and has been widely applied to bio-molecular systems over the last 30 years, prominently so in the study of protein folding pathways [61].

The application of MD in elucidating the structure of N–terminal histone tails has been limited, and has only been applied to the H3 and H4 tails at the time of writing. Most early work was based on coarse-grained models [62, 63] that were used to study chromatin folding, and did therefore not provide any structural detail on the histone tails. Recently there has been an increase in all-atom MD studies of the tails, and with the development of force field parameters for most of the predominant PTMs [64], more studies are likely to follow.

LaPenna and co-workers simulated a 25-residue H3 tail peptide in the presence and absence of 10 bp of DNA [65]. The peptide exhibited a wide range of structures with a high α– and 3_{10}-helical content in the presence of DNA. In agreement with the secondary structure prediction (see Figure 1.3), most of the residues, except for residues 10-15, were found in a helical structure. No β-strand content was observed. The presence of DNA increased the average helical content in the peptide, and resulted in compact, rod-like structures, despite only 4-5 bp of DNA directly interacting with the peptide [65].

Liu and Duan incorporated PTMs into their MD study of the H3 tail [66], using an 18-residue H3 variant identical to the major H3, except for 2 N-terminal glycine residues. Five PTM states were studied in the H3 peptide: unmodified, K4me2, K9me2, K4me2-K9me2, and K4Ac-K9Ac-K14Ac. The peptides preferred α-helical regions with a similar structure: a shared α-helix between K9 and T11 with the rest in an extended conformation. The singly di-methylated peptides did not differ significantly from the unmodified peptide. The doubly di-methylated peptide, however, showed a decrease in α-helical and an increase in β-strand content, although the biological relevance of a simultaneous K4 and K9 methylation is questionable. The acetylated peptide showed a decrease in helical content compared to the unmodified peptide, and exhibited a β-hairpin as the most populated structure [66]. It is thus evident that "cross-talk" between different modification groups may have a structural basis, where combinations of modifications may stabilize specific secondary structural distributions in the tail that could influence binding of the tails in chromatin.

Lins and Röthlisberger conducted MD studies on tetra- and un-acetylated 23-residue N-terminal H4 peptides [67]. The starting conformation for the two peptides was a canonical α-helix, which was found to be more stable in the tetra–acetylated peptide than in the un-acetylated peptide. A small β-hairpin was formed that spanned residues 4-12 in the tetra-acetylated peptide, which remained stable for approximately 2 ns of a 20 ns simulation [67]. Taken together with results from the previous studies, the histone tails seem able to stably accommodate secondary structure other than only α-helices. This opens the possibility that modifications to residues may be a way of changing the transition of the tails to different secondary structures on the fly, impacting on tail binding and, consequently, chromatin structure, and could thus provide a mechanism for genetic control.

In the most recent MD study, Yang and Arya investigated the effect of K16 acetylation in a 25-residue H4 tail peptide [68]. An α-helical region was formed and stabilized between residue 15 and 20 in the unmodified peptide. An α-helix was formed in the same region in the K16Ac peptide, but, in contrast to another study [67] , the helix exhibited a significantly reduced stability [68]. It is, however, important to note that the authors of the MD studies used a wide range of different simulation protocols and techniques, which makes the comparison of results between studies difficult.

Nevertheless, MD studies suggested that both H3 and H4 tail peptides preferred helix-rich structures. PTMs changed the stability of these structures, and β-strands were also observed in some cases. These studies therefore underscore a possible critical role in PTMs tipping the balance between different secondary structures in the histone tails, which may have a major impact on the function of these tails.

1.4.3 Biophysical methods

Parello and co-workers compared CD spectra obtained from a native and two selectively proteolyzed nucleosome core particles (NCP) to investigate the secondary structure of the N-terminal tails [69]. Clostripain was used to produce a "half-proteolyzed" NCP that lacked the H3 and H4 tails, and a "fully proteolyzed" NCP, that lacked all four core histone tails. The authors established that approximately 60% of the residues in the H3 and H4 tails were in an α-helical conformation, and contributed about 35% to the α-helical content in the whole NCP. It was confirmed that these contributions corresponded to the tails in the bound state in the nucleosome. The individual contributions of the H3 and H4 tails to α-helical content could, however, not be resolved. The H2A and H2B tails were found to be in a random coil conformation. A subsequent NMR study also showed that 31 residues of the H2B tail were unstructured [70].

Ausio and co-workers investigated the contribution of the histone tails to the secondary structure of the octamer, and the effect that acetylation of the tails had on this contribution[71]. The contribution of the tails to the overall α-helical content of the octamer was calculated at 17% by comparing the α-helical content of trypsin digested octamer with an undigested octamer. This value was about half of that reported by Parello and colleagues [69], and was attributed to the use of different experimental conditions. Consequently, it was shown that the overall α-helical content of the nucleosome increased by about 3% as a result of acetylation. This translated to an increase of about 17% in the α-helicity of the tails. An H4 tail peptide corresponding to residue 1-23 was isolated as mono-, di-, tri- and tetra-acetylated isomers, and analysed by CD in an aqueous solution and in trifluoroethanol (TFE), a known stabilizer of α-helices. The unmodified peptide showed an α-helical content of 17% in TFE, which increased to about 24% in the tetra-acetylated peptide in the same solvent. In the aqueous solution the isolated peptides exhibited CD spectra consistent with a random coil conformation, suggesting that the chemical environment of the histone tails played a major role in their structural conformations.

In a combined NMR and CD study Lee and co-workers also showed that a 27-residue synthetic H4 peptide had no defined structure in aqueous solution at physiological pH [72]. However, a pH dependent structural transition was observed at an acidic pH for the native peptide. None of the peptides displayed any regular secondary structures. The acetylated form of this peptide seemed insensitive to pH change, and exhibited two regions of turn-like structures at L10-G13 and R19-L22.

1.5 THE NUCLEOSOME SURFACE – POKER FACE OF CHROMATIN REGULATION

The environment inside the cell nucleus is very crowded, hence the need for chromatin compaction. However, with the demonstrated structure of the histone tails, these tails may be bound to a surface present in this environment. The histone code proposes that the histone tails exclusively bind non-histone protein complexes [11], although experimental evidence exist that showed that this was not universally so [18]. Some interesting observations have hinted at the possibility that the nucleosome surface itself may serve as binding site to histone tails and foreign elements alike [4, 46].

1.5.1 The acidic patch

The most distinguishing feature of the nucleosome surface is a small area of grouped acidic amino acids formed by the H2A-H2B dimer, called the acidic patch [4, 51]. Luger and co-workers reported that the N-terminal tail of H4 was bound to the acidic patch in co-crystal structures [4]. Certainly the most intriguing finding was the mechanism whereby KSHV exploited the acidic patch to ensure the successful migration of its viral genome to daughter cells during chromatin segregation [46]. The LANA peptide of the virus was found to bind to the acidic patch in a hairpin structure. The peptide itself reminded of a histone tail in terms of its basic composition, but lacked any lysine residues, which makes sense in the light of the preference for modification of lysine residues in the chromatin environment [46]. Recent observations also suggest that the nucleosome surface is not structurally static, but allow subtle changes facilitated by the incorporation of some but not all histone variants. The Tremethick group showed that the histone H2A variant H2A.Z promoted the compaction of chromatin at constitutive heterochromatin domains mediated by the chromatin remodeler HP1α [73]. They also showed that the histone H4 N-terminal tail was required by the HP1α to generate the highly folded chromatin fibres. In a subsequent study the same group evaluated the effect of the acidic patch on transcription *in vitro* using the H2A variant H2A.Bbd which naturally lacks the acidic patch[45]. It was found that H2A.Bbd containing nucleosomal arrays could not achieve the higher state of compaction characteristic of the 30-nm fiber, while the arrays containing canonical H2A was able to achieve this level of compaction. It was also observed that the H2A.Bbd arrays could not efficiently repress transcription. Histone mutants with partially restored acidic patches rescued efficient repression of transcription [45].

1.5.2 Molecular Docking

Although the acidic patch seems to play a definitive role in chromatin compaction as a binding receptor, it only forms a small part of the entire nucleosome surface. However, other histone variants such as the H3 variants show very little difference in residues exposed to the surface of the nucleosome, thus it is difficult to justify an experimental probe into exploring the rest of the surface, as it does not seem to show any obvious interaction sites such as a patch of negatively charged residues.

However, another computational method provides a cost and time – efficient way of exploring the entire surface, namely molecular docking. Molecular docking involves using computational techniques to predict the most likely bound complex of two molecules based on their 3D coordinates and atom composition[74, 75]. Traditionally only small molecules where docked onto protein receptors[75]. To save computational time, simple representations of the receptor and ligand were used together with elementary predicting and scoring functions. Protein – peptide docking, however, requires a more complex way of representing, predicting and scoring bound complexes because of the inherent degrees of freedom in even the smallest of peptides[76]. While the field of protein – peptide docking is still in its infancy, it is, nonetheless, an efficient and useful tool when used with appropriate experimental backup[77].

There has thus far been only one attempt to use molecular docking in studying the nucleosome surface as a potential docking receptor. Yang and Arya followed their previously mentioned Molecular Dynamics experiments of the H4 N-terminal tail up with docking fragments of the tail structures obtained onto the acidic patch [68]. Fragments of 8 residues (16 – 23) were used and 4 docking experiments were performed. An unmodified fragment and a fragment with lysine 16 acetylated were constrained to an α – helix found during the MD experiments, and were subsequently docked to the nucleosome surface. Next, the unconstrained fragments of both unmodified and modified peptides were docked. It was found that the α – helical fragments bound more favourably to the acidic patch than the unconstrained peptides, and that acetylation of K16 disrupted binding to the acidic patch [68].

1.6 HISTONE TAILS AND HUMAN HEALTH

A link between chromatin and human disease is long established. In recent times thousands of studies have been published reporting on the role of epigenetics in human disease. This role is varied and fundamental. Epigenetics was shown to be involved in development, trans-generational inheritance, memory formation, psychiatric disorders, autism spectrum disorders, carcinogenesis, cardiovascular diseases and a slew of heritable diseases including Fragile X syndrome, Friedreich's ataxia, Machado-Joseph disease, spinocerebellar ataxia, Huntington's disease and myotonic dystrophy, to provide but a significantly truncated representative list. Epigenetics have also been implicated in longevity in model eukaryotic organisms [78]. Many excellent reviews have recently appeared on epigenetics and human health [79-81]. Because of the extensive role of epigenetics in human disease, modulators of epigenetic modifications suitable for therapy have become pharmacologically highly prized [10]. A multitude of modifiers, including deactylase and demethylate inhibitors, are currently in various phases of clinical trials, and many show extremely promising results.

Many of the epigenetic therapeutic agents direct a change in gene expression level of numerous genes, where miss-expression is associated with a diseased state. The precise mechanism whereby the epigenetic modification alters gene expression level is often not fully understood. Some modifiers are now known to induce structural transitions in the core histone tails. For instance, the binding of Ni^{2+} to the sequence 15-AKRHRK-20 in the tail of H4, showed a drastic structural shift in the conformation of the peptide [82]. The binding of Ni^{2+} to a 22-residue H4 tail peptide had the same effect as acetylation on the α-helical content of the peptide [83]. This is an interesting observation since Ni is a known carcinogen which seems to act on the epigenetic level. This suggests that the epigenetic link between some human diseases and chromatin may not simply be the chemical modifications of the core histone tails that subsequently act as binding surfaces for transcription-related enzymes, but may also occur due to changes in the stable secondary structures of the histone tails which may impact not only on transcription, but also other genetic processes of the DNA molecule.

1.7 CONCLUSIONS AND INTRODUCTION TO CURRENT STUDY

There is significant evidence that the core histone tails are partially structured [69, 71], and that they are involved in intra- and inter-nucleosomal as well as in inter-fibre contacts [42, 54, 56]. It seems likely that the H4 tail binds to the lateral surface of an adjacent nucleosome in chromatin [4], and may act as a molecular tether, stabilising the architecture of the 30 nm fibre [18, 36]. It is further known that acetylation of H4 K16 abolished formation of the 30nm fibre [18]. Although this may simply involve a reduced electrostatic attraction between the acidic surface and the acetylated lysine residue, it is also possible that acetylation may disrupt secondary structures required for docking to the acidic patch or to sites in its vicinity. Alternatively, acetylation may stabilise an extended α-helix, diminishing the reach of the H4 tail, and limiting contact to the adjacent nucleosome.

Although no H3 mediated inter-nucleosome contacts were seen in X-ray crystallographic studies, this tail was, nevertheless, shown to bind intra-nucleosomally as well as between fibres [42, 54]. The predicted presence of two α-helices, demarcated by clusters of sites targeted for epigenetic modification, appears intriguing. Although, clearly, the recognition and binding of specific protein domains such as chromo and bromo domains to methylated and acetylated lysine residues are well established, and recruit proteins that serve crucial biochemical functions, the cross-linking data suggests that the H3 as well as the H2A and H2B tails are also involved in binding to DNA and/or protein surfaces in chromatin [42, 50, 52]. The binding of chromatin-associated proteins and enzymes to the histone tails may therefore only reflect a part of the functionality of the tails, which may also make a direct structural contribution to chromatin organization. One may therefore speculate that specific PTMs, stabilizing a specific distribution of secondary structures, are required for binding of the tail in chromatin. Removal of these PTMs may destabilise the structure, disrupt binding, and allow subsequent association of other regulatory proteins with the released tail. Conversely, specific PTMs may favour defined structures that allow an exact binding in chromatin, which may then provide a combined molecular surface that is recognised and bound by other regulatory factors. It is thus evident from the studies cited above that our understanding of the biochemical role of the core histone tails is incomplete, and that the tails may be multi-functional molecular entities that impact on chromatin structure and genetic function in a way that is only partially appreciated. This opens the exciting possibility of a different angle on the role of epigenetics in human disease, and the development of therapies that target histone tail structures and patterns of association as opposed to only the enzymes that are recruited by a fraction of the epigenetic marks.

Thus we will investigate the structure of the histone H3 N-terminal tail using MD simulations and the nucleosome as a potential binding site for the docking of the H3 tail using molecular docking.

1.8 REFERENCES

1. Van Holde, K. (1989). *Chromatin* Academic Press.
2. Woodcock, C. L. & Dimitrov, S. (2001). Higher-order structure of chromatin and chromosomes. *Curr. Opin. Genet. Dev.* **11**, 130-135.
3. Arents, G., Burlingame, R. W., Wang, B. C., Love, W. E. & Moudrianakis, E. N. (1991). The nucleosomal core histone octamer at 3.1 Å resolution: a tripartite protein assembly and a left-handed superhelix. *Proc. Natl. Acad. Sci. U. S. A.* **88**, 10148-10152.
4. Luger, K., Mader, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. (1997). Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* **389**, 251-260.
5. Phillips, D. M. P. (1961). Acetyl groups as N-terminal substituents in calf-thymus histones. *Biochem. J.* **80**, 40P.
6. Phillips, D. M. P. (1963). The presence of acetyl groups in histones. *Biochem. J.* **87**, 258-263.
7. Murray, K. (1964). The occurrence of epsilon-N-methyl lysine in histones. *Biochemistry.* **3**, 10-15.
8. Allfrey, V. G., Faulkner, R. R. & Mirsky, A. E. (1964). Acetylation and methylation of histones and their possible role in the regulation of RNA synthesis. *Proc. Natl. Acad. Sci. U. S. A.* **51**, 786-794.
9. Kouzarides, T. (2007). Chromatin modifications and their function. *Cell.* **128**, 693-705.
10. Kundu, T. T. (2007). Small molecular modulators in epigenetics: implications in gene expression and therapeutics. In *Chromatin and disease* (Kundu, T. T. & Dasgupta, D., eds), pp. 399-430, Springer, New York.

11. Strahl, B. D. & Allis, C. D. (2000). The language of covalent histone modifications. *Nature* **403**, 41-45.

12. Bannister, A. J., Zegerman, P., Partridge, J. F., Miska, E. A., Thomas, J. O., Allshire, R. C. & Kouzarides, T. (2001). Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. *Nature* **410**, 120-124.

13. Forneris, F., Binda, C., Vanoni, M. A., Battaglioli, E. & Mattevi, A. (2005). Human histone demethylase LSD1 reads the histone code. *J. Biol. Chem.* **280**, 41360-41365.

14. Zheng, S., Wyrick, J. J. & Reese, J. C. (2010). Novel trans-tail regulation of H2B ubiquitylation and H3K4 methylation by the N terminus of histone H2A. *Mol. Cell Biol.* **30**, 3635-3645.

15. Dover, J., Schneider, J., Tawiah-Boateng, M. A., Wood, A., Dean, K., Johnston, M. & Shilatifard, A. (2002). Methylation of histone H3 by COMPASS requires ubiquitination of histone H2B by Rad6. *J. Biol. Chem.* **277**, 28368-28371.

16. Briggs, S. D., Xiao, T., Sun, Z. W., Caldwell, J. A., Shabanowitz, J., Hunt, D. F., Allis, C. D. & Strahl, B. D. (2002). Gene silencing: trans-histone regulatory pathway in chromatin. *Nature.* **418**, 498.

17. Bartke, T., Vermeulen, M., Xhemalce, B., Robson, S. C., Mann, M. & Kouzarides, T. (2010). Nucleosome-Interacting Proteins Regulated by DNA and Histone Methylation., pp. 470-484.

18. Robinson, P. J., An, W., Routh, A., Martino, F., Chapman, L., Roeder, R. G. & Rhodes, D. (2008). 30 nm chromatin fibre decompaction requires both H4-K16 acetylation and linker histone eviction. *J. Mol. Biol.* **381**, 816-825.

19. Shogren-Knaak, M., Ishii, H., Sun, J. M., Pazin, M. J., Davie, J. R. & Peterson, C. L. (2006). Histone H4-K16 acetylation controls chromatin structure and protein interactions. *Science.* **311**, 844-847.

20. Allahverdi, A., Yang, R., Korolev, N., Fan, Y., Davey, C. A., Liu, C. F. & Nordenskiöld, L. (2011). The effects of histone H4 tail acetylations on cation-induced chromatin folding and self-association. *Nucleic Acids Research* **39**, 1680-1691.

21. Dion, M. F., Altschuler, S. J., Wu, L. F. & Rando, O. J. (2005). Genomic characterization reveals a simple histone H4 acetylation code. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 5501-5506.

22. Liu, C. L., Kaplan, T., Kim, M., Buratowski, S., Schreiber, S. L., Friedman, N. & Rando, O. J. (2005). Single-nucleosome mapping of histone modifications in *S. cerevisiae*. *PLoS. Biol.* **3**, e328.

23. Wang, X., He, C., Moore, S. C. & Ausio, J. (2001). Effects of histone acetylation on the solubility and folding of the chromatin fiber. *J. Biol. Chem.* **276**, 12764-12768.

24. Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Natale, D. A., O'Donovan, C., Redaschi, N. & Yeh, L. S. (2004). UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.* **32**, D115-D119.

25. Walker, I. O. (1984). Differential dissociation of histone tails from core chromatin. *Biochemistry.* **23**, 5622-5628.

26. Thoma, F., Koller, T. & Klug, A. (1979). Involvement of histone H1 in the organization of the nucleosome and of the salt-dependent superstructures of chromatin. *J. Cell Biol.* **83**, 403-427.

27. Finch, J. T. & Klug, A. (1976). Solenoidal model for superstructure in chromatin. *Proc. Natl. Acad. Sci. U. S. A.* **73**, 1897-1901.

28. Worcel, A., Strogatz, S. & Riley, D. (1981). Structure of chromatin and the linking number of DNA. *Proc. Natl. Acad. Sci. U. S. A.* **78**, 1461-1465.

29. Woodcock, C. L., Frado, L. L. & Rattner, J. B. (1984). The higher-order structure of chromatin: evidence for a helical ribbon arrangement. *J. Cell Biol.* **99**, 42-52.

30. Williams, S. P., Athey, B. D., Muglia, L. J., Schappe, R. S., Gough, A. H. & Langmore, J. P. (1986). Chromatin fibers are left-handed double helices with diameter and mass per unit length that depend on linker length. *Biophys. J.* **49**, 233-248.

31. Dorigo, B., Schalch, T., Kulangara, A., Duda, S., Schroeder, R. R. & Richmond, T. J. (2004). Nucleosome arrays reveal the two-start organization of the chromatin fiber. *Science* **306**, 1571-1573.

32. Schalch, T., Duda, S., Sargent, D. F. & Richmond, T. J. (2005). X-ray structure of a tetranucleosome and its implications for the chromatin fibre. *Nature* **436**, 138-141.

33. Daban, J. R. & Bermudez, A. (1998). Interdigitated solenoid model for compact chromatin fibers. *Biochemistry.* **37**, 4299-4304.

34. Robinson, P. J., Fairall, L., Huynh, V. A. & Rhodes, D. (2006). EM measurements define the dimensions of the "30-nm" chromatin fiber: evidence for a compact, interdigitated structure. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 6506-6511.

35. Caterino, T. L. & Hayes, J. J. (2011). Structure of the H1 C-terminal domain and function in chromatin condensation. *Biochem. Cell Biol.* **89**, 35-44.

36. Dorigo, B., Schalch, T., Bystricky, K. & Richmond, T. J. (2003). Chromatin fiber folding: requirement for the histone H4 N-terminal tail. *J. Mol. Biol.* **327**, 85-96.

37. Harborne, N. & Allan, J. (1983). Modulation of the relative trypsin sensitivities of the core histone 'tails'. *FEBS Lett.* **155**, 88-92.

38. Schiltz, R. L., Mizzen, C. A., Vassilev, A., Cook, R. G., Allis, C. D. & Nakatani, Y. (1999). Overlapping but distinct patterns of histone acetylation by the human coactivators p300 and PCAF within nucleosomal substrates. *J. Biol. Chem.* **274**, 1189-1192.

39. Herrera, J. E., Schiltz, R. L. & Bustin, M. (2000). The accessibility of histone H3 tails in chromatin modulates their acetylation by P300/CBP-associated factor. *J. Biol. Chem.* **275**, 12994-12999.

40. Ducker, C. E. & Simpson, R. T. (2000). The organized chromatin domain of the repressed yeast a cell-specific gene *STE6* contains two molecules of the corepressor Tup1p per nucleosome. *EMBO J.* **19**, 400-409.

41. Edmondson, D. G., Smith, M. M. & Roth, S. Y. (1996). Repression domain of the yeast global repressor Tup1 interacts directly with histones H3 and H4. *Genes Dev.* **10**, 1247-1259.

42. Kan, P. Y., Lu, X., Hansen, J. C. & Hayes, J. J. (2007). The H3 tail domain participates in multiple interactions during folding and self-association of nucleosome arrays. *Mol. Cell Biol.* **27**, 2084-2091.

43. Moore, S. C. & Ausio, J. (1997). Major role of the histones H3-H4 in the folding of the chromatin fiber. *Biochem. Biophys. Res. Commun.* **230**, 136-139.

44. Chadwick, B. P. & Willard, H. F. (2001). A novel chromatin protein, distantly related to histone H2A, is largely excluded from the inactive X chromosome. *J. Cell Biol.* **152**, 375-384.

45. Zhou, J., Fan, J. Y., Rangasamy, D. & Tremethick, D. J. (2007). The nucleosome surface regulates chromatin compaction and couples it with transcriptional repression. *Nat. Struct. Mol. Biol.* **14**, 1070-1076.

46. Barbera, A. J., Chodaparambil, J. V., Kelley-Clarke, B., Joukov, V., Walter, J. C., Luger, K. & Kaye, K. M. (2006). The nucleosomal surface as a docking station for Kaposi's sarcoma herpesvirus LANA. *Science* **311**, 856-861.

47. Chodaparambil, J. V., Barbera, A. J., Lu, X., Kaye, K. M., Hansen, J. C. & Luger, K. (2007). A charged and contoured surface on the nucleosome regulates chromatin compaction. *Nat. Struct. Mol. Biol.* **14**, 1105-1107.

48. Sperling, J. & Sperling, R. (1978). Photochemical cross-linking of histones to DNA nucleosomes. *Nucleic Acids Res.* **5**, 2755-2773.

49. Jackson, V. (1999). Formaldehyde cross-linking for studying nucleosomal dynamics. *Methods.* **17**, 125-139.

50. Lee, K. M. & Hayes, J. J. (1997). The N-terminal tail of histone H2A binds to two distinct sites within the nucleosome core. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 8959-8964.

51. Davey, C. A., Sargent, D. F., Luger, K., Maeder, A. W. & Richmond, T. J. (2002). Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 Å resolution. *J. Mol. Biol.* **319**, 1097-1113.

52. Zheng, C. & Hayes, J. J. (2003). Intra- and inter-nucleosomal protein-DNA interactions of the core histone tail domains in a model system. *J. Biol. Chem.* **278**, 24217-24224.

53. Usachenko, S. I., Bavykin, S. G., Gavin, I. M. & Bradbury, E. M. (1994). Rearrangement of the histone H2A C-terminal domain in the nucleosome. *Proc. Natl. Acad. Sci. U. S. A.* **91**, 6845-6849.

54. Zheng, C., Lu, X., Hansen, J. C. & Hayes, J. J. (2005). Salt-dependent intra- and internucleosomal interactions of the H3 tail domain in a model oligonucleosomal array. *J. Biol. Chem.* **280**, 33552-33557.

55. Ebralidse, K. K., Grachev, S. A. & Mirzabekov, A. D. (1988). A highly basic histone H4 domain bound to the sharply bent region of nucleosomal DNA. *Nature.* **331**, 365-367.

56. Kan, P. Y., Caterino, T. L. & Hayes, J. J. (2009). The H4 tail domain participates in intra- and internucleosome interactions with protein and DNA during folding and oligomerization of nucleosome arrays. *Mol. Cell Biol.* **29**, 538-546.

57. McGuffin, L. J., Bryson, K. & Jones, D. T. (2000). The PSIPRED protein structure prediction server. *Bioinformatics.* **16**, 404-405.

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

59. Wei, Y., Mizzen, C. A., Cook, R. G., Gorovsky, M. A. & Allis, C. D. (1998). Phosphorylation of histone H3 at serine 10 is correlated with chromosome condensation during mitosis and meiosis in *Tetrahymena*. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 7480-7484.

60. Goto, H., Tomono, Y., Ajiro, K., Kosako, H., Fujita, M., Sakurai, M., Okawa, K., Iwamatsu, A., Okigaki, T., Takahashi, T. & Inagaki, M. (1999). Identification of a novel phosphorylation site on histone H3 coupled with mitotic chromosome condensation. *J. Biol. Chem.* **274**, 25543-25549.

61. Adcock, S. A. & McCammon, J. A. (2006). Molecular dynamics: survey of methods for simulating the activity of proteins. *Chem. Rev.* **106**, 1589-1615.

62. Arya, G. & Schlick, T. (2006). Role of histone tails in chromatin folding revealed by a mesoscopic oligonucleosome model. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 16236-16241.

63. Korolev, N., Lyubartsev, A. P. & Nordenski÷ld, L. (2006). Computer Modeling Demonstrates that Electrostatic Attraction of Nucleosomal DNA is Mediated by Histone Tails., pp. 4305-4316.

64. Grauffel, C., Stote, R. H. & Dejaegere, A. (2010). Force field parameters for the simulation of modified histone tails. *J. Comput. Chem.* **31**, 2434-2451.

65. LaPenna, G., Furlan, S. & Perico, A. (2006). Modeling H3 histone N-terminal tail and linker DNA interactions. *Biopolymers* **83**, 135-147.

66. Liu, H. & Duan, Y. (2008). Effects of post-translational modifications on the structure and dynamics of histone H3 N-terminal peptide. *Biophys. J.* **94**, 4579-4585.

67. Lins, R. D. & Röthlisberger, U. (2006). Influence of Long-Range Electrostatic Treatments on the Folding of the N-Terminal H4 Histone Tail Peptide. *J. Chem. Theory Comput.* **2**, 246-250.

68. Yang, D. & Arya, G. (2011). Structure and binding of the H4 histone tail and the effects of lysine 16 acetylation. *Phys. Chem. Chem. Phys.* **13**, 2911-2921.

69. Banères, J. L., Martin, A. & Parello, J. (1997). The N tails of histones H3 and H4 adopt a highly structured conformation in the nucleosome. *J. Mol. Biol.* **273**, 503-508.

70. Nunes, A. M., Zavitsanos, K., Del, C. R., Malandrinos, G. & Hadjiliadis, N. (2009). Interaction of histone H2B (fragment 63-93) with Ni(ii). An NMR study. *Dalton Trans.* **11**, 1904-1913.

71. Wang, X., Moore, S. C., Laszckzak, M. & Ausio, J. (2000). Acetylation increases the alpha-helical content of the histone tails of the nucleosome. *J. Biol. Chem.* **275**, 35013-35020.

72. Bang, E., Lee, C. H., Yoon, J. B., Lee, D. W. & Lee, W. (2001). Solution structures of the N-terminal domain of histone H4. *J. Pept. Res.* **58**, 389-398.

73. Fan, J. Y., Rangasamy, D., Luger, K. & Tremethick, D. J. (2004). H2A.Z Alters the Nucleosome Surface to Promote HP1α-Mediated Chromatin Fiber Folding., pp. 655-661.

74. Morris, G. M. & Lim-Wilby, M. (2008). Molecular Docking. In *Molecular Modeling of Proteins* (Kukol, A., ed), pp. 365-382, Humana Press, Totowa.

75. Halperin, I., Ma, B., Wolfson, H. & Nussinov, R. (2002). Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins* **47**, 409-443.

76. Andrusier, N., Mashiach, E., Nussinov, R. & Wolfson, H. J. (2008). Principles of flexible protein – protein docking. *Proteins* **73**, 271-289.

77. Wu, H., Min, J., Lunin, V. V., Antoshenko, T., Dombrovski, L., Zeng, H., Allali-Hassani, A., Campagna-Slater, V. r., Vedadi, M., Arrowsmith, C. H., Plotnikov, A. N. & Schapira, M. (2010). Structural Biology of Human H3K9 Methyltransferases. *PLoS ONE* **5**, e8570.

78. Dang, W., Steffen, K. K., Perry, R., Dorsey, J. A., Johnson, F. B., Shilatifard, A., Kaeberlein, M., Kennedy, B. K. & Berger, S. L. (2009). Histone H4 lysine 16 acetylation regulates cellular lifespan. *Nature* **459**, 802-807.

79. Watanabe, Y. & Maekawa, M. (2010). Methylation of DNA in cancer. *Adv. Clin. Chem.* **52**, 145-167.

80. Luco, R. F., Allo, M., Schor, I. E., Kornblihtt, A. R. & Misteli, T. (2011). Epigenetics in alternative pre-mRNA splicing. *Cell.* **144(1)**, 16-26.

81. Kurdistani, S. K. (2011). Histone modifications in cancer biology and prognosis. *Prog. Drug Res.* **67**, 91-106.

82. Zoroddu, M. A., Kowalik-Jankowska, T., Kozlowski, H., Molinari, H., Salnikow, K., Broday, L. & Costa, M. (2000). Interaction of Ni(II) and Cu(II) with a metal binding sequence of histone H4: AKRHRK, a model of the H4 tail. *Biochim. Biophys. Acta.* **1475**, 163-168.

83. Zoroddu, M. A., Medici, S. & Peana, M. (2009). Copper and nickel binding in multi-histidinic peptide fragments. *J. Inorg. Biochem.* **103**, 1214-1220.

CHAPTER 2

Development of tools for the analysis of Molecular Dynamics Trajectories in YASARA

2.1 INTRODUCTION

YASARA (Yet Another Scientific Artificial Reality Application) is a proprietary software application used to perform a wide variety of structural biology related tasks [1]. These tasks include, but are not limited to, explicit MD simulations, molecular docking and homology modeling. It features a user friendly GUI, its own scripting language: Yanaconda, limited parallelization and integrated visualization. Using this application thus eliminates the need to use a wide variety of different software for a given experiment.

However, YASARA is limited in terms of MD trajectory analysis tools, especially for large trajectories. For example, it does not feature the implementation of any clustering algorithm for clustering structures in a MD trajectory. YASARA does support extension through a Python wrapper. Python (www.python.org) is a scripting language based the C programming language, and is used extensively in the bioinformatics [2]. Thus, it is possible to use YASARA functions in custom Python scripts to analyze MD trajectories and perform other molecular modeling tasks.

Scripting in Python is preferred over scripting in Yanaconda, since Python handles file I/O operations more easily and supports object orientation.

The following tools were required for the current study: (i) tools for the analysis of MD trajectories produced by YASARA and (ii) tools for the efficient management and storage of data produced by the trajectory analysis tools.

The implementation of these tools will be discussed broadly and will not include utility applications that were created. All the source code for the scripts used in the study can be found on the CD included with this document or at http://cbio.ufs.ac.za

All scripts were designed with YASARA Structure version 11.6.16 as the last tested version.

2.2 AN OBJECT ORIENTATED MODEL FOR CONDUCTING MD TRAJECTORY ANALYSIS USING PYTHON AND YASARA

2.2.1 MD Trajectories in YASARA

YASARA natively stores trajectories in multiple *.sim files: binary files containing the atom positions and velocities of the simulated system in single precision. These files are known as snapshots and correspond to the state of the system at a particular time point in the simulation. When a snapshot is loaded, the stored atom velocities and positions are restored to the atoms and properties at the specified simulation time of the system. This means that the atoms and the simulation cell should already be present. The atoms and the simulation cell are usually stored in a *.sce file, which is loaded first. Once loaded, analysis can be performed on the system at the specific time in the simulation. For instance: the secondary structure of residues in a simulated protein can be determined at the time the snapshot was taken.

2.2.2 The trajectory object: Traj()

The trajectory object represents an MD trajectory generated by YASARA. The object contains the location of the starting *.sce file and the *.sim snapshot files. Additionally, it contains a dictionary with the values of all the parameters used to set up the simulation, and can be used to set up the simulation environment in YASARA, using the provided function. The trajectory object contains a list of snapshot objects as well as a list containing method objects. The trajectory also contains an analysis function which analyses each snapshot object with each method object. It can also save the results of all analysis done to disk. Further details about the analysis and save function are given with the method object description.

2.2.3 The snapshot object: Snapshot()

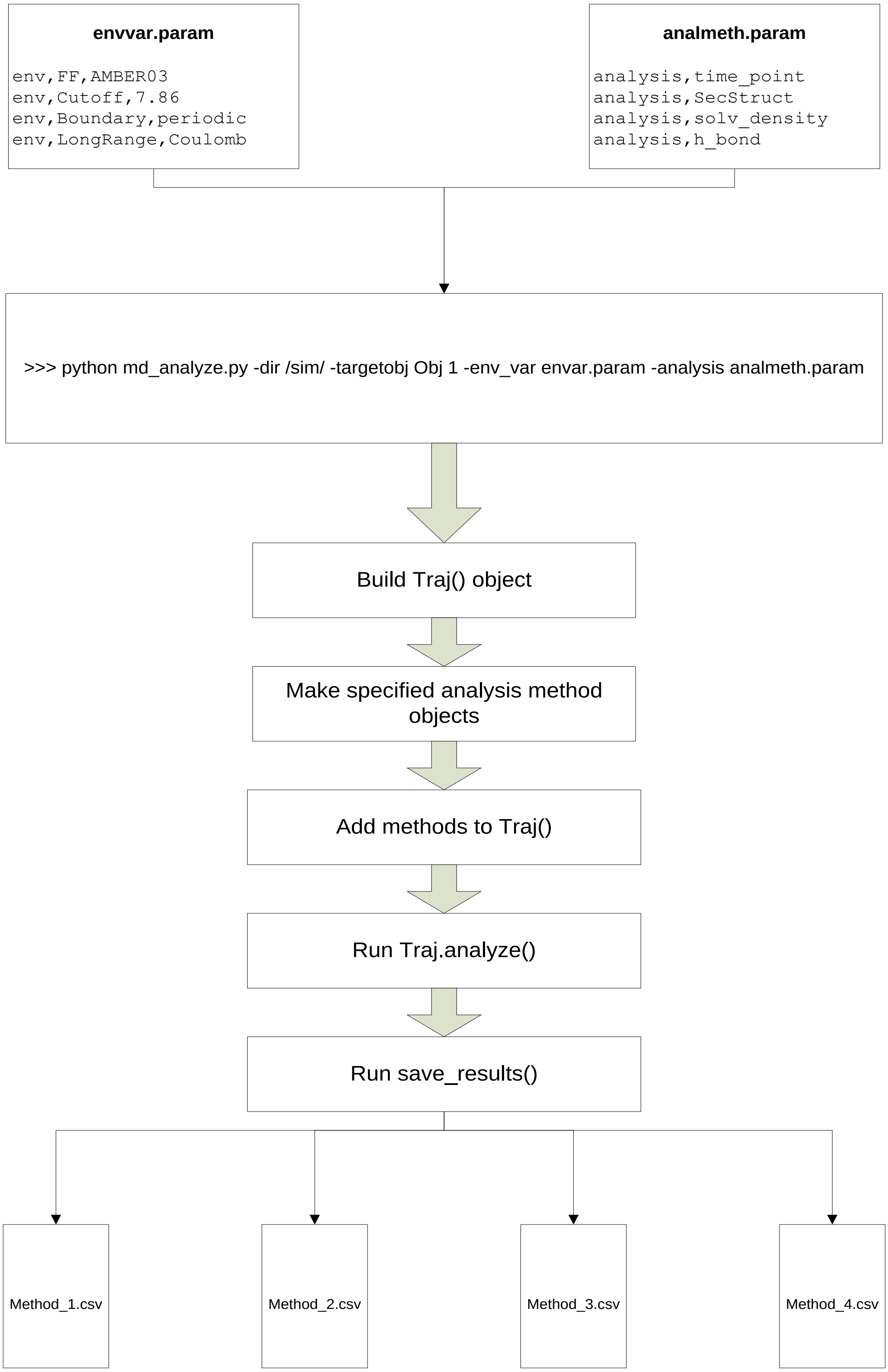
The snapshot object represents a single snapshot in a MD trajectory generated by YASARA. It contains the time point which the snapshot occupies in the trajectory, as well as the filename of the stored *.sim file. The snapshot object can load the snapshot into YASARA and also stores results from any analysis done on the snapshot.

2.2.4 The analysis object: Analysis()

The analysis class is an abstract base class. This means that for every analysis method a new class can be created which will yield an object of type Analysis. Each derived class is required to have 2 functions: DoAnalysis() and DumpData(). DoAnalysis() contains the actual analysis method which is usually performed on a given snapshot and saved in the snapshot object. DumpData() is required for saving all the stored analysis data in the trajectory object (which is subsequently stored in each snapshot object) to CSV text files. The significance of this abstract base class is that the model becomes flexible in terms of the addition and implementation of new analysis methods.

2.2.5 Implementation of the model: md_analyze.py

Implementation of the above objects for the analysis was carried out using the script md_analyze.py. Figure 2.1 describes the workflow of the script. The script is executed with 2 parameter files: the first contains the parameters used to set up the simulation environment in YASARA, and the second contains the list of analysis methods to be applied to the trajectory. The script creates a trajectory object and the analysis methods specified. The analysis methods are subsequently added to the trajectory object. The traj() method analyze is used to load the trajectory snapshots from disk and to subsequently apply each analysis method to each snapshot. The results are saved to separate files corresponding to each analysis method by using the method save_results() in the trajectory object.   The full source code can found on the included CD under the directory "md_analysis".



```
env,FF,AMBER03
env,Cutoff,7.86
env,Boundary,periodic
env,LongRange,Coulomb
```

```
analysis,time_point
analysis,SecStruct
analysis,solv_density
analysis,h_bond
```

```
>>> python md_analyze.py -dir /sim/ -targetobj Obj 1 -env_var envar.param -analysis analmeth.param
```

Figure 2.1 Implementation of the object orientated model for the analysis of a MD trajectory using YASARA

2.3 SIMDB : A RELATIONAL DATABASE FOR THE STORAGE OF MD TRAJECTORY ANALYSIS RESULTS

MD trajectories are a rich source of information about a system studied. Unfortunately, data management challenges increases with an increase in the amount of data available. A relational database design, simDB , was implemented to handle the management of data generated from the analysis of the MD trajectories generated in this study.

2.3.1 Database schema and implementation

The complete database schema is given in Figure 2.2. The tables will be described briefly.

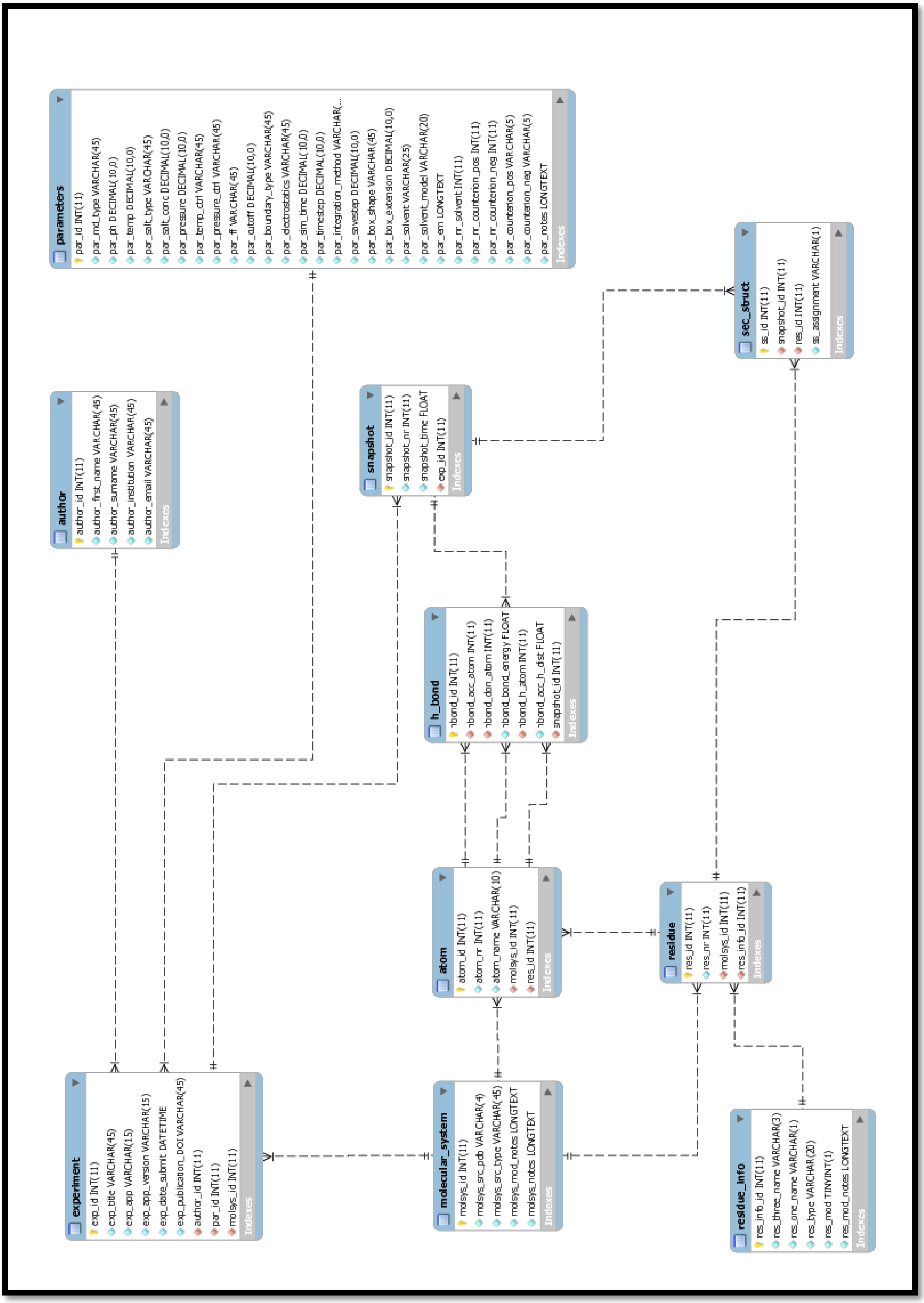
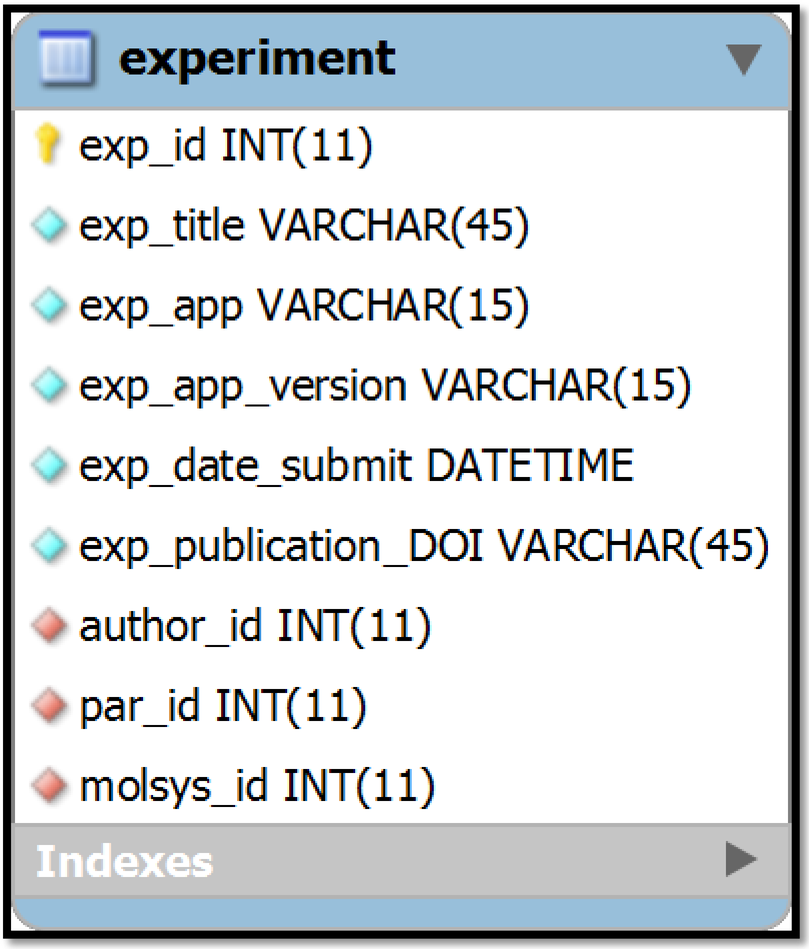


Figure 2.2 The simDB database schema. Yellow key icons denote primary keys. Light Blue diamond icons denote ordinary attributes. Light Red diamond icons denote foreign keys.

2.3.1.1 Experiment



experiment

exp_id INT(11)

exp_title VARCHAR(45)

exp_app VARCHAR(15)

exp_app_version VARCHAR(15)

exp_date_submit DATETIME

exp_publication_DOI VARCHAR(45)

author_id INT(11)

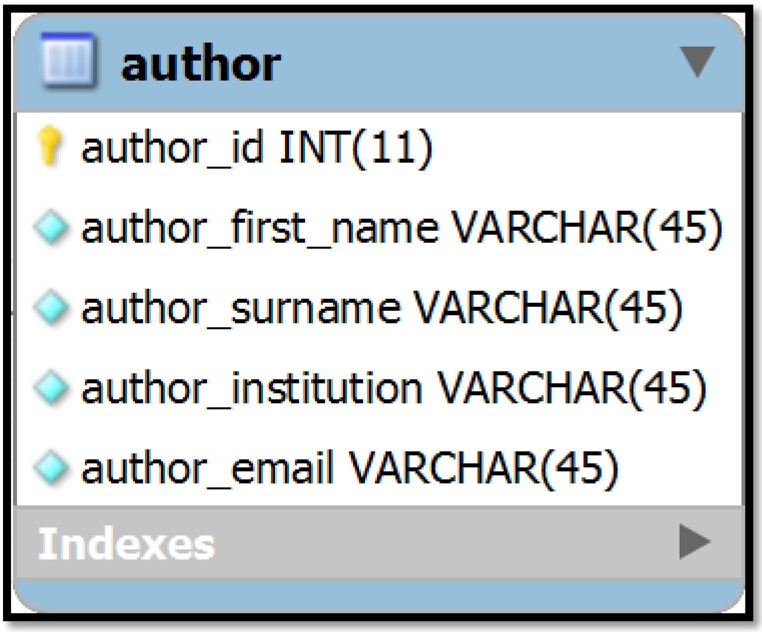
par_id INT(11)

molsys_id INT(11)

Figure 2.3 The experiment table

The experiment table is the central table in the database (Figure 2.3). It represents an MD experiment, and contains attributes describing the experiment, such as the title and the application used. An experiment has an author, a parameter set and a molecular system (the atoms in the system studied) and can only contain one of each. An experiment can also contain multiple snapshots. The primary key used for experiment is an integer value which automatically increments when a record is added. Most queries involve the use of the primary key, exp_id.

2.3.1.2 Author



author ▼

author_id INT(11)

author_first_name VARCHAR(45)

author_surname VARCHAR(45)

author_institution VARCHAR(45)

author_email VARCHAR(45)

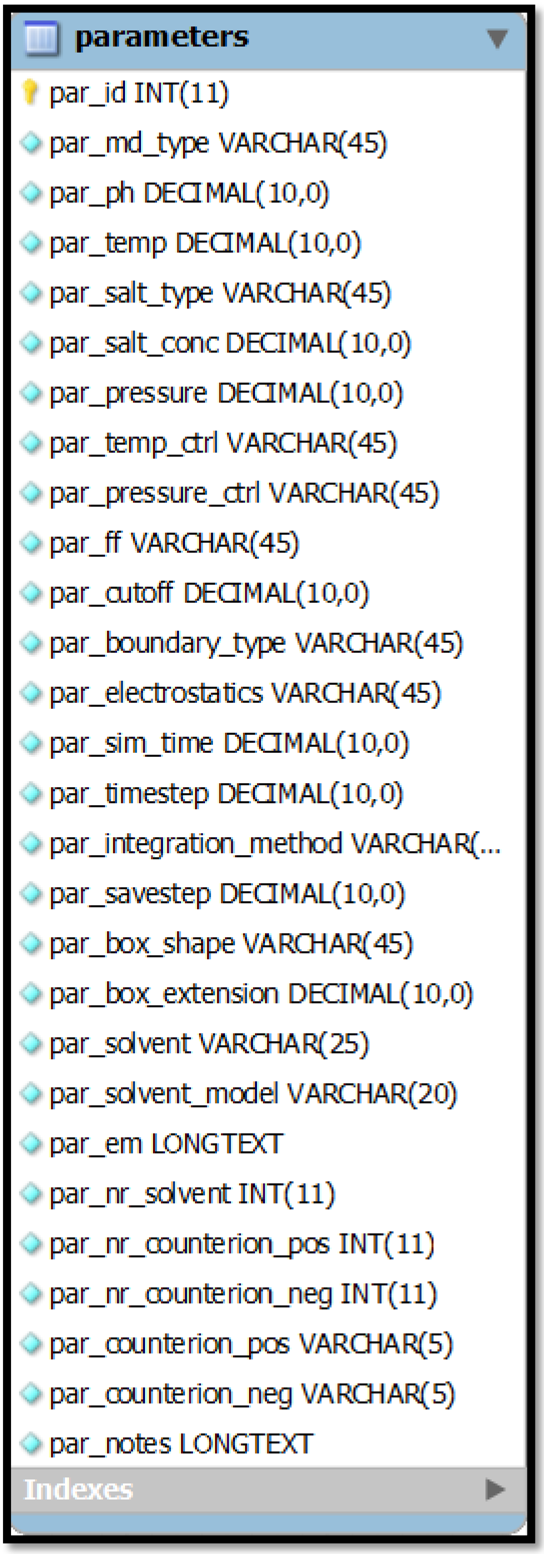
Figure 2.4 The author table

The author table contains information about the author of one or multiple experiments (Figure 2.4). For example: author_email stores the email address of the author.

As mentioned previously, one experiment can have only one author, however one author may author one or more experiments.

This table was implemented with the expectation that the current database will be expanded to be used as a repository for researchers to deposit MD trajectory analysis data for access to the broader scientific community.

2.3.1.3 Parameter



parameters

par_id INT(11)
par_md_type VARCHAR(45)
par_ph DECIMAL(10,0)
par_temp DECIMAL(10,0)
par_salt_type VARCHAR(45)
par_salt_conc DECIMAL(10,0)
par_pressure DECIMAL(10,0)
par_temp_ctrl VARCHAR(45)
par_pressure_ctrl VARCHAR(45)
par_ff VARCHAR(45)
par_cutoff DECIMAL(10,0)
par_boundary_type VARCHAR(45)
par_electrostatics VARCHAR(45)
par_sim_time DECIMAL(10,0)
par_timestep DECIMAL(10,0)
par_integration_method VARCHAR(...
par_savestep DECIMAL(10,0)
par_box_shape VARCHAR(45)
par_box_extension DECIMAL(10,0)
par_solvent VARCHAR(25)
par_solvent_model VARCHAR(20)
par_em LONGTEXT
par_nr_solvent INT(11)
par_nr_counterion_pos INT(11)
par_nr_counterion_neg INT(11)
par_counterion_pos VARCHAR(5)
par_counterion_neg VARCHAR(5)
par_notes LONGTEXT

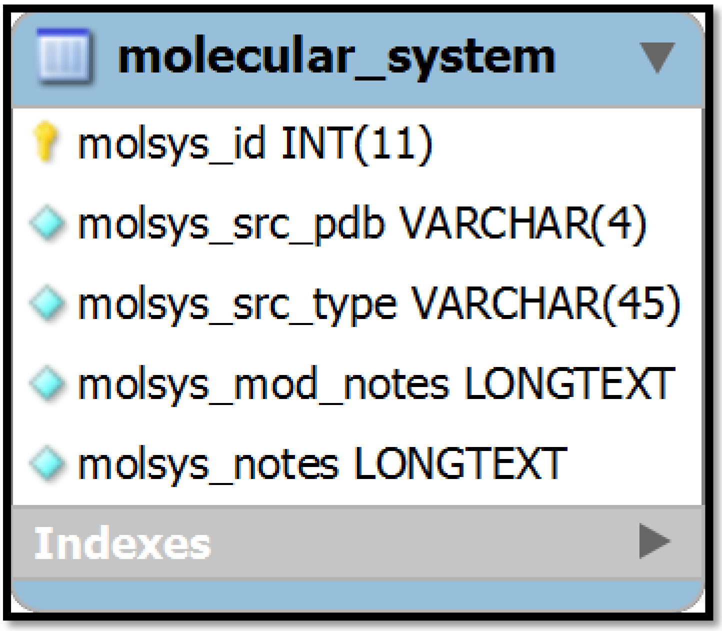
Figure 2.5 The parameter table

The parameter table represents the parameter set used to produce the MD trajectory (Figure 2.5).

For example: par_ff contains the name of the force field used in the MD simulation.

One experiment can only have one parameter set, however a parameter set can be used by multiple experiments.

2.3.1.4 Molecular system



molecular_system

molsys_id INT(11)
molsys_src_pdb VARCHAR(4)
molsys_src_type VARCHAR(45)
molsys_mod_notes LONGTEXT
molsys_notes LONGTEXT

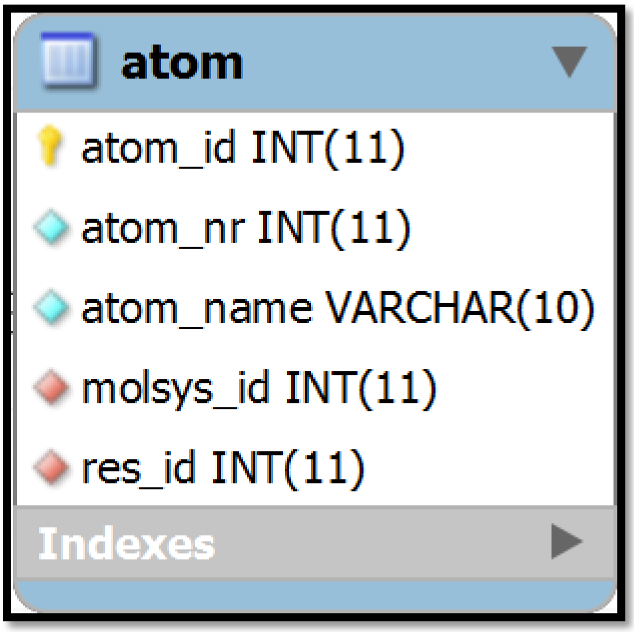
Figure 2.6 The molecular_system table

The molecular_system represents the system studied (Figure 2.6). It contains attributes such as the source protein databank code (http://www.pdb.org) [3], for example. The table also allows one to add information about modifications made to the structure. The molecular system does not contain information about the solvent and counter-ions used in the MD simulation.

An experiment can only have one molecular system, while a molecular system may be used in more than one experiment, for example in a duplicate experiment.

The molecular system will contain atoms and residues, which is described in separate tables.

2.3.1.5 Atom



atom

atom_id INT(11)

atom_nr INT(11)

atom_name VARCHAR(10)

molsys_id INT(11)

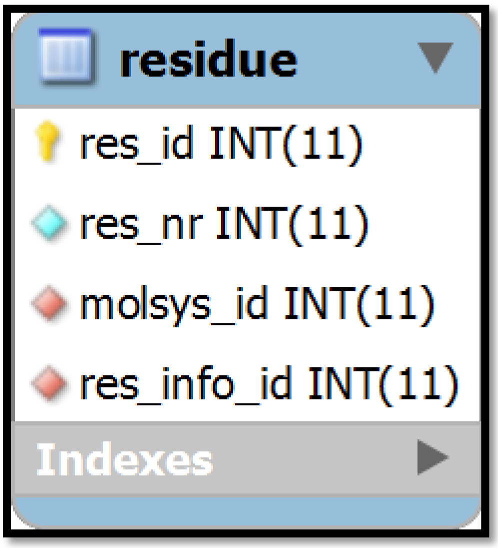
res_id INT(11)

Figure 2.7 The atom table

The atom table represents a single atom in a molecular system (Figure 2.7). It contains attributes such as the atom name and its number in the particular system, for example.

An atom can only belong to one molecular system. An atom can also belong to a residue; however it can only belong to one residue.

2.3.1.6 Residue



residue

res_id INT(11)

res_nr INT(11)

molsys_id INT(11)

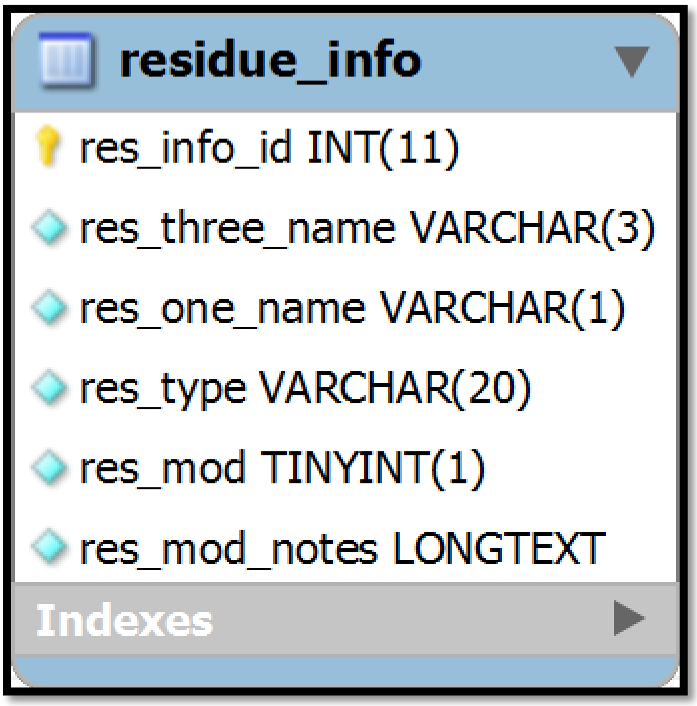
res_info_id INT(11)

Figure 8 The residue table

The residue table represents a single residue in a molecular system (Figure 2.8). The table contains the attribute res_nr to identify the specific residue in the molecular system, however most information about the residue is stored in the table residue_info.

A residue can only belong to one molecular system and can only have one set of descriptive data.

2.3.1.7 Residue info



residue_info

res_info_id INT(11)

res_three_name VARCHAR(3)

res_one_name VARCHAR(1)

res_type VARCHAR(20)

res_mod TINYINT(1)

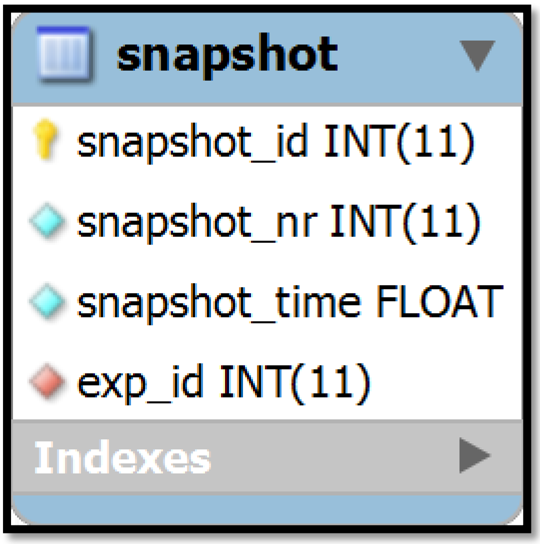
res_mod_notes LONGTEXT

Indexes

Figure 2.9 The residue_info table

The residue_info table simply contains redundant information about residues which is shared by residues across molecular systems (Figure 2.9). This includes information such as residue names and types. New residue templates can also be added for residues which have been modified, for future use. The implementation of this table prevents the unnecessary duplication of data in the database.

2.3.1.8 Snapshot



snapshot

snapshot_id INT(11)

snapshot_nr INT(11)

snapshot_time FLOAT

exp_id INT(11)

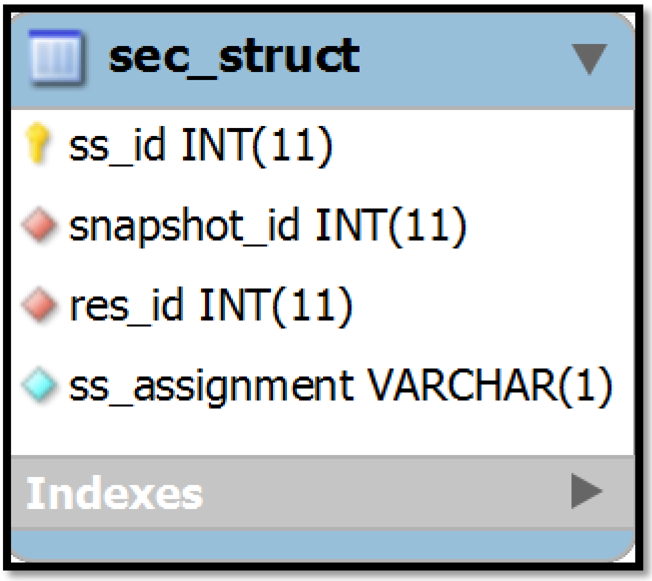
Indexes

Figure 2.10 The snapshot table

The snapshot table represents a single snapshot in a single MD trajectory (Figure 2.10). The attributes describe the snapshot number as well as the time point at which the snapshot was taken.

A snapshot can only belong to one experiment; however a snapshot may contain multiple secondary structure assignments and hydrogen bonds.

2.3.1.9 Sec_struct



sec_struct

ss_id INT(11)

snapshot_id INT(11)

res_id INT(11)

ss_assignment VARCHAR(1)

Indexes

Figure 2.11 The sec_struct table

Table 2.1 Secondary Structure Assignment by YASARA

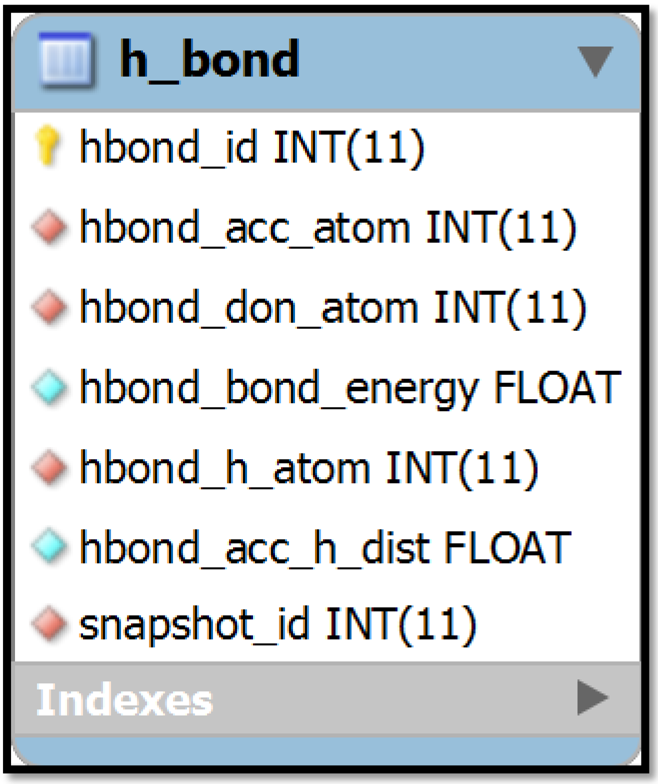
YASARA Assignment	Secondary structure
H	α - Helix
EE/R*	β - strand in a sheet
E	β bridge (minimal β – sheet)
T	Hydrogen bonded turn
G	3/10 Helix
C	All other

* EE is converted to R by the analysis script

The sec_struct table represents a single secondary structure assignment in the MD trajectory (Figure 2.11). The only ordinary attribute is the assignment. The secondary structure assignment is done according to Table 2.1.

First, an assignment belongs to only one residue in only one snapshot. Thus, a snapshot can have multiple residues, and multiple secondary structure assignments, while a residue can belong to multiple snapshots and may also have more than one secondary assignment.

2.3.1.10 H_bond



h_bond

hbond_id INT(11)

hbond_acc_atom INT(11)

hbond_don_atom INT(11)

hbond_bond_energy FLOAT

hbond_h_atom INT(11)

hbond_acc_h_dist FLOAT

snapshot_id INT(11)

Figure 2.12 The h_bond table

The h_bond table represents a hydrogen bond between two atoms in a molecular system (Figure 2.12). A hydrogen bond contains an acceptor atom, a donor atom and the hydrogen atom mediating the bond. The bond may also have a distance as well as a certain energy value associated with it. The table attributes thus mirrors the actual attributes of the hydrogen bond.

A hydrogen bond can belong to only one snapshot, while a snapshot may contain many hydrogen bonds. An atom can participate in many hydrogen bonds, while a particular hydrogen bond can only have one atom associated with one attribute.

2.3.2 Methods (software used)

MySQL Workbench (http://www.mysql.com) was used to design the database and create the SQL script used to create the database. SimDB was implemented and managed in the MySQL server version 5.1.32 (http://www.mysql.com). Scripts written in Python (www.python.org) was used to import data produced by the analysis scripts described previously, and to subsequently retrieve data from the database. All software used is Open Source software.

2.4 DISCUSSION AND CONCLUSION

Python was used together with YASARA to produce a script which analyzed the MD trajectory and saved the results to CSV formatted text files. Subsequently a relational database was designed to store and manage the analysis data generated.

Data in the CSV text files were imported into simDB. Additional python scripts were subsequently used to extract the desired data from the database and to graph the data. This workflow is shown in Figure 2.13.

The object orientated analysis script was designed specifically for use with YASARA. Thus the code as is will not be applicable for more general usage. However, the general object orientated model is adaptable enough to be applied in a more general sense.

The implementation of simDB in this study, though sufficient for the current use, is seen as starting point for a broader application. It is hoped that the database can be expanded by the development of a web-interface, as well as import scripts for other simulation programs such as GROMACS (http://www.gromacs.org). This would enable other researchers to deposit their MD trajectory data into the database, which could aid in the availability of simulation data for broader research use.

The development of software tools was shown for the construction of a pipeline to analyze, store and re-retrieve data from a MD trajectory using YASARA. These tools have expanded the use of YASARA as well as laid the foundation for a more general-use database to store and manage MD trajectory analysis data.

Note that the software for all code developed are available at http://cbio.ufs.ac.za

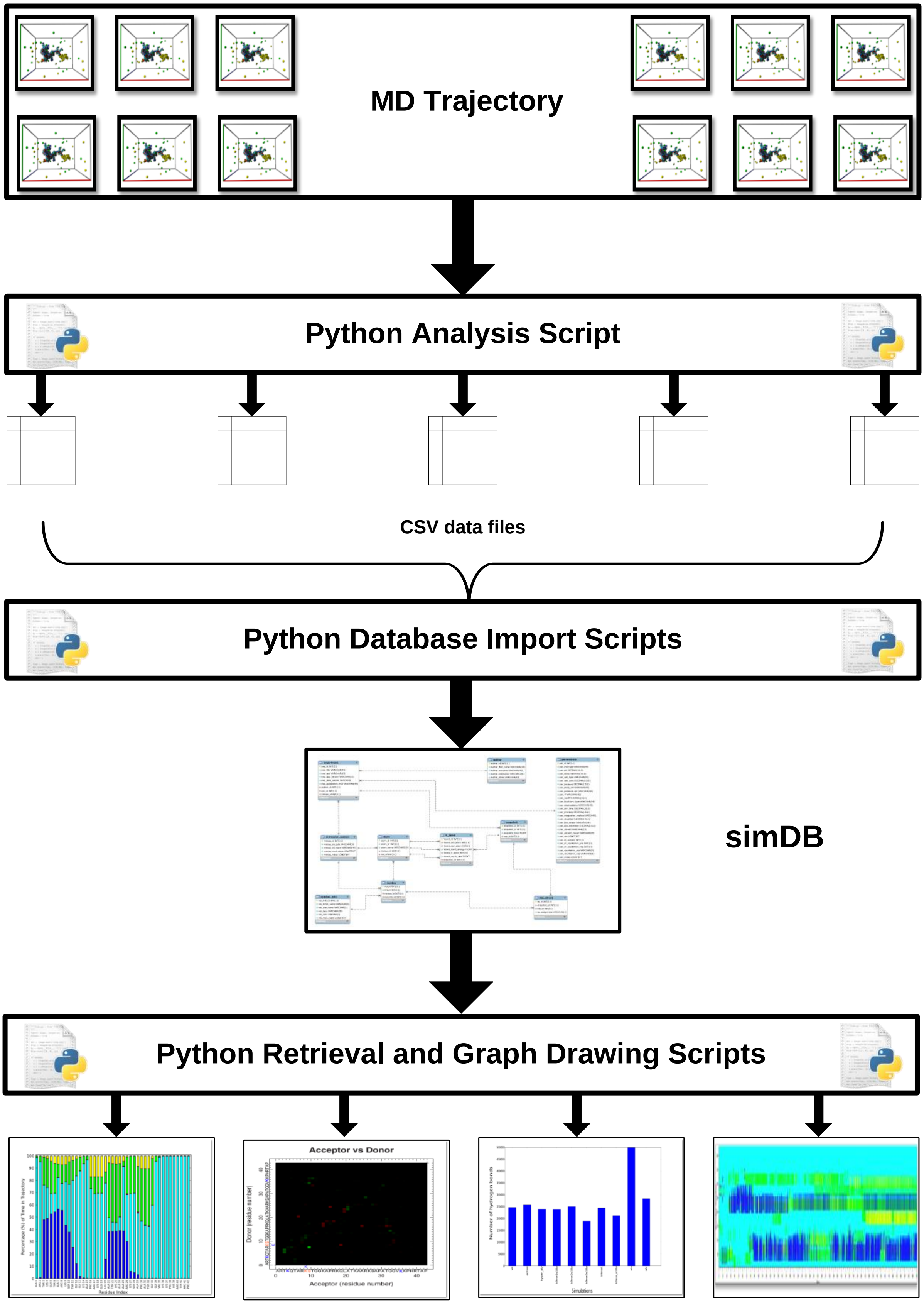


Figure 2.13 Workflow of the analysis and storage of a MD trajectory using the tools developed

2.5 REFERENCES

1. Krieger, E., Koraimann, G. . & Vriend, G. (2002). Increasing the precision of comparative models with YASARA NOVA – a self-parameterizing force field. *Proteins* **47**, 393-402.
2. Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B. & de Hoon, M. J. L. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422-1423.
3. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research* **28**, 235-242.

CHAPTER 3

Validation of the Docking Method

3.1 INTRODUCTION

Molecular docking is a computational technique used to predict the most probable binding conformation between two molecules (the receptor and the ligand) given the geometrical coordinates of the two molecules [1]. This technique has been widely used to great effect in both academic and pharmacological research areas. We was interested in employing a docking approach to study the possibility that the N-terminal tail extension of histone H3 could bind to a position in the nucleosome.

However, results obtained from protein-peptide/protein docking are often spurious. This generally stems from a high degree of conformational flexibility in both receptor and ligand, the relative importance of solvent molecules at the binding interface and conformational changes in the ligand and/or the receptor upon binding [2-6].

AutoDock is a molecular docking software tool which is widely used in the scientific community [7] and version 4.2 is implemented in YASARA to perform docking [8]. AutoDock has been applied in protein-protein docking studies [9, 10]. Thus the YASARA implemented AutoDock was selected to be used in the current study.

Briefly, AutoDock pre-calculates grid maps for use in rapid energy evaluations during the docking experiment [11]. A grid map is a 3 dimensional lattice surrounding the receptor containing a collection of regularly spaced grid points, which stores the potential energy of a probe atom at that position for each of the atoms in the macromolecule (see Figure 3.1). The probe atoms are the constituent atoms of the ligand molecule, thus multiple grid maps are calculated for each docking experiment.

AutoDock uses a Lamarckian Genetic Algorithm (LGA) to search for binding conformations or docked poses [7, 11]. Each trail or run in the docking experiment starts with a population with a size defined by the parameter *ga_pop_size*. The number of trails in each docking experiment sets how many docked poses will be generated by the experiment and is defined by the parameter *ga_run*. The members of the populations are changed (or mutated by) from the starting conformation by the search algorithm and their energies are subsequently evaluated. If the energy is more favorable than the previous generation, the mutation is retained, else the mutation is discarded. This process is repeated until a cumulative number of energy evaluations are reached which is defined by the parameter *ga_num_evals*.

In essence docked poses are allowed to evolve towards the energetically most favorable solution.

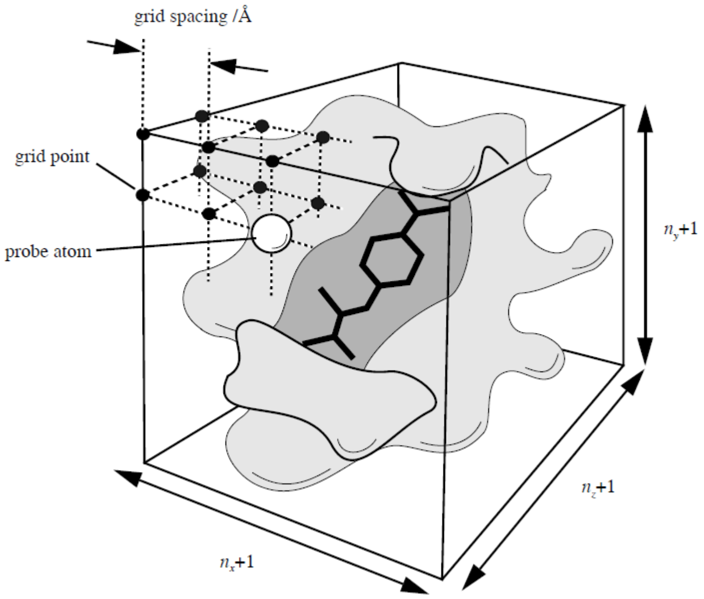


Figure 3.1. A grid map in AutoDock. All the important features of the map are indicated. The ligand bound in the active site of the receptor is shown in the middle of the search space. This figure was adapted from [11]

We were interested in docking some of the conformations of the N-terminal histone H3 tail, derived from MD studies, onto the nucleosome surface, to investigate the possibility of an interaction. To verify the accuracy of the AutoDock procedure, I decided to dock a peptide to the nucleosome surface, where the binding specificity and actual binding position of this peptide had already been elucidated by X-ray crystallography. Such a defined approach would also allow optimization of the docking parameters.

Kaposi's Sarcoma Herpes Virus Latency Associated Nuclear Antigen (KSHV-LANA), an arginine rich 14 – residue peptide, was experimentally shown to bind to the nucleosome surface [12]. The intended experimental system was the 15 – residue N-terminal tail of histone H3 docked to the nucleosome surface. It was therefore decided to dock the KSHV-LANA peptide to the nucleosome surface as a test to establish the closeness of fit of the docked prediction and the actual position of the peptide in the crystal. Thus the crystal structure of KSHV-LANA (PDB – code: 1ZLA) was used as the validation complex.

The aim of the docking method was to evaluate the entire nucleosome surface and the surrounding DNA for possible binding sites of the histone H3 N-terminal tip. The existence, orientation, structure and position of a binding site are unknown. Thus, more emphasis was placed in this chapter on the method being able to distinguish and identify the real binding site and orientation from non-binding sites and orientations.

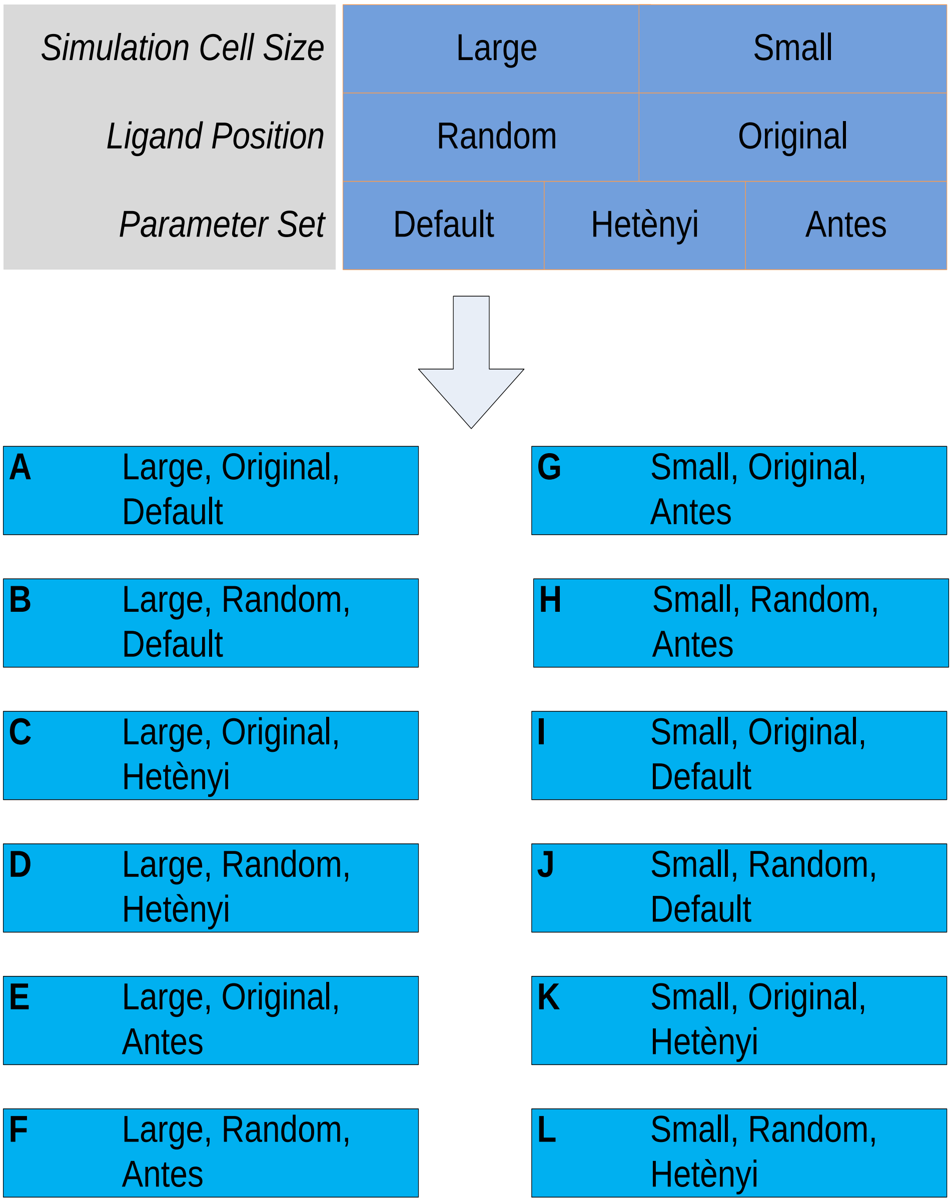
3.2 METHODS

3.2.1 Structure Preparation:

The crystal structure (PDB - code: 1ZLA) was loaded into YASARA. Crystallographic waters were deleted and the structure was cleaned using the *CleanAll* command. Next both ligand (Chain K) and receptor (All Chains minus K) were saved into separate PDB files.

Different parameters (see section 3.2.2) were combined to yield the experiments shown in Figure 3.2 and Figure 3.3. Figure 3.2 represents the rigid docking experiments and Figure 3.3 represents the flexible docking experiments.

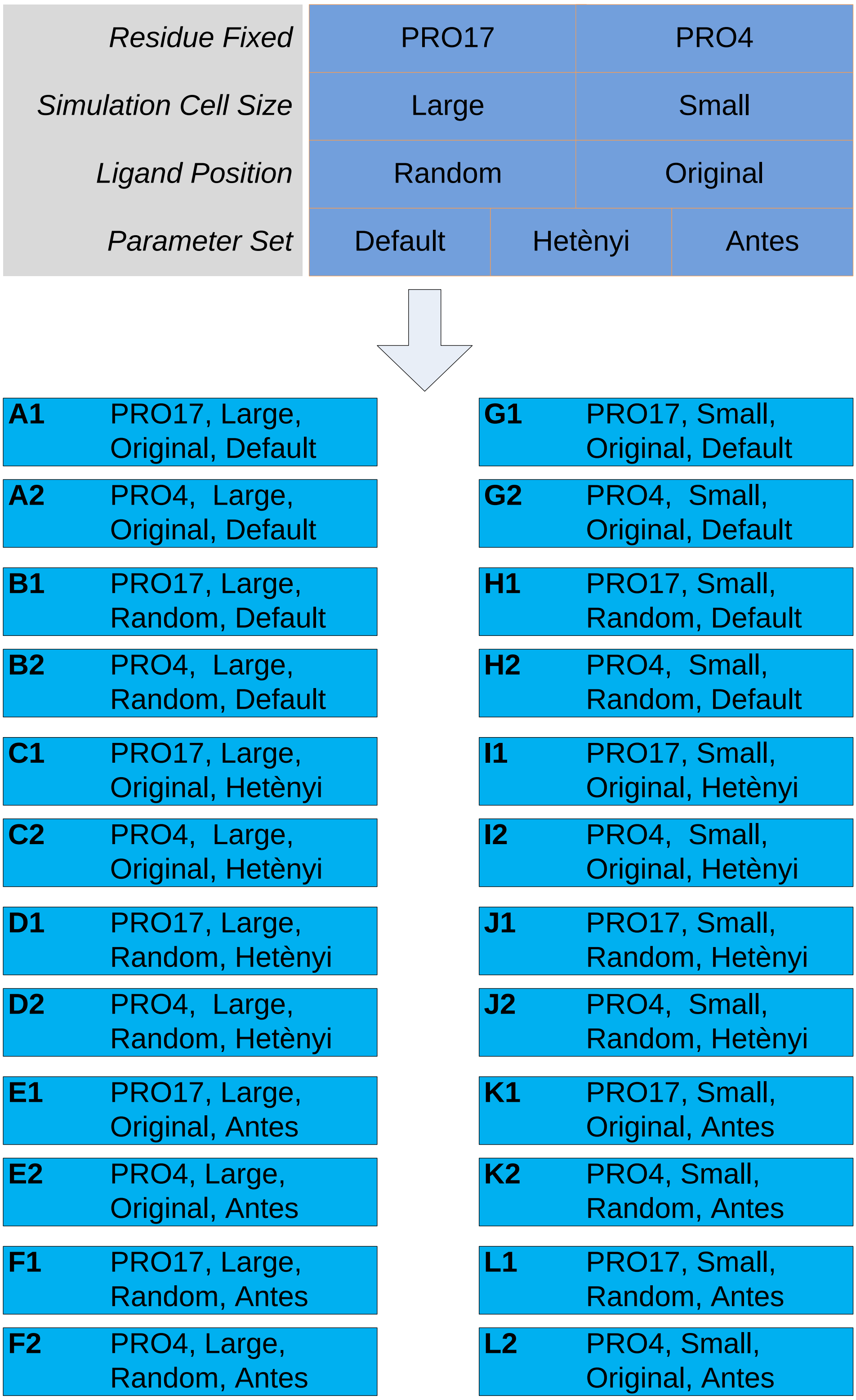
Preparation, docking and analysis were performed using custom Python (http://www.python.org) scripts together with YASARA. All scripts are located on the included CD-ROM.



Simulation Cell Size	Large		Small
Ligand Position	Random		Original
Parameter Set	Default	Hetènyi	Antes

A Large, Original, Default	**G** Small, Original, Antes
B Large, Random, Default	**H** Small, Random, Antes
C Large, Original, Hetènyi	**I** Small, Original, Default
D Large, Random, Hetènyi	**J** Small, Random, Default
E Large, Original, Antes	**K** Small, Original, Hetènyi
F Large, Random, Antes	**L** Small, Random, Hetènyi

Figure 3.2 The generation of the rigid docking experiments performed. Combinations of three different variables were combined to yield 12 experiments. The experiments are labeled A-L.



Residue Fixed	PRO17	PRO4	
Simulation Cell Size	Large	Small	
Ligand Position	Random	Original	
Parameter Set	Default	Hetènyi	Antes

A1	PRO17, Large, Original, Default	**G1**	PRO17, Small, Original, Default
A2	PRO4, Large, Original, Default	**G2**	PRO4, Small, Original, Default
B1	PRO17, Large, Random, Default	**H1**	PRO17, Small, Random, Default
B2	PRO4, Large, Random, Default	**H2**	PRO4, Small, Random, Default
C1	PRO17, Large, Original, Hetènyi	**I1**	PRO17, Small, Original, Hetènyi
C2	PRO4, Large, Original, Hetènyi	**I2**	PRO4, Small, Original, Hetènyi
D1	PRO17, Large, Random, Hetènyi	**J1**	PRO17, Small, Random, Hetènyi
D2	PRO4, Large, Random, Hetènyi	**J2**	PRO4, Small, Random, Hetènyi
E1	PRO17, Large, Original, Antes	**K1**	PRO17, Small, Original, Antes
E2	PRO4, Large, Original, Antes	**K2**	PRO4, Small, Random, Antes
F1	PRO17, Large, Random, Antes	**L1**	PRO17, Small, Random, Antes
F2	PRO4, Large, Random, Antes	**L2**	PRO4, Small, Original, Antes

Figure 3.3 The generation of the flexible docking experiments performed. Four different variables were combined to yield 24 experiments. The experiments are labeled A1-L1 and A2-L2.

3.2.2 Parameters tested

Various factors were involved in validating and tweaking the docking method. These were:

3.2.2.1 Ligand Flexibility

A rigid ligand was used in the docking experiments to see whether AutoDock could correctly dock the ligand to the receptor given the correct docking structure.

AutoDock allows for ligand flexibility [7]. However, there is a limit to the degree of flexibility that can be incorporated into the ligand. This limit is usually around 10 residues depending on the side – chains present in the peptide. Because of this limitation, one residue has to be fixed in space. Thus the C-terminus and the N-terminus were fixed and docked separately. The aim with the flexible docking was to see whether AutoDock could reproduce both the binding position and the structure of the ligand.

3.2.2.2 Docking Search Area Size

YASARA uses a simulation cell to denote the docking search area in AutoDock. Thus the terms “simulation cell”, “cell” and “docking search area” are used interchangeably below.

YASARA gives a warning when the simulation cell is larger than 47Å x 47Å x 47Å. However the core particle has dimensions of approximately 100Å x 40Å x 100Å. Increasing the size of the search area would mean reducing the accuracy of the docking procedure, because of a lower grid map density [7].

One solution tested was a grid-based approach. This involved dividing the simulation cell into smaller, overlapping cells and running multiple simulations instead of running one simulation in one cell. The two cell sizes tested were denoted "Large" (100Å x 75Å x 100Å) and "Small" (90Å x 40Å x 90Å).

3.2.2.3 Ligand Orientation and Position

It was necessary to also test what effect the position and orientation of the ligand relative to the binding site would have on the ability of AutoDock to find the correct binding orientation. Thus we ran all experiments (excluding the grid docking experiments) using the ligand that was moved 20Å along the Y – axis relative to the binding position in the crystal structure, and a random ligand position, implemented by placing the ligand into the cell using the command: *FillCellObj ligand, Copies=1, RandomOri=Yes* in YASARA.

3.2.3 Docking Algorithm Parameters

AutoDock uses a Lamarckian Genetic Algorithm which is highly customizable [7]. It was shown that changing different parameters from their default values could increase accuracy[10]. Three parameter sets were tested to determine which set produced the best results. The three parameter sets consisted of the default and two other sets found in literature. The first parameter set from literature was the parameter set used by Hetènyi and van der Spoel in the blind docking of flexible peptides to protein targets[10]. The second parameter set was used by Antes in evaluating the relative performance of newly developed docking software, DynaDock, against AutoDock[13].

The parameter values are shown in Table 3.1.

Table 3.1 Different parameter sets used in experiments with AutoDock

	Parameter			
Name	***ga_run***	***ga_pop_size***	***ga_num_evals***	**Reference**
default	50	150	25 000 000	-
Hetenyi	100	250	10 000 000	[10]
Antes	400	400	10 000 000	[13]

3.2.4 Grid – based Docking

To solve the problem of losing accuracy with larger simulation cells, the large simulation cell covering the entire surface of the nucleosome was divided into smaller cells. This was achieved by using YASARA's function to render rectangular boxes. This grid is not to be confused with the grid maps which AutoDock generates as part of its docking algorithm. Each grid consisted of 9 grid cells with the dimensions 50Å x 75Å x 50Å and 45Å x 45Å x 45Å for the large and the small grid, respectively. The grid was placed over the simulation box in both cases as seen in Figure 3.4.

Each row and column contained 3 of these cells, placed in a manner to form two distinct cells (Figure 3.4 I) and overlapping cells in both the rows and columns (Figure 3.4 II). The rows were indicated by the letters A – C and the columns were indicated by the numbers 1-3. Thus, each experiment was split into 9 separate experiments and performed independently.

3.2.5 Docking Analysis

The original crystal structure (1ZLA.pdb) was used as reference structure and super imposed on the nucleosome containing the docked structures. The Root Mean Square Deviation (RMSD) value was then calculated between the ligand α – carbons of each pose and the ligand α – carbons in the KSHV-LANA reference structure.

In addition to the RMSD value, a binding energy value was also calculated for each pose and this served as the score whereby poses were ranked. It is important to note that in contrast to the stand-alone version of AutoDock and many other programs, YASARA reports its free energy values as positive instead of negative values. For example: a value of -5.00 kcal/mol in AutoDock will correspond to a value of 5.00 kcal/mol in YASARA.

The poses were clustered by AutoDock using a 5 Å RMSD cut-off. The binding energy and RMSD with KSHV LANA were also calculated for each representative structure of the clusters obtained.

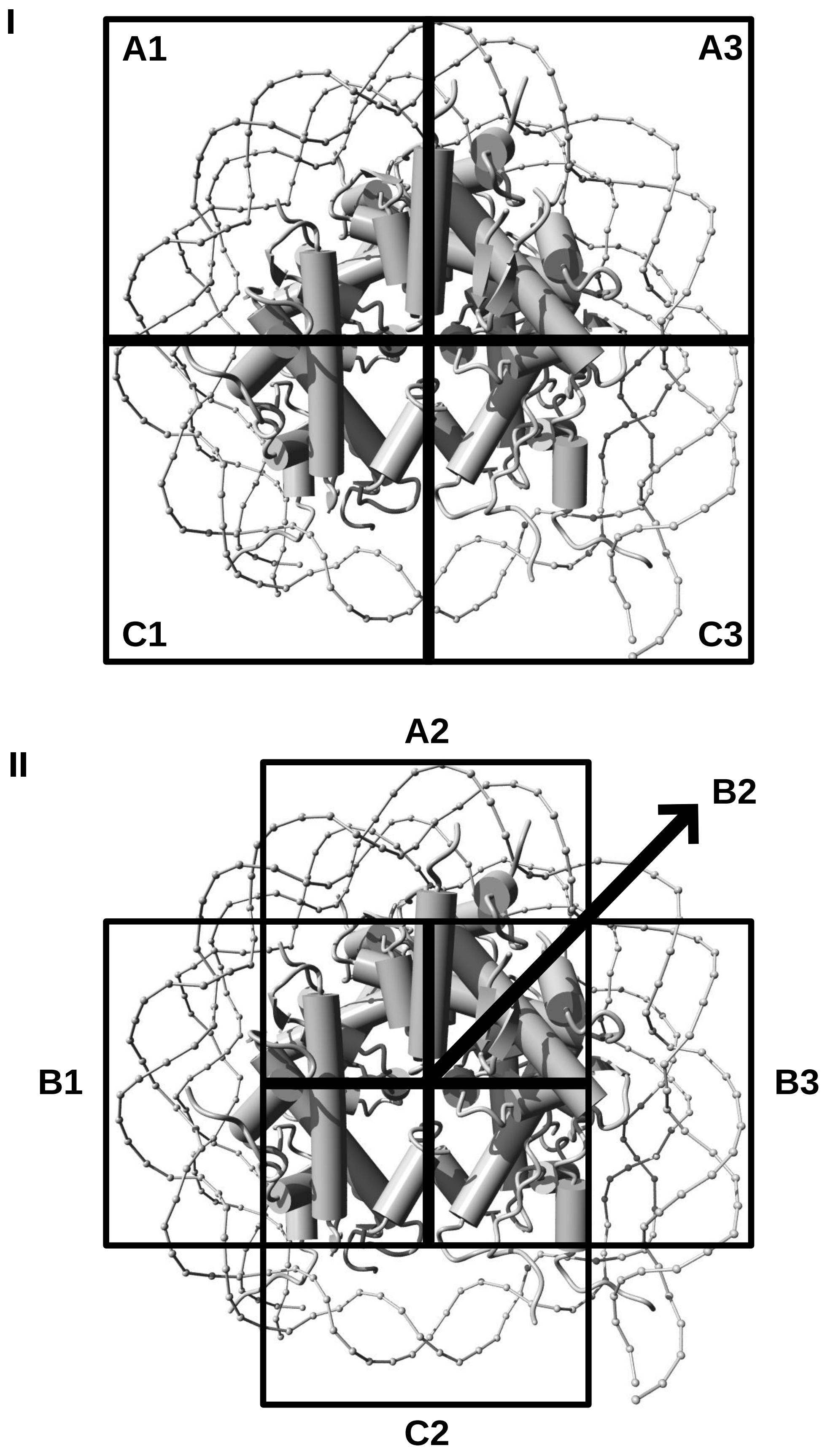


Figure 3.4 The grid covering the nucleosome surface and DNA. I shows the non – overlapping grid cells and II shows the overlapping grid cells.

Molecular graphics created with YASARA (www.yasara.org) and POVRay (www.povray.org)

3.3 RESULTS

3.3.1 Non-grid – based docking experiments

3.3.1.1 Rigid Docking

We first performed a rigid docking experiment using the entire nucleosome surface as a potential receptor. Our aim was to establish whether Autodock could at least correctly dock KSHV LANA to its crystal structure position and orientation given the rigid crystal structure. The results are summarized in Table 3.2 and Figure 3.5. From Figure 3.5 it can be seen that docked positions very similar to that of the ligand in the crystal structure was found in all experiments except in experiment A and B (see Figure 3.2). This was confirmed by the RMSD values associated with the binding positions in Table 3.2, where all RMSD values with the exception of experiment A and B were below 1 Å. An RMSD value of 2.0 Å is generally regarded as a successful redocking pose [14]. In experiment A the ligand was docked on the side of the DNA and in experiment B it was docked in between the two DNA gyres. The binding energy values associated with these poses were approximately 2.5 kcal/mol lower than the lowest binding energy value obtained by a correct binding pose. From these results it is clear that rigid docking with AutoDock could reproduce the crystal binding pose of KSHV-LANA on the nucleosome surface, rank it as the best pose and distinguish it from other binding poses. Referring to Table 3.2, it is seen that using a small search area yielded higher binding energy values and lower RMSD values, and also allowed the default parameter set to find the crystal binding orientation. The starting orientation of the ligand had a marginal effect on the values obtained. With both small and large simulation cells, however, the Antes parameter set seemed to neutralize this effect the best. The best values were obtained with experiment K, which used a small search space, the original starting position of the ligand and the Antes parameter set. Though there were differences in the values obtained, these differences were marginal, with the binding energy values never going beyond 0.5 kcal/mol and all RMSD values remained below 1 Å. It was thus shown that Autodock could correctly predict the crystal binding position and orientation of KSHV LANA to the nucleosome given the rigid crystal structure.

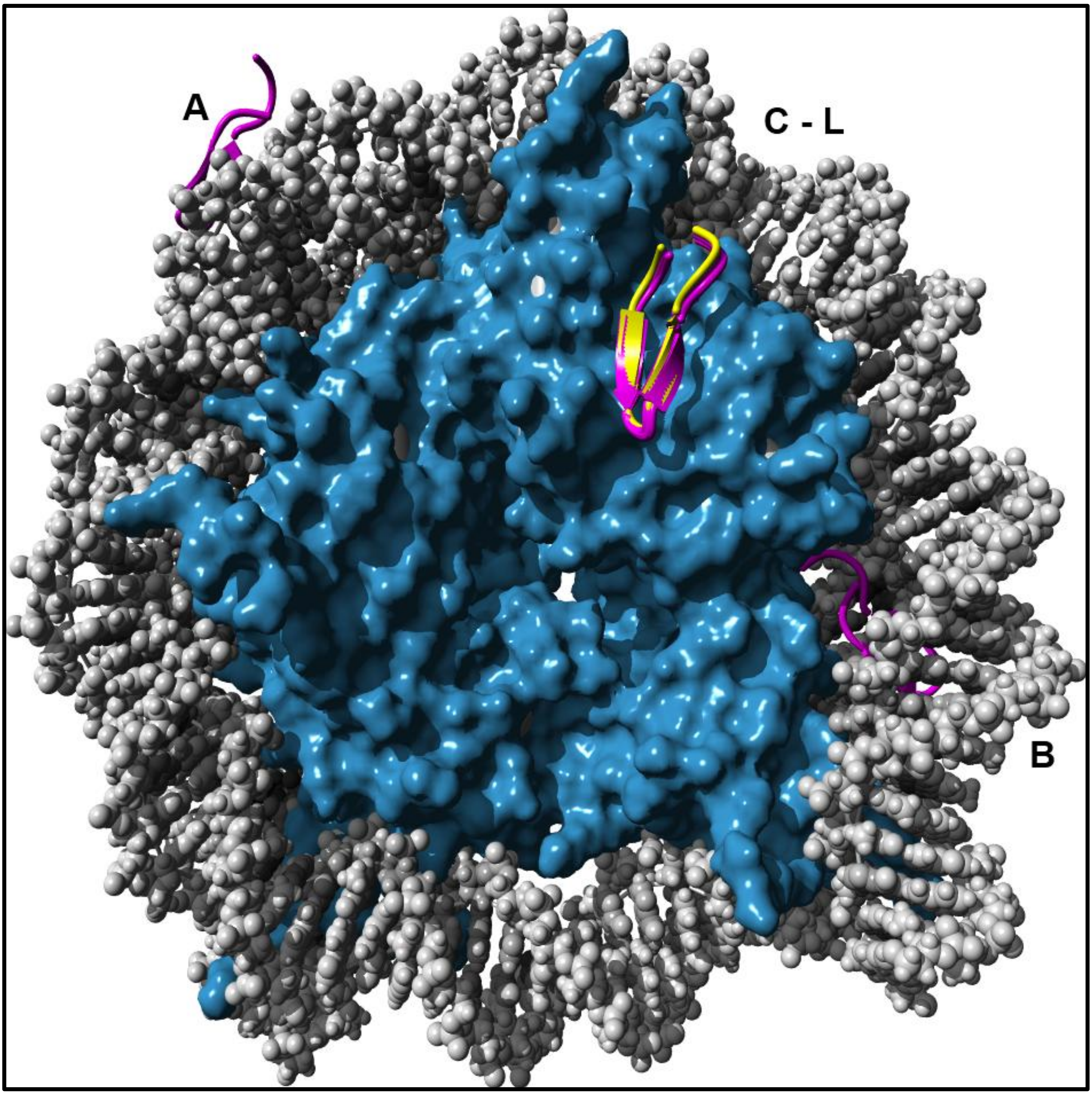


Figure 3.5 Top ranked docking poses obtained by rigid docking experiments. All experiments except A and B was able to reproduce the crystal binding position and orientation The light blue surface represents the histone octamer and the gray ball structures the nucleosomal DNA. Magenta stick structures represent the docked poses which are labeled according to Figure 3.2. The yellow stick structure represents the original crystal docking pose as found in 1ZLA. Molecular graphics created with YASARA (www.yasara.org) and POVRay (www.povray.org)

Table 3.2 Binding energy and RMSD values for the top ranked docking poses for the rigid non-grid experiments.

	Experiment			***Binding Energy (kcal/mol)***	***RMSD (Å)***
	Simulation Cell Size	***Ligand Position***	***Parameter Set***		
A	Large	Original	Default	4.41	52.09
B	Large	Random	Default	4.82	38.18
C	Large	Original	Hetenyi	7.05	0.83
D	Large	Random	Hetenyi	7.15	0.85
E	Large	Original	Antes	7.16	0.90
F	Large	Random	Antes	7.12	0.87
G	Small	Original	Default	7.21	0.73
H	Small	Random	Default	7.47	0.45
I	Small	Original	Hetenyi	7.46	0.42
J	Small	Random	Hetenyi	7.09	0.70
K	Small	Original	Antes	7.49	0.44
L	Small	Random	Antes	7.45	0.43

3.3.1.2 Flexible Docking

We did not know what the binding structure (if any existed) of the H3 N-terminal tail would be and subsequently ran an experiment to establish whether AutoDock could find, or in this case retain, the true binding structure(s). To this end we removed the restraints on the ligand, allowing movement of the ligand to correctly identify the binding position of the KSHV-LANA test structure. The results are summarized in Figure 3.6 and Table 3.3. As is seen in Figure 3.3, flexible docking with AutoDock was unable to reproduce the crystal binding position and structure of KSHV-LANA). The characteristic β – hairpin structure in the peptide was completely abolished and the structure was peeled open. Subsequently this open structure was predominantly bound to or near the nucleosomal DNA. Fixing either the N-terminus or the C-terminus had a significant effect on the best docked poses found, with the most profound difference witnessed for experiment G: G1 was docked almost directly on the opposite side of the nucleosome compared to G2, and G2 had a binding energy value (see Table 3) of more than 4 kcal/mol higher than that of G1.

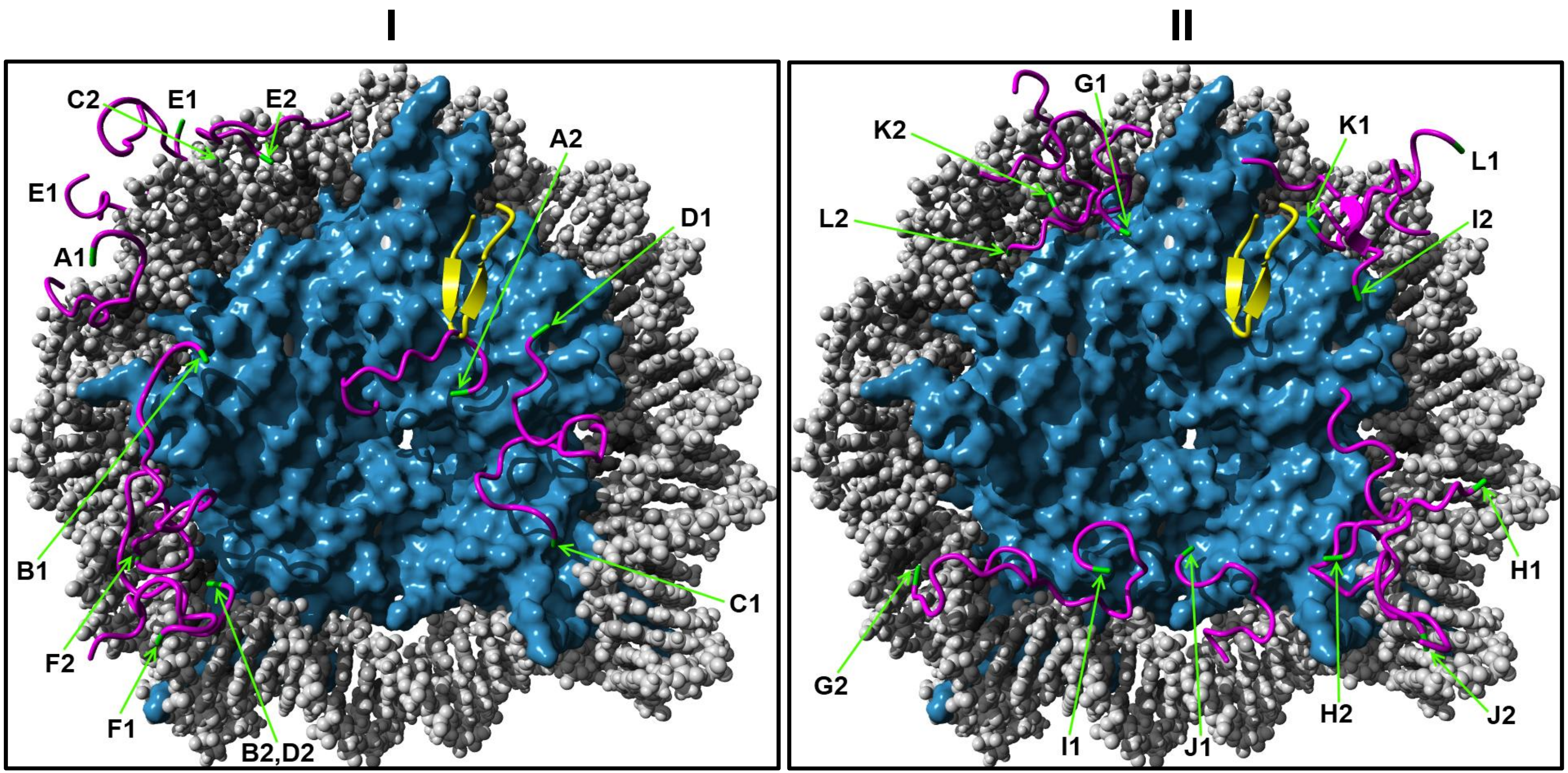


Figure 3.6 Top ranked docking poses obtained by the flexible docking experiments. I represent the experiments in the large simulation cell and II represents the experiments in the small simulation cell. In both I and II the original crystal binding position, orientation and ligand structure was not reproduced. The light blue surface represents the histone octamer and the gray ball structures the nucleosomal DNA. Magenta stick structures represent the docked poses which are labeled according to Figure 3.3. The green caps on the docking poses represent the residue which was fixed during each experiment. The yellow stick structure represents the original crystal docking pose as found in 1ZLA.

Molecular graphics created with YASARA (www.yasara.org) and POVRay (www.povray.org)

Table 3.3 Binding energy and RMSD values for the top ranked docking poses for the flexible non-grid experiments.

	Experiment				***Binding Energy (kcal/mol)***	***RMSD (Å)***
	Simulation Cell Size	***Ligand Position***	***Parameter Set***	***Fixed Residue***		
A1	Large	Original	Default	PRO17	3.65	48.73
A2	Large	Original	Default	PRO4	5.81	20.48
B1	Large	Random	Default	PRO17	4.31	47.47
B2	Large	Random	Default	PRO4	5.09	55.77
C1	Large	Original	Hetenyi	PRO17	5.75	30.22
C2	Large	Original	Hetenyi	PRO4	5.25	48.46
D1	Large	Random	Hetenyi	PRO17	3.43	23.60
D2	Large	Random	Hetenyi	PRO4	4.09	36.88
E1	Large	Original	Antes	PRO17	5.67	54.99
E2	Large	Original	Antes	PRO4	5.44	36.48
F1	Large	Random	Antes	PRO17	5.32	61.55
F2	Large	Random	Antes	PRO4	5.80	51.44
G1	Small	Original	Default	PRO17	6.64	31.91
G2	Small	Original	Default	PRO4	10.93	55.43
H1	Small	Random	Default	PRO17	7.34	42.94
H2	Small	Random	Default	PRO4	6.48	44.56
I1	Small	Original	Hetenyi	PRO17	6.39	46.49
I2	Small	Original	Hetenyi	PRO4	4.65	16.54
J1	Small	Random	Hetenyi	PRO17	5.49	47.00
J2	Small	Random	Hetenyi	PRO4	5.78	32.63
K1	Small	Original	Antes	PRO17	8.19	25.35
K2	Small	Original	Antes	PRO4	7.31	34.32
L1	Small	Random	Antes	PRO17	4.74	26.53
L2	Small	Random	Antes	PRO4	5.75	28.27

3.3.2 Grid – based Docking

AutoDock could not reproduce the crystal structure with a flexible ligand and we believed that this stemmed from the reduced accuracy obtained by the use of a simulation box larger than 47Åx47Åx47Å. Since AutoDock implements an interaction grid map with a constant number of divisions fitted to the docking box where this box exceeded 47Åx47Åx47Å, docking to a large surface such as a nucleosome in a single large box may miss productive docking positions that fall between grid vertices. We thus devised a strategy to reduce the simulation cell size below 47Åx47Åx47Å and at the time increase the amount of sampling done across the nucleosome surface, which could be useful when docking the unknown H3 tail.

Subsequently I divided the large simulation box into smaller, partially overlapping docking boxes (see Figure 3.4), and performed docking with the ligand in each of the boxes independently. This approach was followed for both a rigid and a flexible ligand to ascertain whether an increase in accuracy and binding energy could be obtained.

3.3.2.1 Rigid Grid – based Docking

Nine partially overlapping docking boxes were defined (see Figure 3.4), and the KSHV-LANA peptide independently docked in each of these boxes. The results are shown in Figure 3.7 and summarized in Table 3.4. With grid - docking, the cells that covered the original KSHV-LANA crystal binding position were A2, A3, B2 and B3. In all experiments the crystal binding orientation was faithfully reproduced in cells A2, B2 and B3 (see Figure 3.6). Experiment F (see Table 3.2) was the only parameter set in which the crystal binding position and orientation was reproduced. As with the single rigid docking experiments, experiments with a smaller docking volume produced higher binding energy values and lower RMSD values (see Table 3.4). In the smaller docking volume, the parameter sets had no significant effect on the binding energy values and the RMSD values. Using grid - based docking yielded higher binding energy values compared to the non-grid docking experiments (see Table 3.2 and Table 3.4). Experiments in grid cells not covering the original crystal binding position yielded top ranking poses bound to or very close to the nucleosomal DNA. Since the KSHV-LANA peptide is arginine rich, binding poses with the negatively charged DNA is to be expected. It is however important to note that the crystal binding position and orientation yields significantly higher (~ 3 kcal/mol) binding energy values compared to the non-specific binding poses with the DNA. Thus AutoDock was able to distinguish between a true binding orientation and a non-specific one using a rigid grid-based docking approach.

3.3.2.2 Flexible Grid – based Docking

However, with flexible grid – docking the characteristic β – hairpin structure of KSHV – LANA was again abolished and the crystal binding position and orientation was not reproduced by any of the parameter sets (see Figure 3.8). With the exception of cells B2 and C2, all top ranked docked poses tended to be bound with or near to the nucleosomal DNA. The docked pose producing the highest binding energy value was obtained in cell B1 using the Antes parameter set in the smaller search space (Experiment F; see Table 3.5). In this pose the ligand is bound to the nucleosomal DNA with some contact with the edge of the nucleosomal surface (see Figure 3.7). Across all the experiments the majority of top ranked docking poses were found to be associated with the DNA. Thus flexible docking could not reproduce the binding position, structure and orientation of our test peptide and could subsequently not suitable for use in investigate the binding of the H3 tail to the nucleosome surface.

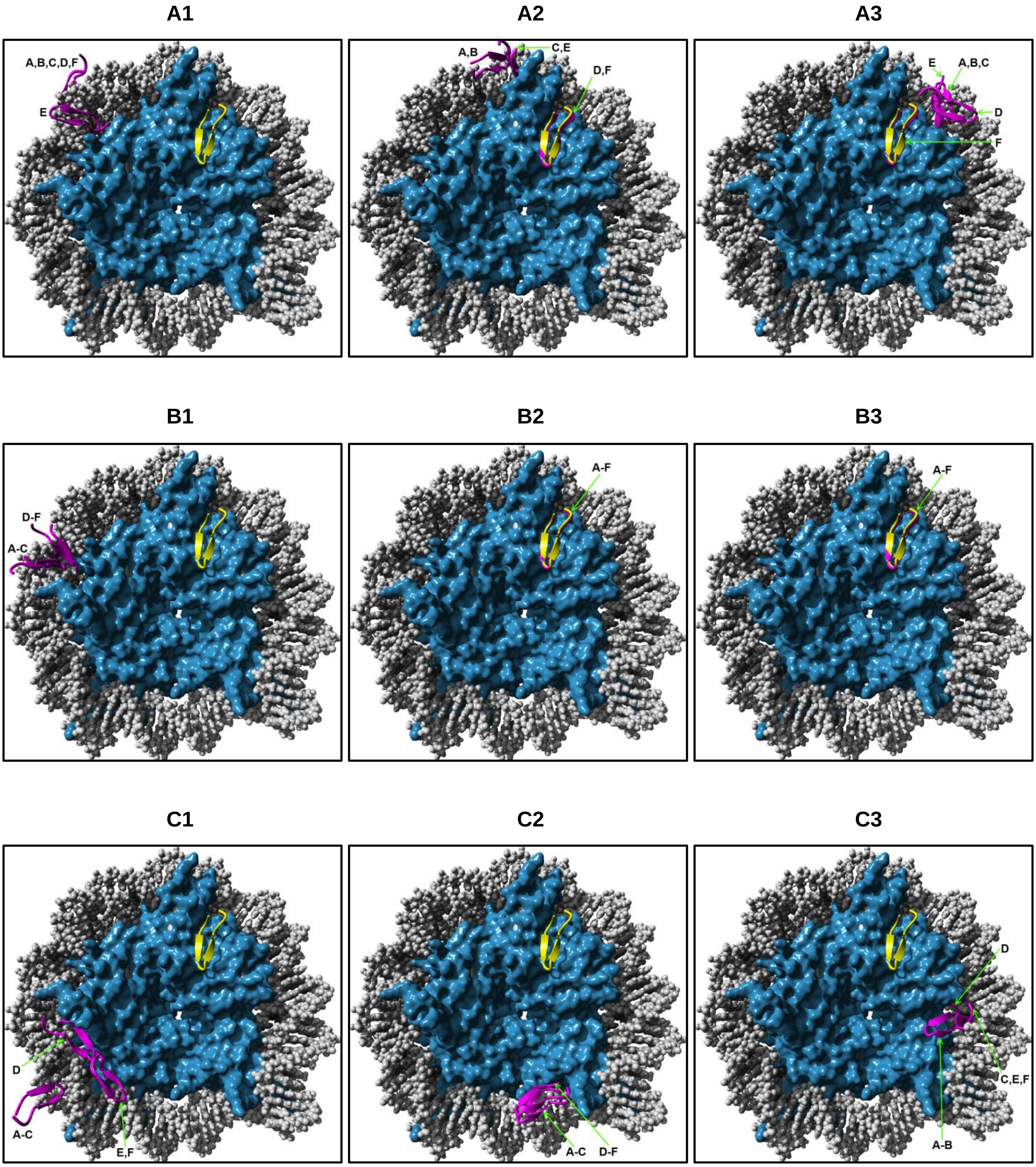


Figure 3.7 Top ranked docking poses obtained with rigid grid-based docking experiments. Cells A2, A3, B2 and B3 cover the crystal binding position. The light blue surface represents the histone octamer and the gray ball structures the nucleosomal DNA. Magenta stick structures represent the docked poses which are labelled according to Figure 3. The yellow stick structure represents the original crystal docking pose as found in 1ZLA.

Molecular graphics created with YASARA (www.yasara.org) and POVRay (www.povray.org)

Table 3.4 Binding Energy (kcal/mol) and C RMSD (Å) values obtained using rigid grid docking and variable box size and parameter sets. Highlighted values indicate the cell which covers the crystal binding site of KSHV-LANA on the nucleosome.

Experi - ment	A		B		C		D		E		F	
	Large Default		Large Hetenyi		Large Antes		Small Default		Small Hetenyi		Small Antes	
Cell	Binding Energy (kcal/mol)	α - C RMSD (Å)	Binding Energy (kcal/mol)	α - C RMSD (Å)	Binding Energy (kcal/mol)	α - C RMSD (Å)	Binding Energy (kcal/mol)	α - C RMSD (Å)	Binding Energy (kcal/mol)	α - C RMSD (Å)	Binding Energy (kcal/mol)	α - C RMSD (Å)
A1	5.20	52.18	5.14	52.23	5.18	52.21	5.68	51.86	5.35	41.59	5.62	51.84
A2	4.49	35.00	4.46	35.02	4.96	30.79	8.48	0.54	5.90	30.90	8.37	0.85
A3	5.00	26.55	4.98	26.56	4.95	26.53	5.89	21.96	5.96	23.49	9.62	0.71
B1	4.24	50.96	4.24	50.99	4.20	50.92	4.71	44.97	4.70	44.95	4.69	44.97
B2	9.65	0.69	9.67	0.70	9.61	0.70	10.53	0.66	10.48	0.67	10.53	0.67
B3	8.45	0.80	8.42	0.82	8.41	0.68	10.27	0.65	10.24	0.64	10.24	0.67
C1	4.15	68.83	4.11	68.84	4.15	68.81	5.20	50.46	5.15	51.09	5.21	50.85
C2	5.81	49.38	5.76	49.39	5.76	49.38	4.70	47.80	5.76	49.39	4.68	47.81
C3	6.13	31.77	6.11	31.78	6.51	35.83	7.14	36.87	7.37	35.86	7.26	35.90

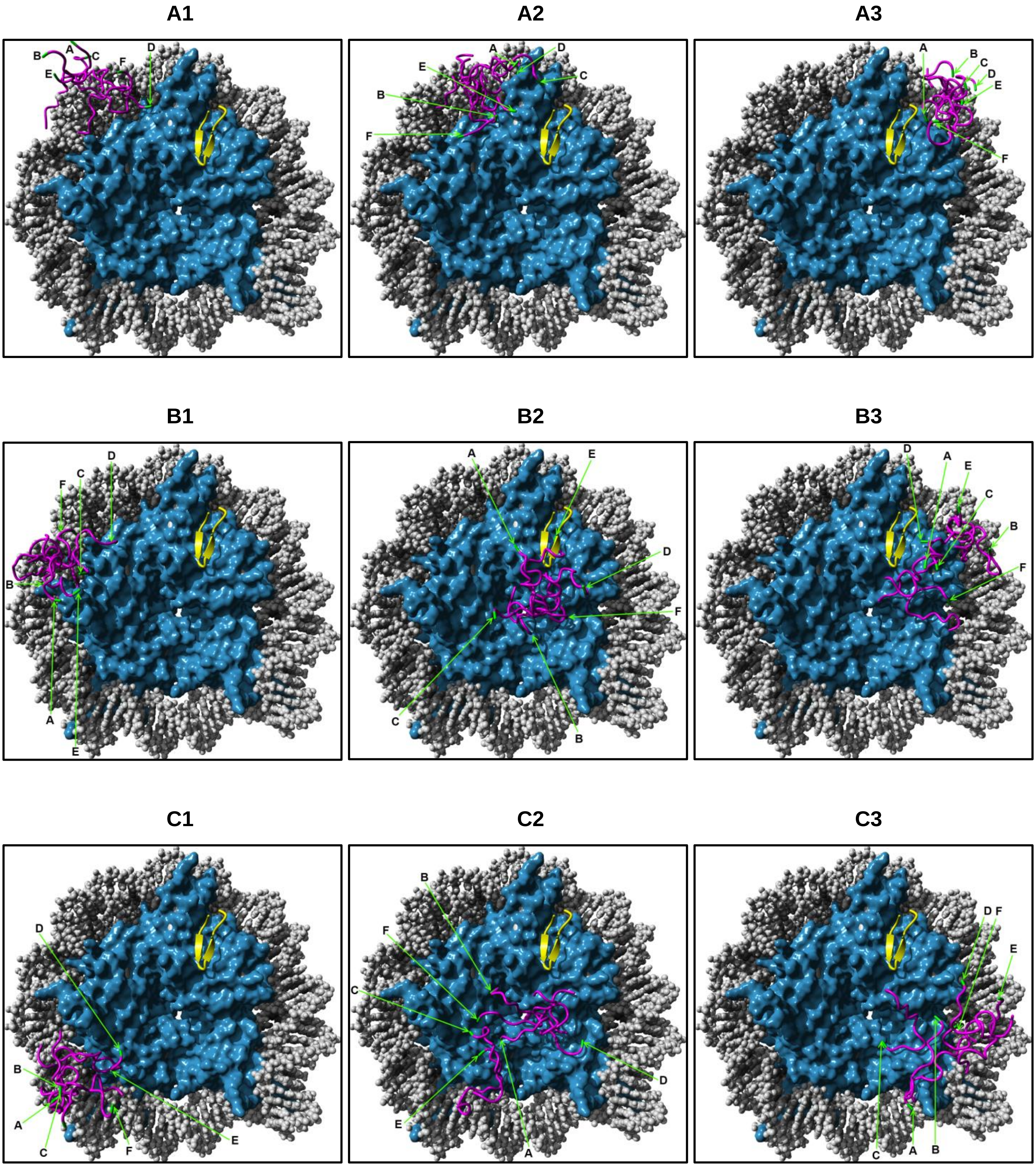


Figure 3.8 Top ranked docking poses obtained with flexible grid-based docking experiments. Cells A2, A3, B2 and B3 cover the crystal binding position. The light blue surface represents the histone octamer and the gray ball structures the nucleosomal DNA. Magenta stick structures represent the docked poses which are labeled according to Figure 2. The green caps on the docking poses represent the residue which was fixed during each experiment. The yellow stick structure represents the original crystal docking pose as found in 1ZLA.

Molecular graphics created with YASARA (www.yasara.org) and POVRay (www.povray.org)

Table 3.5 Binding Energy (kcal/mol) and C RMSD (Å) values obtained using flexible grid docking and variable simulation cell size and parameter sets. Highlighted values indicate the cell which covers the crystal binding site of KSHV-LANA on the nucleosome.

Experiment	**A**		**B**		**C**		**D**		**E**		**F**	
	Large Default		**Large Hetenyi**		**Large Antes**		**Small Default**		**Small Hetenyi**		**Small Antes**	
Cell	**Binding Energy (kcal/mol)**	**RMSD (Å)**	**Binding Energy (kcal/mol)**	**RMSD (Å)**	**Binding Energy (kcal/mol)**	**RMSD (Å)**	**Binding Energy (kcal/mol)**	**RMSD (Å)**	**Binding Energy (kcal/mol)**	**RMSD (Å)**	**Binding Energy (kcal/mol)**	**RMSD (Å)**
A1	11.59	52.66	7.60	48.74	9.35	46.28	12.47	35.00	10.73	39.51	10.28	43.36
A2	6.97	32.30	6.37	32.29	7.57	27.35	10.09	29.31	9.42	28.21	11.45	29.80
A3	9.57	23.30	7.03	27.72	7.03	27.82	9.41	31.97	8.62	24.55	9.42	19.94
B1	4.90	48.70	7.89	52.60	7.19	44.31	9.24	40.26	8.60	47.02	16.20	44.88
B2	7.80	18.76	5.94	24.20	5.62	24.42	9.15	20.60	8.80	14.83	9.08	23.51
B3	4.24	17.44	7.10	26.81	9.32	20.08	9.38	13.76	11.14	20.47	8.45	25.29
C1	6.36	58.72	7.98	62.06	8.26	60.37	11.14	49.56	9.01	52.79	8.41	56.99
C2	4.77	28.24	3.62	25.51	6.33	42.92	9.54	27.35	7.41	49.12	5.66	25.90
C3	8.69	41.53	7.28	39.82	7.27	27.60	10.66	32.90	9.19	38.33	9.34	36.65

3.4 DISCUSSION

Protein – peptide docking is still a developing field and from the results obtained in this chapter it is clear that docking programs must be systematically verified before use. The crystal structure KSHV – LANA bound to the acidic patch on the nucleosome [12] was used as reference structure to test whether AutoDock could correctly predict the crystal binding pose of a small basic rich peptide with the nucleosome surface. By varying ligand flexibility, the overall size of the search space, starting orientation of the ligand and the parameter set used, a better insight into the importance of each parameter to the veracity of the docking result was obtained.

Initially docking with the entire surface defined as search space in one experiment was attempted. Keeping the ligand rigid, and docking with all combinations of the remaining variables, allowed the faithful reproduction of the crystal pose in all experiments, with the exception of experiments using the default parameter set in the larger search space. This was probably due to a combined effect arising from a less dense grid map combined with a low number of energy evaluations and trails. Ligand orientation and choice of parameter set had a minor effect on the binding energy values and RMSD values obtained. However using a smaller search space yielded higher binding energy values and RMSD values and did allow the experiments using the default parameter set to locate the crystal binding position and orientation. This was probably due to the smaller search space having a denser grid map.

In addition to finding the correct binding position and orientation, as is the case with rigid docking, flexible docking adds the additional criteria of finding the correct bound structure of the ligand [1]. However, when repeating the previous batch of experiments with a flexible ligand, the characteristic β – hairpin structure were abolished and the crystal binding position and orientation could not be found in any of the experiments.

One of the aims of the overall study was to scan the nucleosome surface for potential binding sites of the N-terminal tip of the H3 tail. Thus, we had no prior knowledge of the binding site if any such site did, indeed, exist. This is termed blind docking [10]. Thus, too thoroughly and efficiently sample the surface was of utmost importance. The single search space experiments were insufficient, as the number of grid points in AutoDock remains constant after the search space is extended beyond 47Åx47Åx47Å (thus grid map density decreases), while a search space of approximately 100Åx100Åx40Å covered the entire nucleosome surface. To solve this problem, the search space used in the previous experiments was split into a grid consisting of overlapping, smaller search spaces, or grid cells. This essentially split a single experiment up into 9 separate experiments, increasing computational time, but increasing sampling efficiency over the entire surface.

All rigid grid-docking experiments were able to find the crystal binding position and orientation in at least 3 of the 4 cells covering the crystal binding pose. Only the experiment using the Antes parameter set with a smaller search space was able to reproduce the crystal binding position and orientation in all 4 cells. With the grid – based docking approach the binding energy values were improved over the single search space experiments. In addition, top ranked binding orientations and positions in the cells not covering the crystal binding pose tended to interact with the nucleosomal DNA, which gave lower binding energy values compared to the reproduced crystal pose. This is of great importance because the basic character of KSHV-LANA mimics the basic character of the H3 N-terminal tip, which could also be expected to interact with negatively charged nucleosomal DNA. Thus AutoDock was able to distinguish between a known surface binding site on the nucleosome surface and non-specific interaction sites with the DNA when using a strongly basic peptide.

Experiments using the flexible ligand in a grid-docking approach once again yielded no reproduction of the crystal ligand structure, binding position or binding orientation. The top ranked ligand structures were folded open and were associated with nucleosomal DNA.

Thus it was clear that grid – based rigid docking should be used instead of flexible docking in the current study. Hetènyi and van der Spoel also found that it was possible to obtain positive blind docking results with rigid docking for peptides up to 30 residues in length [10]. A tetra peptide was the longest peptide that could dock using flexible docking [10]. Thus from the results obtained, docking a residue with the maximum number of flexible torsion angles allowed in AutoDock, yielded results of low accuracy. Only one test system was, however, used. Thus a more comprehensive study, with more systems, should be done to establish whether AutoDock is inefficient with the blind docking of peptides near the maximum amount of flexibility allowed by the program.

The smaller docking search space proved superior in both single search space and grid – based docking experiments and it was therefore decided to utilize a search space (grid) size of 90Å x 40Å x 90Å in the remainder of the study.

The Antes parameter set was computationally the most expensive; however it was able to find the crystal binding position and orientation in all the grid cells covering the crystal binding orientation and position. AutoDock is a very robust program and a single simulation takes approximately 36h using the Antes parameter set on a Core2Quad Q6600 desktop machine. Also with 28 desktop machines available for the study, 3 grids could be run simultaneously. Thus computational expense was not a real concern in this case. Taking all factors into consideration, the Antes parameter set was subsequently used in this study.

3.5 CONCLUSION

In this chapter AutoDock's application in blind protein - peptide docking was evaluated using the crystal structure of KSHV-LANA bound to the nucleosome surface as reference complex. Various variables were investigated such as ligand flexibility, ligand position, search space size and the parameter set used. A novel method, grid-based docking, was also described in which a large search space could be divided into smaller, partially overlapping search spaces to improve sampling of large protein receptors. This method was verified and proved to be superior to the traditional method of docking. These experiments also gave an indication of the binding energy which could be expected for both the specific binding of a basic peptide to the nucleosome surface

(~ 8 kcal/mol and higher) and for non – specific DNA binding (between ~ 3 and 7 kcal/mol). Finally, based on the results, it was decided to use a rigid, grid-based docking with a total search space of 90Å x 40Å x 90Å and the parameter set described by Antes [13] in the blind docking of the H3 N-terminal tip to the nucleosome surface.

3.6 REFERENCES

1. Halperin, I., Ma, B., Wolfson, H. & Nussinov, R. (2002). Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins* **47**, 409-443.
2. Alexandre MJJ, B. (2006). Flexible protein – protein docking. *Current Opinion in Structural Biology* **16**, 194-200.
3. Andrusier, N., Mashiach, E., Nussinov, R. & Wolfson, H. J. (2008). Principles of flexible protein – protein docking. *Proteins* **73**, 271-289.
4. Krissinel, E. (2010). Crystal contacts as nature's docking solutions. *Journal of Computational Chemistry* **31**, 133-143.
5. Moreira, I. S., Fernandes, P. A. & Ramos, M. J. (2010). Protein – protein docking dealing with the unknown. *Journal of Computational Chemistry* **31**, 317-342.
6. Wang, C., Bradley, P. & Baker, D. (2007). Protein – Protein Docking with Backbone Flexibility. *Journal of Molecular Biology* **373**, 503-519.
7. Morris, G. M., Huey, R., Lindstrom, W., Sanner, M. F., Belew, R. K., Goodsell, D. S. & Olson, A. J. (2009). AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *Journal of Computational Chemistry* **30**, 2785-2791.
8. Krieger, E., Koraimann, G. & Vriend, G. (2002). Increasing the precision of comparative models with YASARA NOVA – a self-parameterizing force field. *Proteins* **47**, 393-402.
9. Huang, Z. & Wong, C. F. (2009). Conformational selection of protein kinase A revealed by flexible-ligand flexible-protein docking. *Journal of Computational Chemistry* **30**, 631-644.

10. Hetényi, C. & van der Spoel, D. (2002). Efficient docking of peptides to proteins without prior knowledge of the binding site. *Protein Science* **11**, 1729-1737.

11. Morris, G. M., Goodsell, D. S., Huey, R., Hart, W. E., Halliday, S., Belew, R. K. & Olson, A. J. (2001). *User's Guide - AutoDock: Automated Docking of Flexible Ligands to Receptors Version 3.0.5*.

12. Barbera, A. J., Chodaparambil, J. V., Kelley-Clarke, B., Joukov, V., Walter, J. C., Luger, K. & Kaye, K. M. (2006). The nucleosomal surface as a docking station for Kaposi's sarcoma herpesvirus LANA. *Science* **311**, 856-861.

13. Antes, I. (2010). DynaDock: A new molecular dynamics-based algorithm for protein – peptide docking including receptor flexibility. *Proteins* **78**, 1084-1104.

14. Morris, G. M. & Lim-Wilby, M. (2008). Molecular Docking. In *Molecular Modeling of Proteins* (Kukol, A., ed), pp. 365-382, Humana Press, Totowa.

CHAPTER 4

A Molecular Dynamics analysis of the role of epigenetic modifications on the structure of the histone H3 N-terminal tail

4.1 INTRODUCTION

The structure of the N-terminal tail of histone H3, and the possible role that epigenetic modifications of this tail may have on this structure, has largely been overlooked in the literature, mostly because of the difficulty in directly addressing this question with current biochemical and biophysical techniques. Molecular Dynamics (MD) does, however, provide a valuable insight into the structure of peptides and the effect of post-translational modifications (PTMs) on such structures. Benchmarking studies have shown that modern force fields can be used in MD approaches to faithfully derive the time-average solution structure of several synthetic peptides [1-3]. In this chapter we investigated the structure of the H3 tail by MD, focusing particularly on the effect that epigenetic histone modification patterns associated with different transcriptional states of the DNA molecule, had on this structure. We were specifically interested in studying the possibility that PTMs of the H3 tail could modulate the tail structure, and influence the possible participation of the tail in the stabilization of the higher-order 30 nm fiber structures of chromatin.

Histone H3 has the longest tail and the most PTM sites among the core histones [4]. Many of these PTMs are found individually or in combination, and their appearance is marked by specific cellular events. Guenther and co-workers showed that K4 tri – methylation, K9 acetylation, K14 acetylation and K36 tri – methylation on H3 marked actively transcribed genes in human embryonic cells [5]. On the other hand, Eberlin and co – workers showed that K9 and K27 di – methylation, and S10 and

S28 phosphorylation on H3 produced a unique structure, which occurred exclusively between early prophase and early anaphase of mitosis, where chromatin is highly condensed. Interestingly H3 S10 phosphorylation is found throughout mitosis and meiosis and is found in combination with PTMs marking chromatin condensation, such as K9 methylation [6, 7] as well as with PTMs that mark chromatin decondensation, for example K9 acetylation [8]. Two early biochemical assays showed that hyper – acetylation of the H3 and H4 tails decreased the initial melting temperature of nucleosome cores [9] and decreased the linking number [10] of the nucleosomal DNA. This suggested that the entering – and exiting DNA were more mobile in the acetylated core. Given that the tails were associated with the nucleosome at low salt concentrations [11], and the H3 tail exited the nucleosome between the two nucleosomal DNA gyres, it appears possible that the H3 tail could fold over the exiting DNA. If the H3 tail then proceeded to contact and bind to the lateral surface of the nucleosome, it could act as a molecular bracket, fastening the ends of the nucleosomal DNA to the nucleosome. Acetylation of the H3 tail could, in principle, abolish the binding of the H3 tail in the nucleosome, releasing the bracket, and allow increased movement of the terminal nucleosomal DNA, in keeping with the observed melting and linking number studies. Here we study the structure of the H3 tail to try and understand such structural mechanisms in the contribution of the H3 tail to chromatin structure.

MD was used to investigate the structural effects of PTMs associated with active and inactive genes; S10 phosphorylation; different levels of K9 methylation; and hyper – acetylation on the histone H3 tail. We also verified YASARA's ability to faithfully reproduce expected structures over the long time scale MD runs by using an alanine homopeptide, known to have a strong helical character [12, 13], and a glycine homopeptide, known to have a weak helical character[13, 14].

4.2 METHODS

4.2.1 Structure preparation

All structural manipulation and preparation was carried out using the program YASARA [15].

4.2.2 H3 N-terminal peptides

The highest resolution crystal structure available for the nucleosome core particle (NCP), PDB code: 1KX5 [16], was used as starting structure. The structure was loaded into YASARA and the following actions were performed:

All the molecules except molecule A (first H3 molecule in the file) were deleted. A split point was inserted between residue 43 and residue 44, and all residues after the split point were deleted. The remaining peptide was checked for errors and subsequently cleaned.

This peptide was used as the unmodified (“wild type”) H3 and as starting structure for the modeling of PTMs in the peptide.

4.2.3 Modeling of PTMs

The modifications were not built from scratch. Instead, NMR or X-ray structures were retrieved and the specific modified residues were excised and inserted into the peptide according to the following method:

The relevant source structures were downloaded from the Protein Databank (See Table 4.1)

Table 4.1 PDB source structures of modified residues used in MD simulations. Three letter abbreviations in brackets are indicative of the original names for the residues in the structure files, which were changed for this study.

Post Translational Modification	Source PDB Code	Chain in PDB structure	Residue Number in PDB structure	Residue three letter abbreviation	Reference
Mono-methylated Lysine	3HNA	P	9	M1L (MLZ)	17
Di – methylated Lysine*	2KVM	B	27	M2L (MLY)	18
Tri – methylated Lysine*	2L11	B	9	M3L	19
Acetylated Lysine*	2KWJ	B	14	ALY	20
Phosphorylated Serine	2C1N	C	10	SEP	21

* For NMR structures, the first structure in the file was selected.

The source PDB file was loaded into YASARA and the following method were applied:

The modified residue was identified and all other atoms were deleted. The residue was cleaned using the *CleanAll* command and, where needed, the residue was renamed. The residue was saved in a pdb format structure file.

For insertion of a modified residue into a structure, the following procedure was followed for each modified residue in the structure, after the starting structure was loaded into YASARA:

The modified residue position was identified and split points were inserted on either side of the original unmodified residue. The structure object was split, yielding three new objects: the N-terminal fragment, the original residue and the C-terminal fragment.

The relevant modified residue was loaded into YASARA and renamed and renumbered to resemble the unmodified residue. It was then superimposed onto the unmodified residue. The *CleanAll* command was executed and the original residue was deleted.

The modified residue was joined to the N-terminal object and the C-terminal object was joined to the new object to form one object. At this point no bonds existed between the 3 fragments.

A bond was created between the backbone amide atom of the modified residue and the backbone carbonyl carbon of the N-terminal fragment. A second bond was created between the backbone carbonyl carbon atom of the modified residue and the backbone amide atom of the C-terminal fragment.

The modified residue was renamed to its alternate name, and the structure was cleaned using the *CleanAll* command. All remaining split points in the structure were removed and the resulting structure was saved in a pdb format file.

4.2.4 Control peptides

The unmodified, "wild-type" (WT) peptide was used as template, and all residues were replaced with alanine residues for the positive helix control or with glycine residues for the negative sheet control, using the *SwapRes* command in YASARA. In both cases residues 39 - 43 were re-fixed to restrain the new side chains. The structures were saved in pdb format files.

4.2.5 Experimental systems

The following systems were set up for simulation, using the previously described methodology:

Table 4.2 Systems simulated in the MD experiments

System Title	Modifications
WT	None
HYPER_ALY	All K Ace
ACTIVE	K4 + K36 Me3, K9 + K14 Ace
INACTIVE	K9 + K27 Me2, S10 + S28 Pho
K9ME1_S10PHO	K9 Me1 + S10 Pho
K9ME2_S10PHO	K9 Me2 + S10 Pho
K9ME3_S10PHO	K9 Me3 + S10 Pho
K9ME1	K9 Me1
K9ME2	K9 Me2
K9ME3	K9 Me3
K9ACE_S10PHO	K9 Ace + S10 Pho
ALA_POS_CTRL	None
GLY_NEG_CTRL	None

* Ace – acetylation, Me1 – mono – methylation, Me2 – Di – methylation, Me3 – tri – methylation, Pho - phophorylation

4.2.6 Molecular Dynamics (MD) simulation setup

Systems described in Table 4.2 were simulated using the program YASARA [22].

Residue 39 - 43 were fixed and the structure was placed in a rectangular simulation cell with an extension of 16 Å around all atoms. The final size of the simulation cell was X = 83.21 Å × Y = 54.48 Å × Z = 65.82 Å.

Simulations were run at 298 K using a weakly-coupled Berendsen thermostat [22] and at a pressure of 1 bar using a Solvent Probe pressure control mode in YASARA.

The AMBER03 force field [23] was used. Long-range electrostatics were treated with the Particle Mesh Ewald method [24], and a non-bonded cut-off of 7.86 Å was implemented.

A multiple time step for integration was used, where intra-molecular forces were calculated every 2 fs and inter-molecular forces every (2x1.25) 2.5 fs.

The simulation was run at a pH of 7.0 and the cell was neutralized using sodium and chloride counter-ions. The salt concentration in the simulation cell was 0.154 M. The cell was neutralized using the Neutralization experiment [25] in YASARA.

For modified residues new force field parameters were derived using YASARA Auto SMILES [26-29], YASARA's automatic parameter assignment procedure for organic molecules.

The hydrogen-bond network were optimized (to give more stable trajectories) using the YASARA method.

4.2.7 Energy minimization

To remove bumps and correct the covalent geometry, the structure was energy-minimized with the AMBER03 force field [23], using a 7.86 Å force cut-off and the Particle Mesh Ewald algorithm [24] to treat long-range electrostatic interactions. After removal of conformational stress by a short steepest descent minimization, the procedure continued by simulated annealing (time step 2 fs, atom velocities scaled down by 0.9 every 10th step) until convergence was reached, i.e. the energy improved by less than 0.05 kJ/mol per atom during 200 steps.

4.2.8 MD production run

Each system was simulated on the High Performance Computing (HPC) cluster at the University of the Free State using YASARA. Each system was simulated for a minimum of 500 ns and was run on 24 CPU cores on single nodes. Coordinates were saved every 25 ps, yielding 20 000 time points for each trajectory.

4.2.9 Secondary structure prediction using internet-based tools

To obtain some reference point as to what may be expected in terms of the secondary structure composition of the unmodified WT tail, the secondary structure of the 43 - residue H3 N-terminal tail was predicted using the following applications:

- CFSSP: Chou & Fasman Secondary Structure Prediction Server (http://www.biogem.org/tool/chou-fasman/) [30]
- Jpred - A consensus method for protein secondary structure prediction at University of Dundee (http://www.compbio.dundee.ac.uk/~www-jpred/) [31]
- JUFO – Protein secondary structure prediction from sequence (neural network) (http://www.meilerlab.org/web/view.php?section=0&page=6) [32]
- NSurfP – Protein Surface Accessibility and Secondary Structure Predictions (http://www.cbs.dtu.dk/services/NetSurfP/) [33]
- Porter – University College Dublin (http://distill.ucd.ie/porter/) [34-36]
- PredictProtein – Sequence Analysis, Structure and Function Prediction (http://www.predictprotein.org/) [37, 38]
- Prof – Cascaded Multiple Classifiers for Secondary Structure Prediction (http://www.aber.ac.uk/~phiwww/prof/) [39]
- Scratch Protein Predictor – SSpro and SSpro8 (http://scratch.proteomics.ics.uci.edu/index.html) [40, 41]
- NPS@ Secondary Structure Consensus Prediction – (http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_seccons.html) [42]

  Includes the methods:

  - SOPM (Geourjon and Deléage, 1994)
  - HNN (Guermeur, 1997)
  - GOR IV (Garnier et al., 1996)
  - SOPMA (Geourjon and Deléage, 1995)

  - SIMPA96 (Levin, 1997)
  - DPM (Deléage and Roux, 1987)
  - DSC (King and Sternberg, 1996)
  - GOR I (Garnier et al., 1978)
  - GOR III (Gibrat et al., 1987)
  - PHD (Rost and Sander, 1993)
  - PREDATOR (Frishman and Argos, 1996)
- PSIPRED - Protein secondary structure prediction based on position-specific scoring matrices (http://bioinf.cs.ucl.ac.uk/psipred/) [43]

4.2.10 Secondary Structure and Hydrogen bonding analysis

The secondary structure and hydrogen bonding present in the histone H3 tail was analyzed using YASARA and custom Python (http://www.python.org) scripts as described in Chapter 2. All results were imported into the simDB database, also described in Chapter 2, and downstream analyses used data from the database. Custom Python scripts were used to investigate differences in hydrogen bonding patterns between simulations by constructing frequency matrices and an R script was used to graph the analyses. In each case, except the controls, the matrices were normalized by subtracting the WT matrix.

The source code for all custom scripts can be found on the CD included with this document or at http://cbio.ufs.ac.za

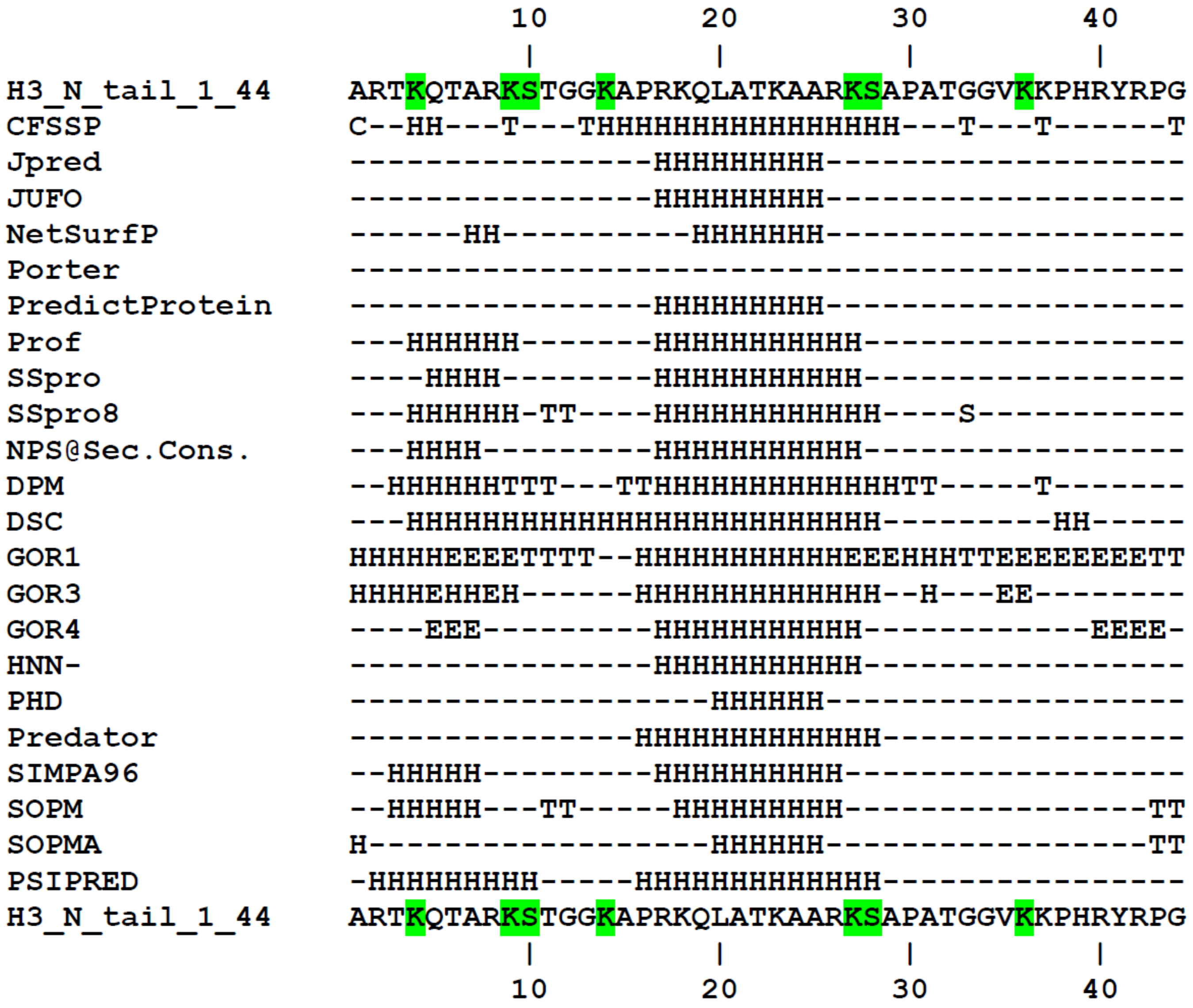
4.2.11 Clustering analysis

All trajectories were converted to "xtc" format using the YASARA script md_convert. Subsequently the GROMACS [44] program g_cluster was used to cluster each trajectory using the single – linkage clustering algorithm with a 15 Å cut-off value.

4.3 RESULTS

4.3.1 Secondary structure prediction

Figure 4.1 shows the secondary structures predicted for the H3 tail. Two α – helices were predicted, the shorter of the two approximately between K4 and S10 and the longer helix approximately between R17 and R26. The longer helix was predicted by more algorithms and was flanked by the modifiable K14, K27 and S28 residues. The shorter helix was flanked by the modifiable K4, K9 and S10 residues. Thus the modified residues seemed to be located in positions where they could potentially influence the secondary structure of the tail. However, this approach gives one-dimensional information on the sequence location of a limited number of secondary structures. Although the identification of the α-helices bordered by residues shown to be subject to reversible modification suggests a possible role for such secondary structures, no insight into secondary structures other than α-helices, β-strands and coiled regions is obtained from such approaches. Also, no information on spatial arrangements of the tail, particularly the stabilization of tertiary structures due to long-range interactions, and the contribution of proline residues, often involved with peptide backbone "kinking" [45], can be gained from secondary structure predictions. Thus, to gain a detailed insight into the structures assumed by the H3 tail, we proceeded to study the tail by MD. We first looked at the unmodified WT tail, and then continued with simulations to systematically study the effect of epigenetic modifications of residues on the tail structure.



```
                              10        20        30        40
                               |         |         |         |
H3_N_tail_1_44     ARTKQTARKSTGGKAPRKQLATKAARKSAPATGGVKKPHRYRPG
CFSSP              C--HH---T---THHHHHHHHHHHHHHHH---T---T------T
Jpred              ----------------HHHHHHHHH-------------------
JUFO               ----------------HHHHHHHHH-------------------
NetSurfP           ------HH----------HHHHHHH-------------------
Porter             --------------------------------------------
PredictProtein     ----------------HHHHHHHHH-------------------
Prof               ---HHHHHH-------HHHHHHHHHHH-----------------
SSpro              ----HHHH--------HHHHHHHHHHH-----------------
SSpro8             ---HHHHHH-TT----HHHHHHHHHHHH----S-----------
NPS@Sec.Cons.      ---HHHH---------HHHHHHHHHHH-----------------
DPM                --HHHHHHTTT---TTHHHHHHHHHHHHHTT-----T-------
DSC                ---HHHHHHHHHHHHHHHHHHHHHHHHH---------HH-----
GOR1               HHHHHEEEETTTT--HHHHHHHHHHHEEEHHHTTEEEEEEEETT
GOR3               HHHHEHHEH------HHHHHHHHHHHHH--H---EE--------
GOR4               ----EEE---------HHHHHHHHHHH------------EEEE-
HNN-               ----------------HHHHHHHHHHH-----------------
PHD                -------------------HHHHHH-------------------
Predator           ---------------HHHHHHHHHHHHH----------------
SIMPA96            --HHHHH---------HHHHHHHHHH------------------
SOPM               --HHHHH---TT-----HHHHHHHHH----------------TT
SOPMA              H------------------HHHHHH-----------------TT
PSIPRED            -HHHHHHHHH-----HHHHHHHHHHHHH----------------
H3_N_tail_1_44     ARTKQTARKSTGGKAPRKQLATKAARKSAPATGGVKKPHRYRPG
                               |         |         |         |
                              10        20        30        40
```

Figure 4.1 Secondary structure prediction of the 44-residue N-terminal tail of histone H3 using the algorithms indicated. The highlighted residues indicate residues which contained PTMs in this study.

4.3.2 Secondary structure during 500 ns explicit MD simulations

4.3.2.1 WT tail

Figure 4.2.A shows the evolution of the secondary structure in the WT H3 tail during an explicit, 500 ns MD simulation. Two α – helices were observed during the MD simulation of the unmodified WT tail, in agreement with the secondary structure prediction results. At the N-terminus of the tail an α – helix was stabilized between T3 and G12. It initially appeared at 8 ns, but did not remain stable and unfolded after 20 ns. It reappeared at 78 ns and was stabilized after 90 ns. Thereafter it remained relatively stable, unfolding into mostly hydrogen bonded turns after 230 ns for 60 ns and after 408 ns for about 12 ns.

The second α – helix was formed between L20 and A29. It first appeared at approximately 27 ns and was stabile up to 85 ns, where after a shorter α - helix between A21 and R26 was stable until 175 ns. Following its unfolding into hydrogen bonded turns for 15 ns and for 40 ns after 325 ns, it remained semi – stable for the rest of the simulation. Figure 4.2.B shows the percentage of time spent in particular secondary structure for each residue during the simulation and subsequently the stabilization of the two α – helices during the simulation as well as the greater length and stability of the N-terminal helix. This MD simulation provided insight into the structures of the unmodified tail. Since the unmodified tail in generally not observed in an eukaryotic cell, this simulation has little biological value, but does provide a basis from which to analyze the effect of PTMs on tail structure.

4.3.2.2 ACTIVE tail

We next studied the structures assumed by the N-terminal tail with epigenetic modifications associated with an "active" tail. Histone H3 with these modifications are typically found in nucleosomes associated with regions of the genome that are actively transcribed.

The secondary structure evolution of the ACTIVE tail (see Figure 4.3.B) during the MD simulation showed the complete destabilization of the two α – helices observed during the WT simulation. A shortened N-terminal α - helix between T3 and R8 made an unstable appearance between 55 ns and 73 ns before disappearing for the rest of the simulation. An α – helix between the acetylated K14 and Q19 appeared shortly after 50 ns and remained semi – stable until approximately 165 ns, disappeared, and only reappeared again briefly after 400 ns for 2 – 3 ns periods. It was replaced by hydrogen bonded turns for the duration of the simulation. Where the WT tail showed almost no β – strands or β – bridges during the simulation, the ACTIVE peptide, in contrast, showed the stabilization of several β – bridges during the simulation. β – Bridges are minimal β – sheets, with

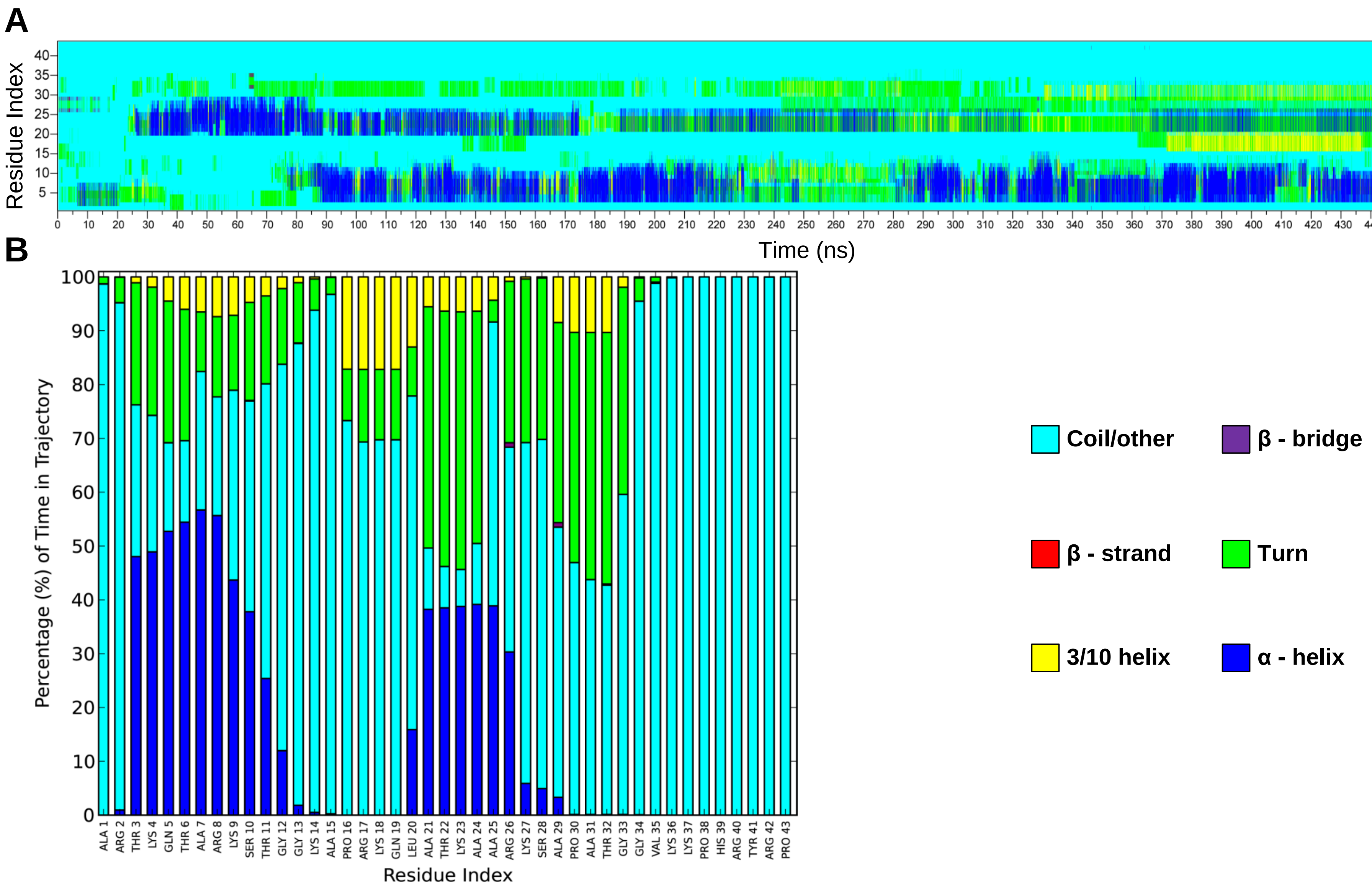


Figure 4.2 Secondary structure composition of the WT tail computed from a 500 ns explicit MD simulation using YASARA with a 250 ps sampling. The time evolution of the secondary structure for each residue during the simulation is shown in A. The percentage of time spent in a secondary structure element by each residue is shown in B.

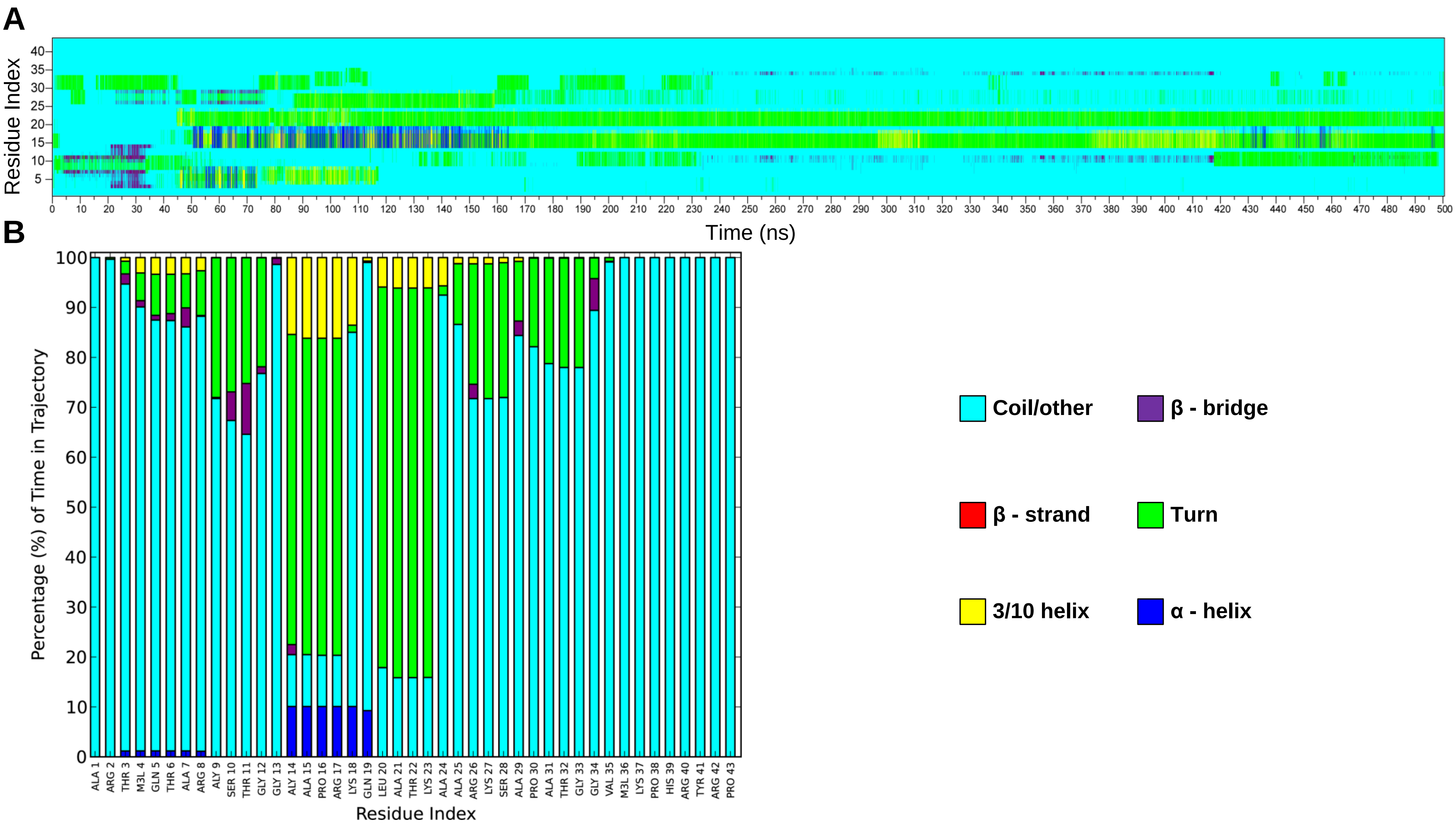


Figure 4.3 Secondary structure composition of the ACTIVE tail computed from a 500 ns explicit MD simulation using YASARA with a 250 ps sampling. The time evolution of the secondary structure for each residue during the simulation is shown in A. The percentage of time spent in a secondary structure element by each residue is shown in B.

an insufficient number of residues in a β – strand conformation to satisfy the definition of a β – sheet. Residues in these β – bridges appeared in regions where the α – helices in the WT tail were located (See Figure 4.3.A). It was, however, difficult to identify the exact structural nature of the β – bridges in the conformation because β – bridges can be formed between residues that were located in distal parts of the sequence. An example of this was seen between 230 ns and 420 ns: The β-bridges were formed between G33 and residues between G12 - K14 (See Figure 4.3). It was, however, not possible to determine the position of the β – bridges where more than two regions simultaneously contained this structural elements. An example was seen between 20 ns and 35 ns. There were two regions occupied by β – bridges: R26 - A29, S10 – K14 and T3 – A7. However, one cannot be completely certain between which residues the β – bridges were formed. The β – bridge partners were determined by using a combination of the secondary structure time evolution plots, the secondary structure histograms, the hydrogen bonding matrices and the clustered structures.

4.3.2.3 INACTIVE tail

In contrast to both the WT tail and the ACTIVE tail, the INACTIVE tail showed the appearance of a single α – helix between L20 and K27 after 15 ns which remained highly stable throughout the entire simulation (See Figure 4.4.A and Figure 4.4.B). Interestingly, this α – helix was accompanied by two regions shown to be in hydrogen bonded turns. The first of these was between K9 and G13 and was less stable than the second, between P30 and G34. The hydrogen bonded turn at P30 to G34 was more stable than the α – helix throughout the entire simulation period, as is seen in Figure 4.4.A. The region in which the N-terminal α - helix was stabilized during the WT simulation, was also devoid of any stable secondary structures.

4.3.2.4 HYPER – ALY tail

We next studied the structures assumed by an H3 tail peptide where the lysine residues were quantitatively acetylated. Although such a saturated acetylation state for the H3 tail peptide has

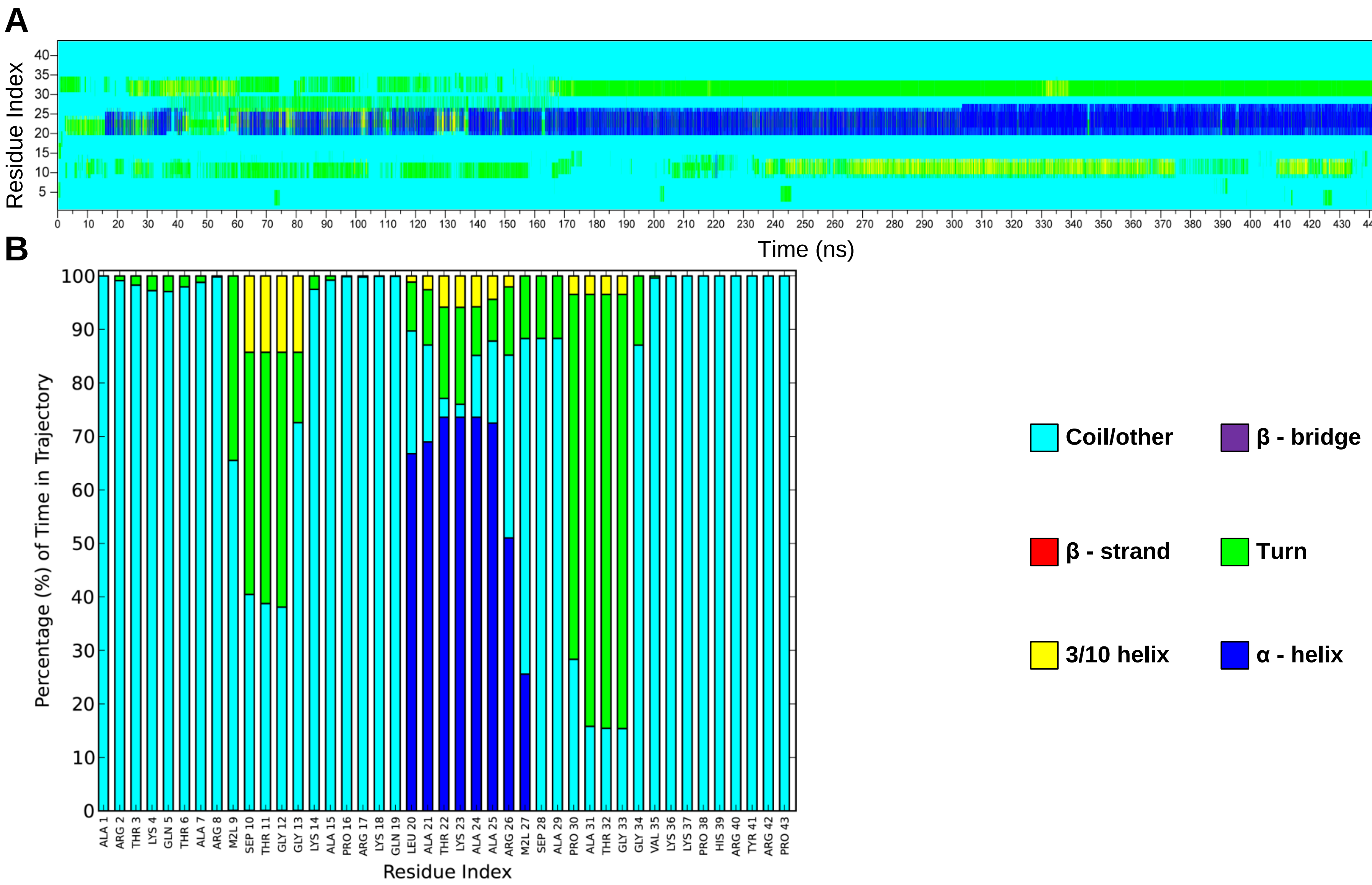


Figure 4.4 Secondary structure composition of the INACTIVE tail computed from a 500 ns explicit MD simulation using YASARA with a 250 ps sampling. The time evolution of the secondary structure for each residue during the simulation is shown in A. The percentage of time spent in a secondary structure element by each residue is shown in B.

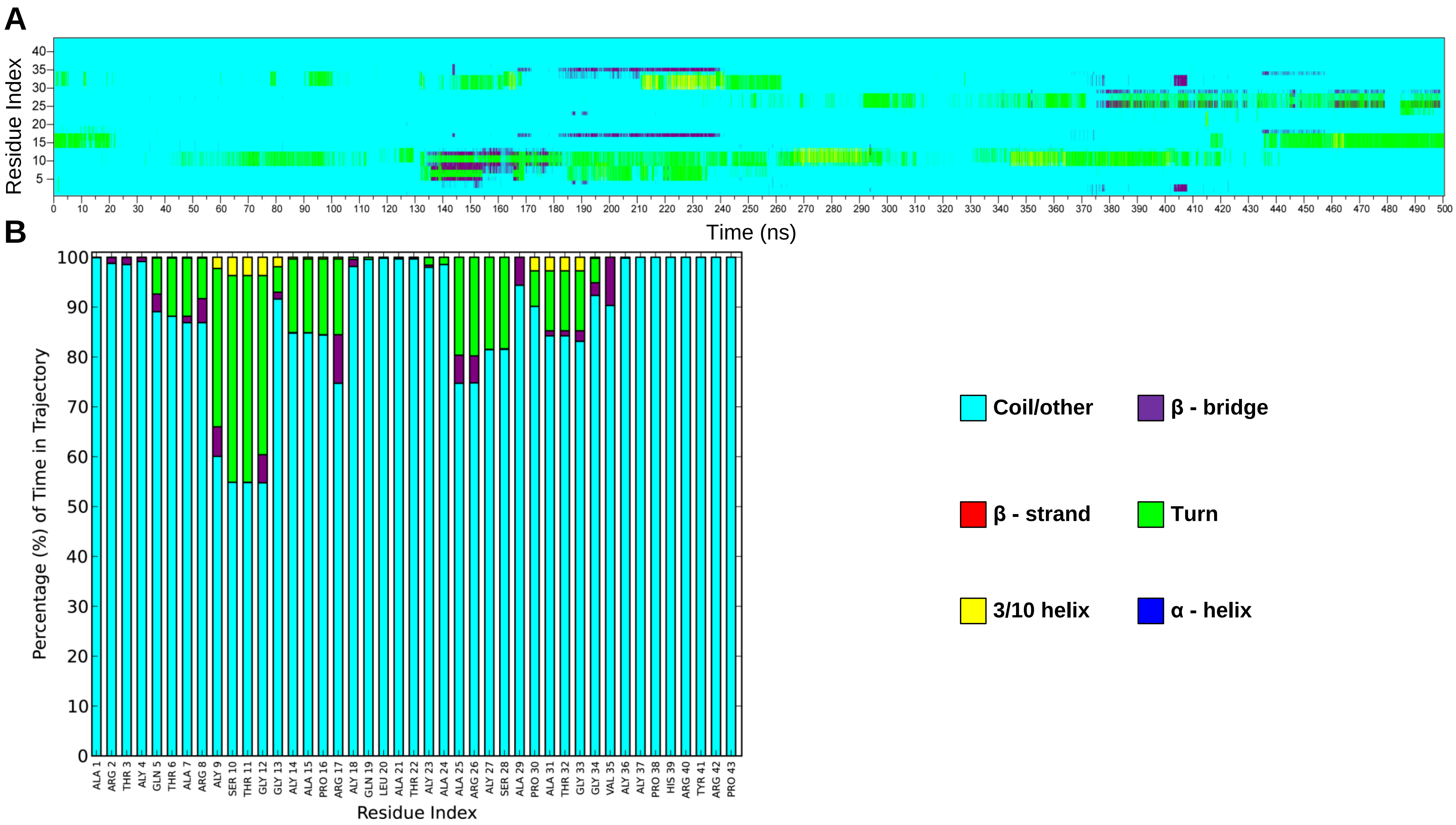


Figure 4.5 Secondary structure composition of the HYPER-ALY tail computed from a 500 ns explicit MD simulation using YASARA with a 250 ps sampling. The time evolution of the secondary structure for each residue during the simulation is shown in A. The percentage of time spent in a secondary structure element by each residue is shown in B.

never been observed in a cell, and is also likely not to be biologically relevant, this extreme modification allowed us to access the sensitivity of the MD approach to study the effect of PTMs on tail structure.

The HYPER-ALY tail did not show the stabilization of any α – helices whatsoever (see Figure 4.5). This was an interesting finding, since it is widely assumed that proximal positively charged lysine residues will destabilize an α-helix, and that acetylation of these residues would therefore contribute to helix stabilization. This assumption is not borne out by the MD simulation.

For the first 130 ns the peptide was devoid of any secondary structure elements, except for hydrogen – bonded turns: P30 – G33 (for ~10 ns), K14 - R17 (for ~20 ns) and K9 – G12 (~40 ns). This was followed by two β – bridges with hydrogen – bonded turns between the bridge – pairs, Q5 – R8 and K9 - G13, for ~15 ns. The Q5 – R8 β - bridge unfolded and the K9 – G13 bridge remained semi – stable for ~25 ns. Subsequently, a β – bridge was stabilized between G34 and R17 for ~ 60 ns and only briefly reappeared at 435 ns for ~25 ns. The β – bridge was accompanied by the stabilization of hydrogen bonded turns in the region between Q5 and G12. At ~ 75 ns a β – bridge with a hydrogen bonded turn in the middle was formed between A25 and A29 and remained semi – stable until the end of the simulation. Finally, the K14 – R17 hydrogen bonded turn observed at the start, reappeared at 435 ns and remained stable until the end of the simulation.

These secondary structural features were varied and comprehensively different from those of the WT tail. An MD simulation thus identified structural differences between peptides that differ at the level of reversible chemical modifications, and it is therefore reasonable to expect that an MD simulation should allow an analysis of the structural impact of epigenetic modifications on the structure of the H3 tail peptide.

4.3.2.5 K9ME1_S10PHO tail

It has often been shown in the literature that phosphorylation of S10 co-localized with K9 methylation [6, 7]. It was further shown that S10 had to be de-phosphorylated before K9 could be demethylated in order to change the epigenetic mark on K9 from a methyl to an acetyl [46], the latter associated with active transcription. It was therefore suggested that S10PHO was a "molecular safety switch" which had to be toggled prior to changing the modification mark on K9. We were interested to study whether, apart from such a proposed role, the phosphorylation of S10 also played a role in the structure of the H3 tail, particularly in combination with a methylated K9. We therefore studied the structure of mono- di and tri-methylated K9 in conjunction with S10PHO.

The K9ME1_S10PHO tail (Figure 4.6) showed the formation of an α – helix between T22 and A29 at ~ 20 ns, which remained stable until 100 ns before unfolding. Over the next 280 ns it refolded briefly before stabilizing at 380 ns and subsequently remained stable until the end of the simulation. A hydrogen bonded turn was also observed throughout the simulation, remaining the most stable between 280 ns and 390 ns. At 80 ns a β – bridge between T6/A7 and G12/G13 with a hydrogen bonded turn between them appeared, and remained semi - stable until the end of the simulation.

The most striking difference between the secondary structures assumed by the tail peptide in the K9ME1 and K9ME1S10PH forms, is the increase in the stability of β-turns in the vicinity of K9, and the virtual complete abolishment of the α-helix centered at position 6 in the un-phosphorylated peptide (compare Figure 4.6 with Figure 4.9).

4.3.2.6 K9ME2_S10PHO tail

We next investigated the effect of S10 phosphorylation on the K9ME2 peptide (Figure 4.7). An α – helix, L20 – R26, was stable between ~125 ns and 250 ns and briefly started reappearing at 480 ns. The region between R26 and A29 was in a hydrogen bonded turn throughout the majority of the simulation. A second region, A31 - G34 was also found to be rich in hydrogen bonded turns throughout the simulation, though less abundant than the former region. Another region rich in

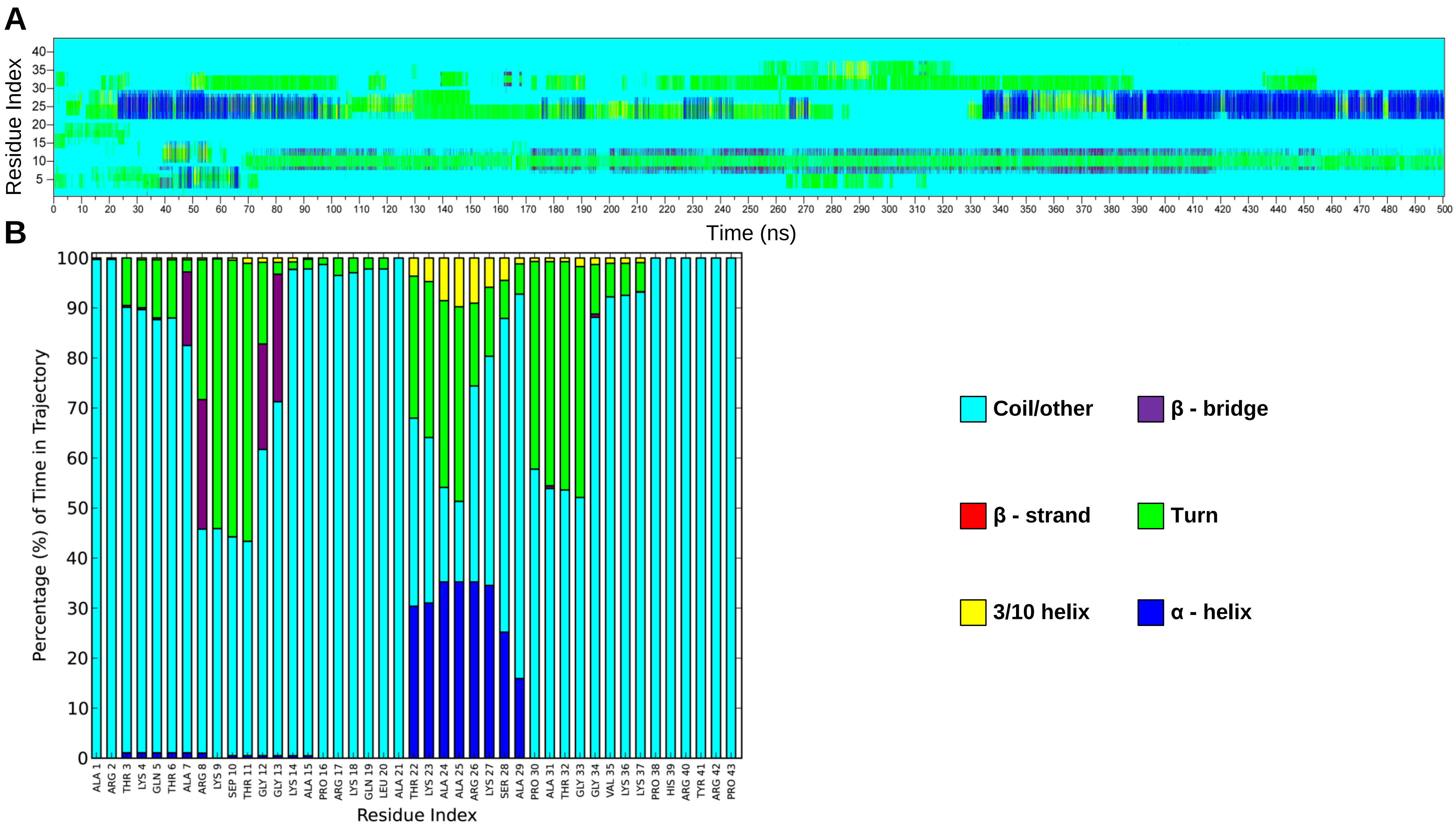


Figure 4.6 Secondary structure composition of the K9ME1_S10PHO tail computed from a 500 ns explicit MD simulation using YASARA with a 250 ps sampling. The time evolution of the secondary structure for each residue during the simulation is shown in A. The percentage of time spent in a secondary structure element by each residue is shown in B.

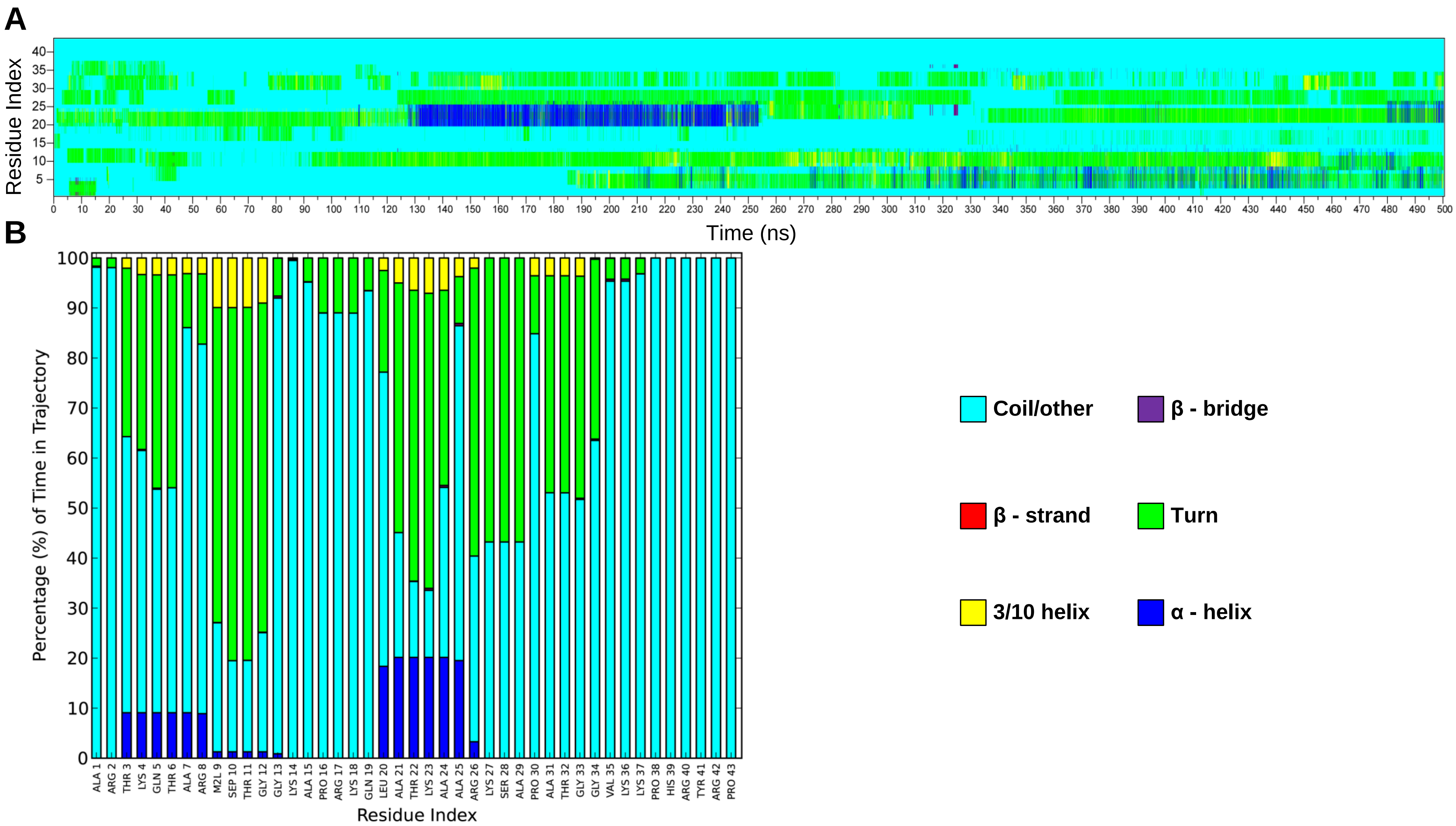


Figure 4.7 Secondary structure composition of the K9ME2_S10PHO tail computed from a 500 ns explicit MD simulation using YASARA with a 250 ps sampling. The time evolution of the secondary structure for each residue during the simulation is shown in A. The percentage of time spent in a secondary structure element by each residue is shown in B.

hydrogen bonded turns was found between K9 and G12, and remained stable throughout the simulation except for the period between ~55ns and ~90 ns. Finally the region between T3 and T6 fluctuated between existing in a hydrogen bonded turn and extending to R8 to form an α – helix from ~185 ns onwards. Again, an increase in the stability of β-turns and a decrease in the occurrence of α-helices were observed in the phosphorylated peptide (compare Figure 4.7 with Figure 4.10).

4.3.2.7 K9ME3_S10PHO tail

We were interested in ascertaining whether phosphorylation of S10 caused the same conformational stability shifts in the peptide that contained a tri-methylated K9. The K9ME3_S10PHO tail showed the formation of various α – helical regions (Figure 4.8). The first of these were observed as early as ~7 ns between K23 and K28 and remained stable until ~20 ns. Another α – helical region, A21 – K27, appeared at ~37 ns and remained stable until 90 ns. The region R8 – A15 was also found in an α – helix from ~143 ns – 180 ns. The most stable α – helix was observed in the region R2 – R8, from ~285 ns to ~414 ns. Interestingly the former α – helix extended its length by one residue on both the C-terminal - and N-terminal end to include A1 and K9. An α – helix was also found in A24 – A29 for the time period ~292 ns to ~340 ns. The final α – helix was observed in the region P16 – A21, between ~363 ns and ~384 ns. A β – bridge was also observed at ~153 ns between A1 and R26 and remained stable until 205 ns. Subsequently only one other β – bridge was observed in P16 - Q19 and A29 – G33. The region K9 – G13 as well as the region T22 – A29 were observed to be rich in hydrogen bonded turns when not involved in other secondary structures described throughout the simulation. The region P30 – G34 was also shown to form hydrogen bonded turns, however it was not as abundant as in the case of the latter regions. The difference between the stabilities of secondary structures in the K9ME3 and K9ME3S10PHO peptides were not as striking as seen for the mono- and di-methylated peptides in the phosphorylated and un-phosphorylated forms, respectively (compare Figure 4.8 and Figure 4.11).

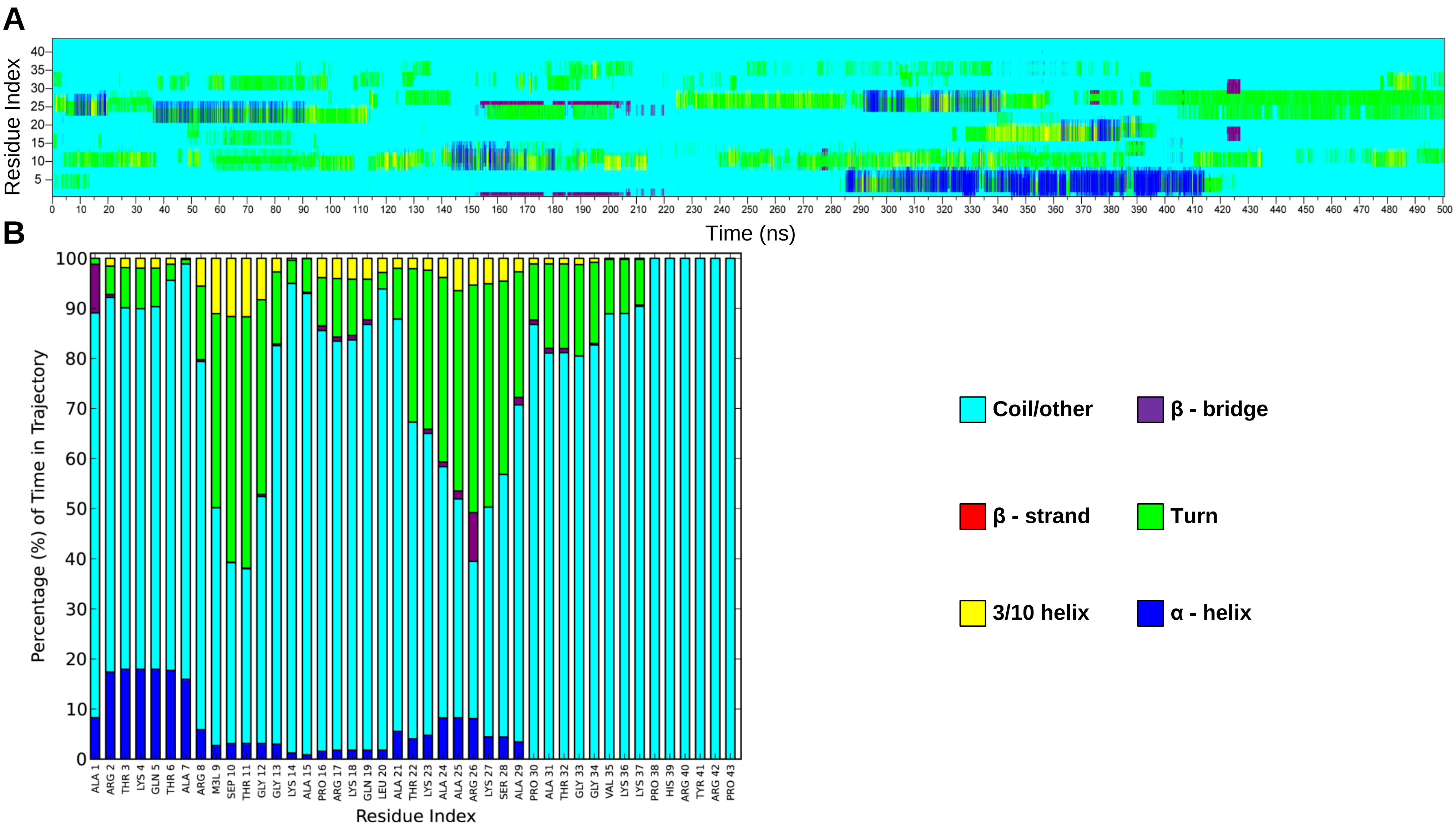


Figure 4.8 Secondary structure composition of the K9ME3_S10PHO tail computed from a 500 ns explicit MD simulation using YASARA with a 250 ps sampling. The time evolution of the secondary structure for each residue during the simulation is shown in A. The percentage of time spent in a secondary structure element by each residue is shown in B.

4.3.2.8 K9ME1 tail

Methylation of histone H3 lysine 9 is associated with transcriptional repression [47]. The mono-, di- and tri-methylated state is generally observed in heterochromatic regions, although the mono-methylated lysine has been detected in areas of low transcriptional activity. There is a comprehensive literature that shows the binding of chromo-domain containing proteins to methylated K9 [47]. Most strikingly, the association of the chromo-domain HP1 protein in *Drosophila* is found in silent genomic regions [48]. Thus, it is clear that such epigenetic modifications are associated with a mechanism of molecular signaling and binding of specific proteins in local regions of the genome. However, we were interested in determining possible differences between the different methylation states of K9 in an H3 peptide, to study the possibility that methylation of this residue could also serve a structural role in epigenetic regulation.

The results of a 500 ns MD simulation run of the H3 peptide with K9ME1 is shown in Figure 4.9. The K9ME1 tail contained a large amount of α – helical content. At ~26 ns an α – helix was stabilized in the region R2 – R8. The helix remained stable until ~160 ns, unfolded into a collection of hydrogen bonded turns, R2 - T6 and then reappeared for 5 ns after 170 ns. Interestingly the α – helical region would occasionally extend to both N – terminal - and C – terminal ends to include A2 and K9. The former α – helix extended toward the C – terminal end and formed a larger α – helix in the region A1 – G13 at 180 ns of which the core (R2 – T11) α – helix remained stable until ~330 ns. After ~330 ns the extended α – helix unfolded to stabilize the R2 – T6 α – helix again until 350 ns. In turn the α –helix unfolded into a hydrogen bonded turn region, which remained in the simulation until ~395 ns. Subsequently the region only contained random coil content until the end of the simulation. A third α – helix was also observed in the region P16 – A21. It originated in a series of hydrogen – bonded turns at 25 ns and gradually folded into an α – helix at ~125 ns. The α – helix was semi – stable up to ~230 ns, after which it was more stable until unfolding at ~280 ns. The α – helix refolded again at ~295 ns and remained stable until ~405 ns, after which it gradually started to unfold into a series of hydrogen bonded turns at ~475 ns. Another α – helix

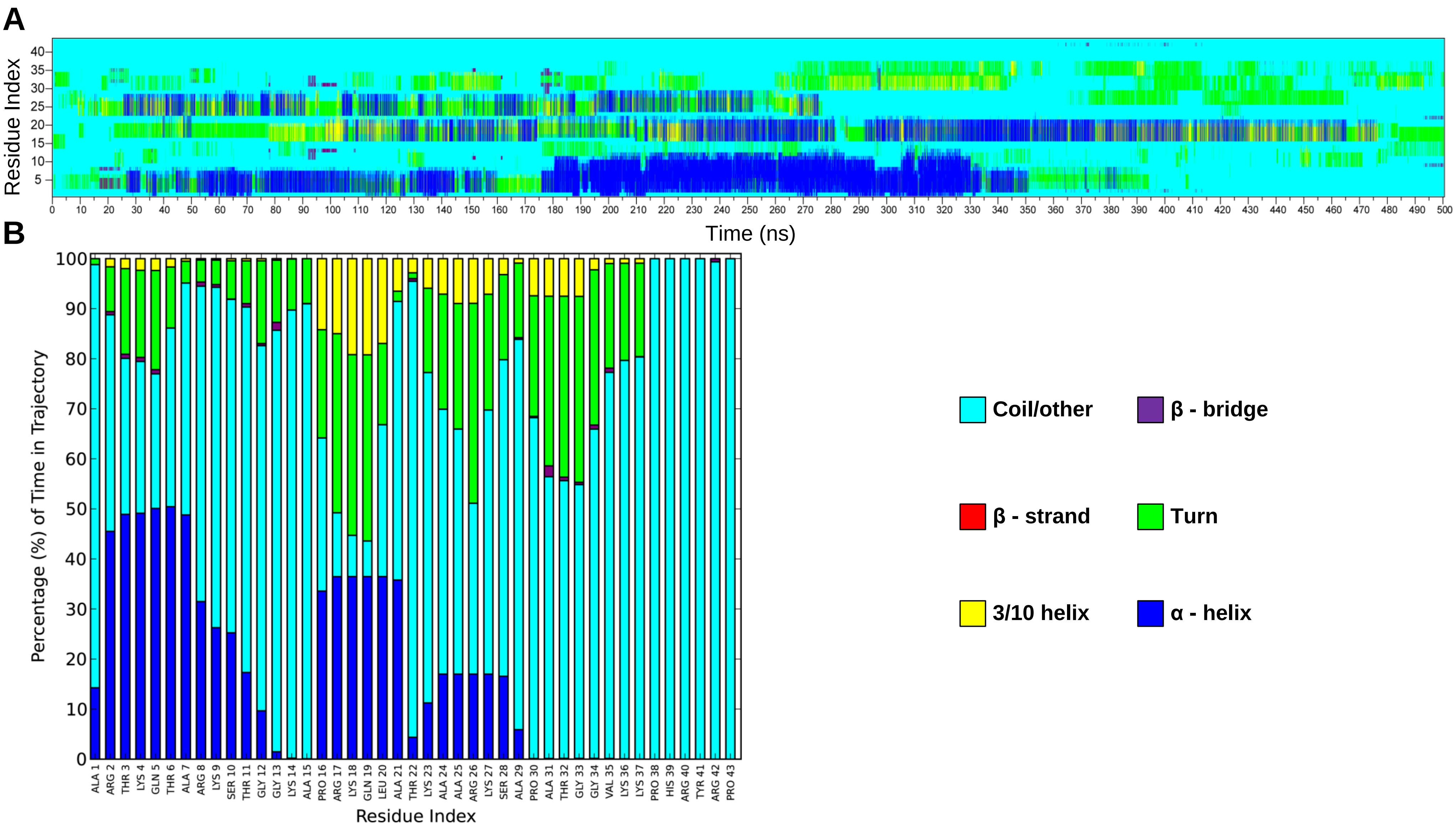


Figure 4.9 Secondary structure composition of the K9ME1 tail computed from a 500 ns explicit MD simulation using YASARA with a 250 ps sampling. The time evolution of the secondary structure for each residue during the simulation is shown in A. The percentage of time spent in a secondary structure element by each residue is shown in B.

was formed in the region T22 – A29 at ~25 ns and only remained semi – stable, constantly switching between an α – helix and a series of hydrogen bonded turns for variable lengths of time. After ~280 ns this region remained in a coil conformation with hydrogen bonded turns appearing again between ~370 ns and ~465 ns. Two regions of hydrogen bonded turns were also observed during the simulation: A31 – GLY34 and V35 – K37. Often times these regions were joined to form a single larger region of hydrogen bonded turns. Some β – bridge elements were also observed. These, however, appeared only briefly, and thus appeared not be of significance. The most striking difference between the stability of secondary structures over the 500 ns simulation period between the WT and the K9ME1 tail peptide is the splitting of the α-helix centered at approximately residue 24 in the WT tail into two α-helices that meet at approximately residue 24 in the K9ME1 simulation. This simulation showed that, in principle, a single methylation adduct on K9 could have a significant impact on the tail peptide structure.

4.3.2.9 K9ME2 tail

The K9ME2 tail was also rich in α – helical content (see Figure 4.10). First, an α – helix was formed in the region R2 – R8. It formed at ~5 ns and remained stable until ~128 ns, where after it unfolded into a series of hydrogen bonded turns and remained as such until ~165 ns. At ~165 ns it refolded and remained stable until ~215 ns and unfolded again. It appeared briefly at ~222 ns and was present until ~243 ns before unfolding again.

At ~295 ns an α – helix emerged in the region T3 – G12 and remained stable until ~365 ns, where approximately half of the N-terminal side of the α – helix unfolded at ~385 ns. From ~385 ns the extended α – helix started refolding, reaching stability at ~ 415 ns and remaining stable until ~465 ns. At ~465 ns the α – helix started to unfold, forming a series of hydrogen bonded turns by ~475 ns, where after it remained until the end of the simulation.

A second main region of α – helical content folded from a series of hydrogen – bonded turns in the region, A21 – R26, and first appeared at ~130 ns. Like the former main α – helical region, this α – helix was also extended to include the region L20 – A29. After an initial constant switching

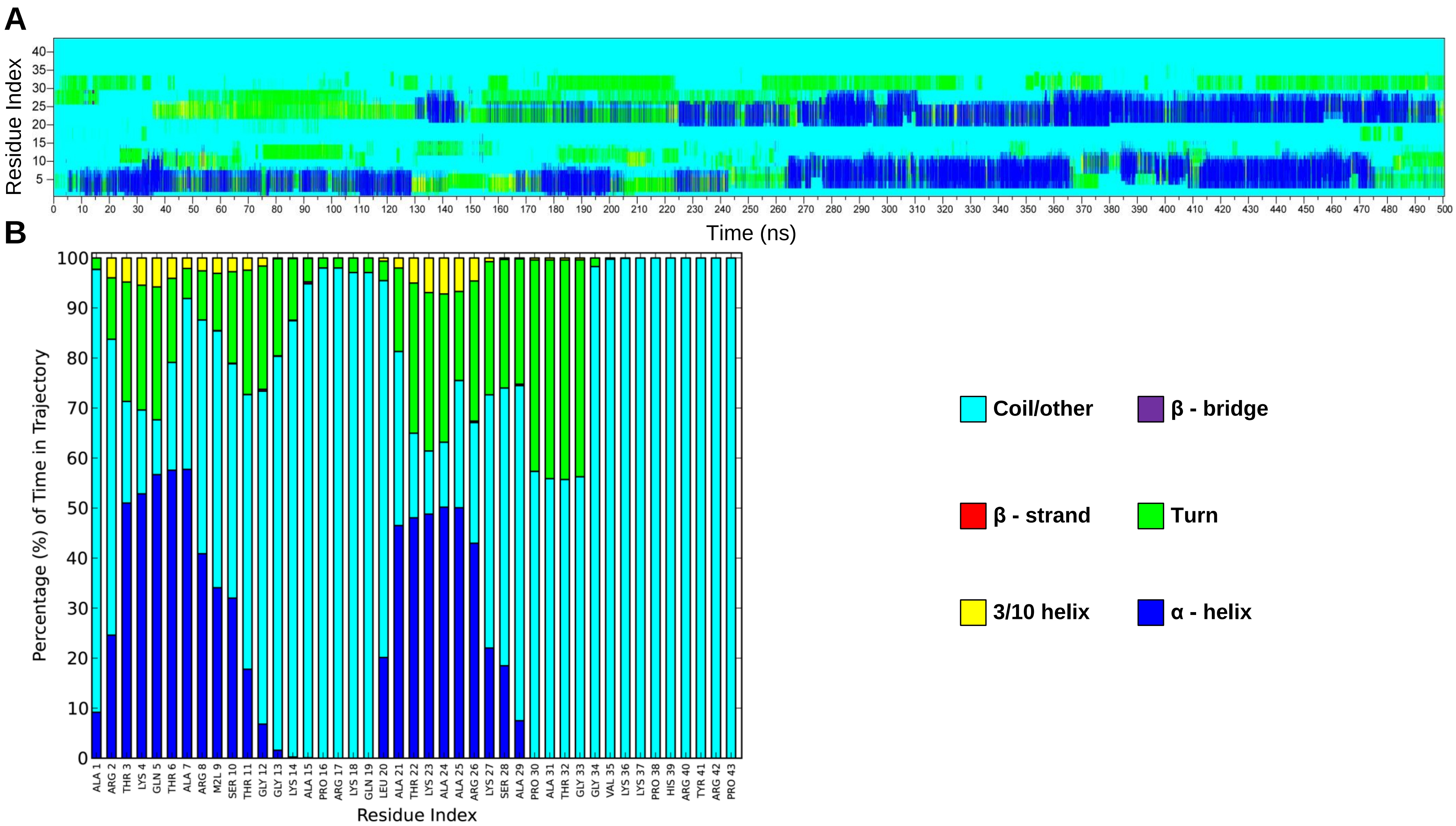


Figure 4.10Secondary structure composition of the K9ME2 tail computed from a 500 ns explicit MD simulation using YASARA with a 250 ps sampling. The time evolution of the secondary structure for each residue during the simulation is shown in A. The percentage of time spent in a secondary structure element by each residue is shown in B.

between α – helical structure and a hydrogen – bonded turn between ~145 ns and ~225 ns, the region stably adopted an α – helical structure until ~498 ns. As seen in most other simulations, there was the presence of a region of hydrogen – bonded turns in the region P30 – G33 throughout most of the simulation. The most striking difference between the K9ME1 and K9ME2 modified tail peptides was the disappearance of β-turns in the region centered on approximately residue 19, and the joining of the two α-helices centered on positions 19 and 26 into a single α-helix centered on position 24 in K9ME2, similar to that observed in the WT peptide.

4.3.2.10 K9ME3 tail

The K9ME3 simulation showed a very low α – helical content in contrast to a more pronounced presence of β – bridges (see Figures 4.10). At first a β – bridge with a central hydrogen – bonded turn, Q19 – K23 was formed at ~13 ns and remained stable until ~25 ns. This was followed by the formation of a β – bridge forming between Q19 and A24-K27, which appeared at ~30 ns and remained stable until ~ 50 ns, where after it unfolded. At ~60 ns a new β – bridge was formed between R17 and A25 and remained stable until ~175 ns. During the time periods ~253 - ~258 ns, ~300 - ~318 ns and ~373 - ~475 a very unstable α – helix, P16 – A21, was present. At ~320 ns at new β – bridge was formed between A15 and A29, which remained stable until the end of the simulation. Interestingly this β – bridge at times expanded to include G13/K14/P16 and A31/P30/S28. The tail was also very rich in hydrogen bonded turns. The main regions identified were R2 – T6, K9 – G13, L20 – A24, P30 – G34 and T32 – K36, which periodically occupied the structure of the tail during the simulation.

The almost complete absence of α-helical regions in the K9ME3 peptide compared to the K9ME2 peptide was the most striking difference. This significant difference was intriguing, since the K9ME2 and K9ME3 tails were shown to co-localise in the genome, suggesting identical functions. This apparent discrepancy may suggest that methylation of lysine may be a more important molecular beacon for the recruitment of non-histone proteins to specific regions in the genome as opposed to playing a structural role.

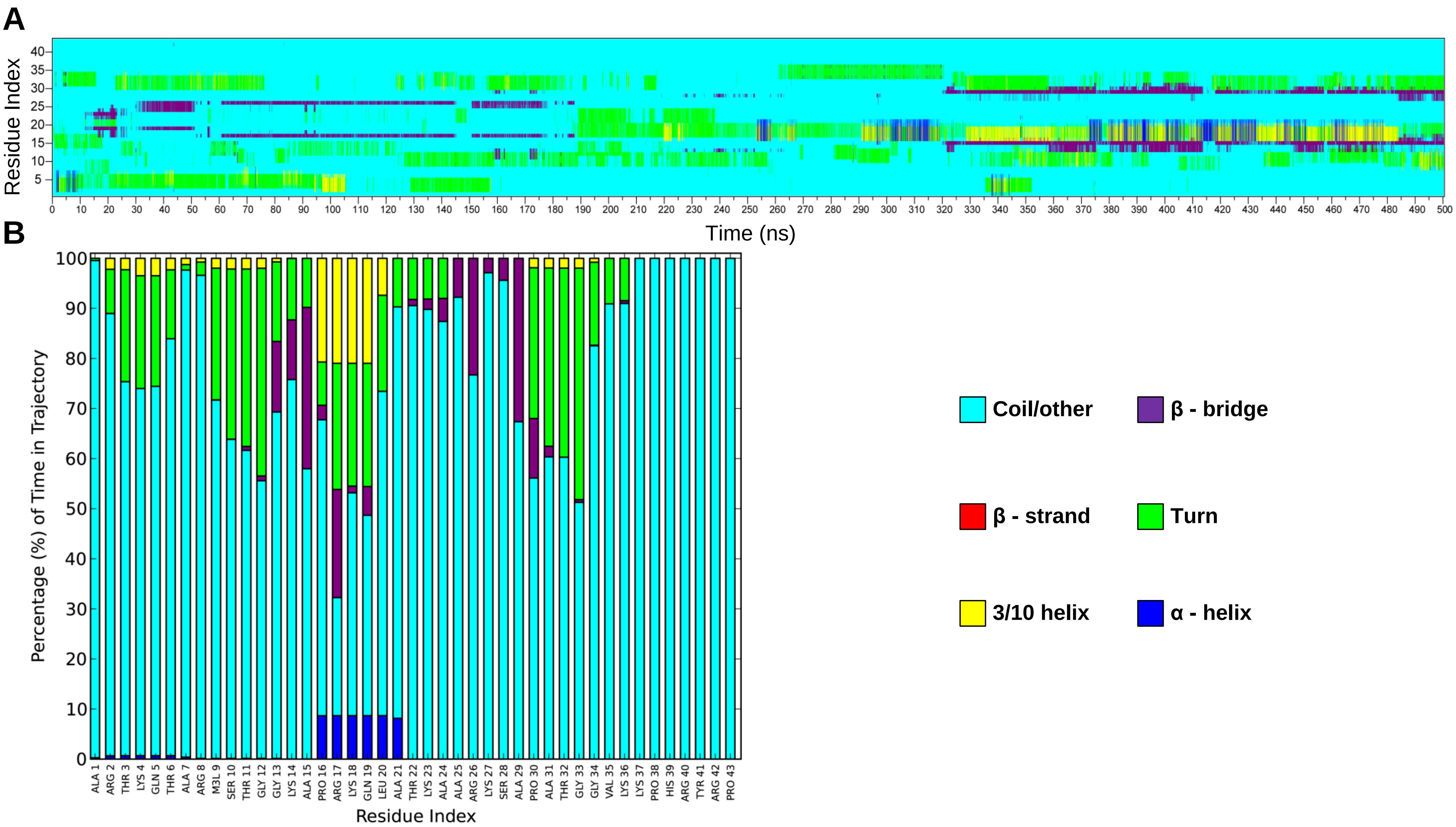


Figure 4.11 Secondary structure composition of the K9ME3 tail computed from a 500 ns explicit MD simulation using YASARA with a 250 ps sampling. The time evolution of the secondary structure for each residue during the simulation is shown in A. The percentage of time spent in a secondary structure element by each residue is shown in B.

4.3.2.11 K9ACE_S10PHO tail

The K9ACE_S10PHO tail displayed some α – helical content, however overall the tail was very hydrogen bonded turn rich (see Figure 4.12). An α – helical region, R2 – R8 was first observed at ~77 ns and remained stable until ~90 ns before unfolding into hydrogen bonded turns. It reappeared sporadically until ~ 280 ns. A very unstable α- helical region, S10 – A15, also briefly made an appearance between ~280 and ~330 ns. The final α – helical region, L20 – R26 initially appeared at ~90 ns and unfolded into hydrogen bonded turns after ~100 ns. After a brief appearance between ~240 ns and ~245 ns, the α – helical region refolded after ~325 ns and remained semi – stable until the end of the simulation. The major regions involved in hydrogen bonded turns (which did not fold into α - helices) were K9 – G12, K14 – R17 and A29 – G34. These hydrogen bonded turns were present throughout the simulation.

4.3.2.12 ALA_POS_CTRL tail

We finally tested the accuracy of the MD simulations to provide information on the secondary structures that different regions of the tails could assume. As a basic test we wanted to confirm that a peptide sequence known to prefer an α-helical conformation indeed displayed extensive α-helical stability during the MD run. We made use of a peptide in the same starting conformation as the H3 tail peptide used above, but with all residues substituted with alanine. As expected, the ALA_POS_CTRL tail were exclusively found in an α – helix (see Figure 4.13). Three major α – helices were formed during the first ~45 ns, after which the two central matrices merged into one, followed by the merger of the two terminal tips with the larger α – helix by ~ 175 ns. During the rest of the simulation the tail was found in one large α – helix, with some instability at the terminal ends of the peptide, as expected.

4.3.2.13 GLY_NEG_CTRL tail

In contrast to the ALA_POS_CTRL tail, the GLY_NEG_CTRL, which was expected to strongly resist folding into an α-helical conformation, indeed showed the absence of any α – helical content

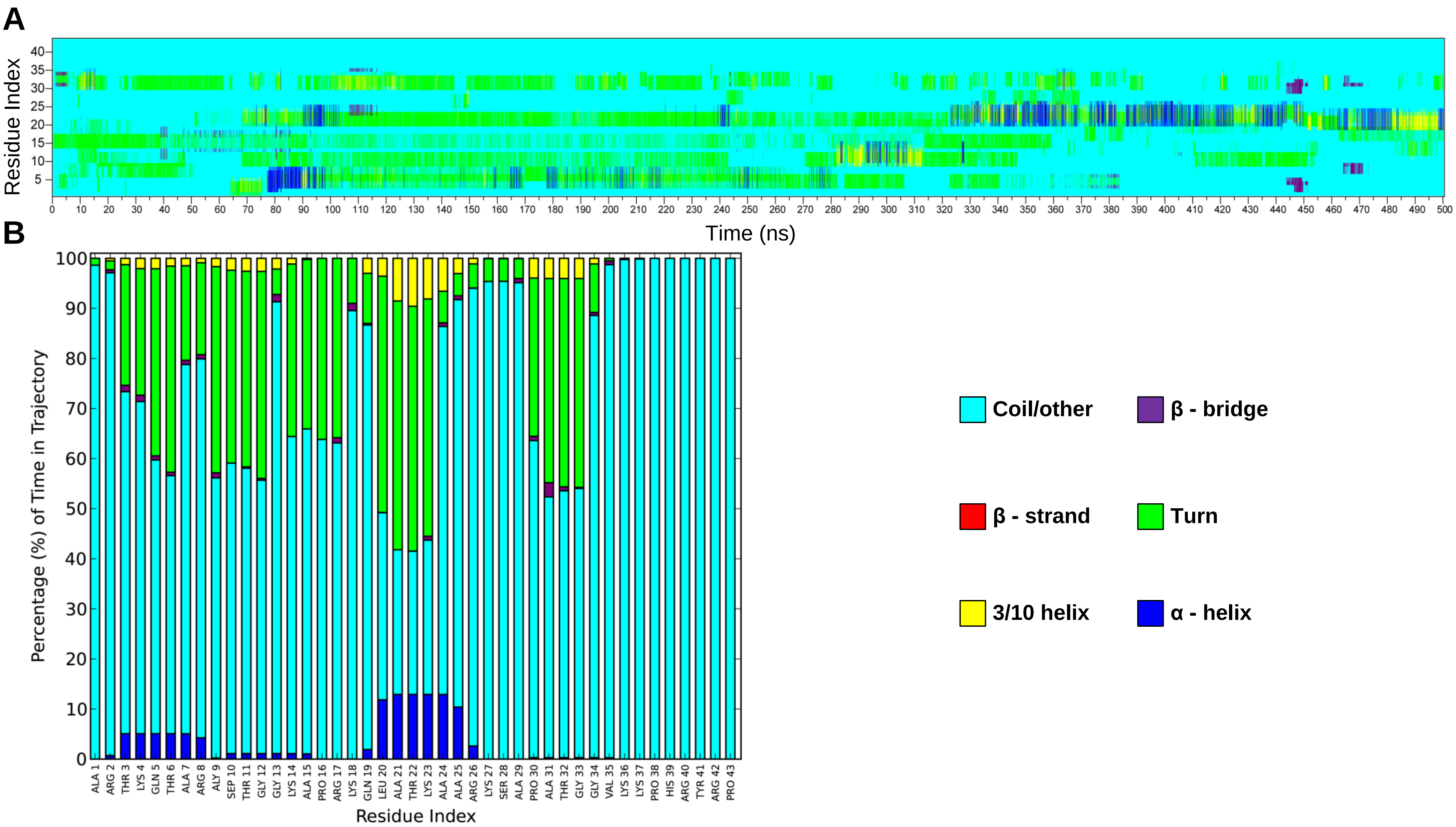


Figure 4.12 Secondary structure composition of the K9ACE_S10PHO tail computed from a 500 ns explicit MD simulation using YASARA with a 250 ps sampling. The time evolution of the secondary structure for each residue during the simulation is shown in A. The percentage of time spent in a secondary structure element by each residue is shown in B.

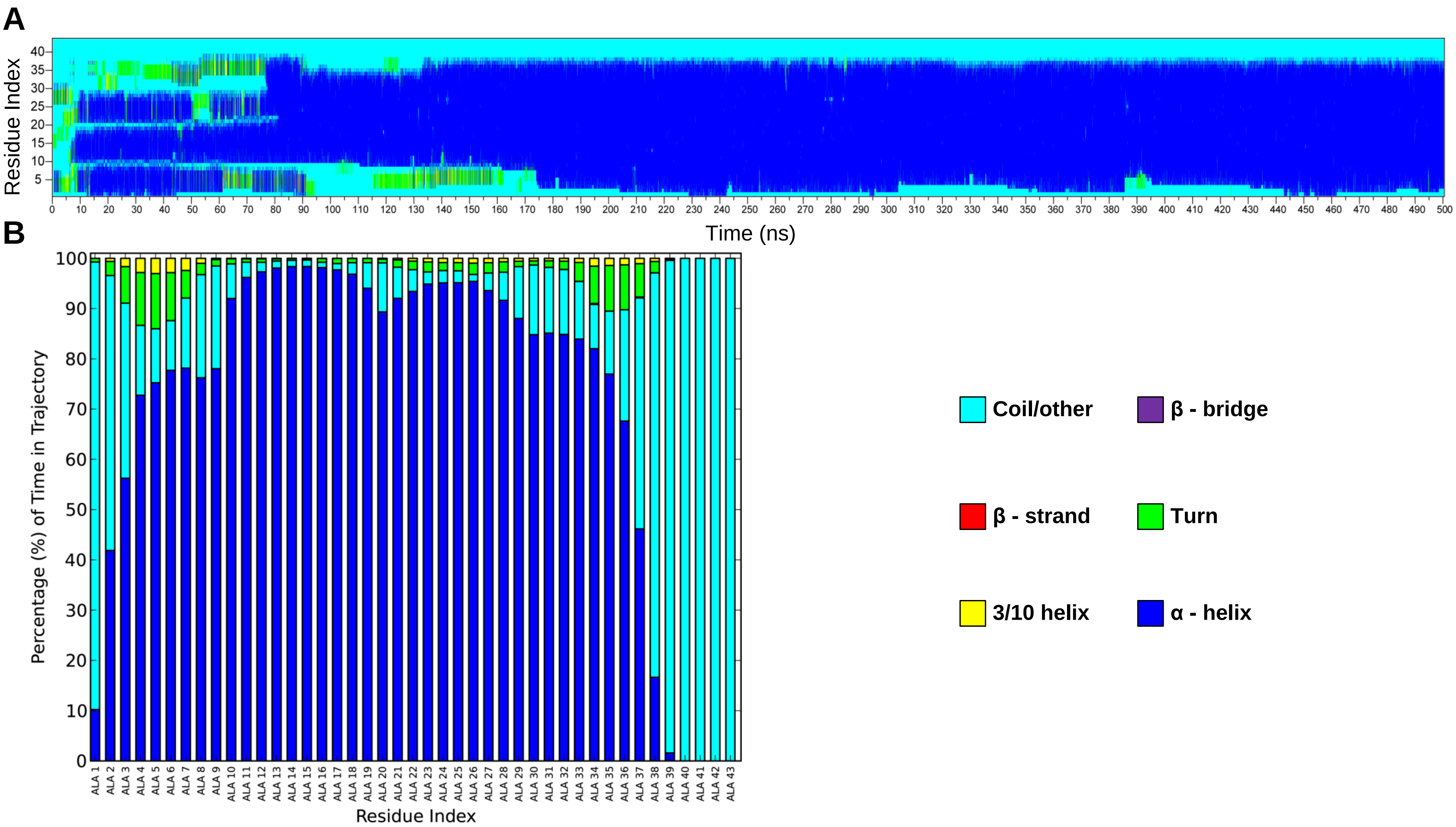


Figure 4.13 Secondary structure composition of the ALA_POS_CTRL tail computed from a 500 ns explicit MD simulation using YASARA with a 250 ps sampling. The time evolution of the secondary structure for each residue during the simulation is shown in A. The percentage of time spent in a secondary structure element by each residue is shown in B.

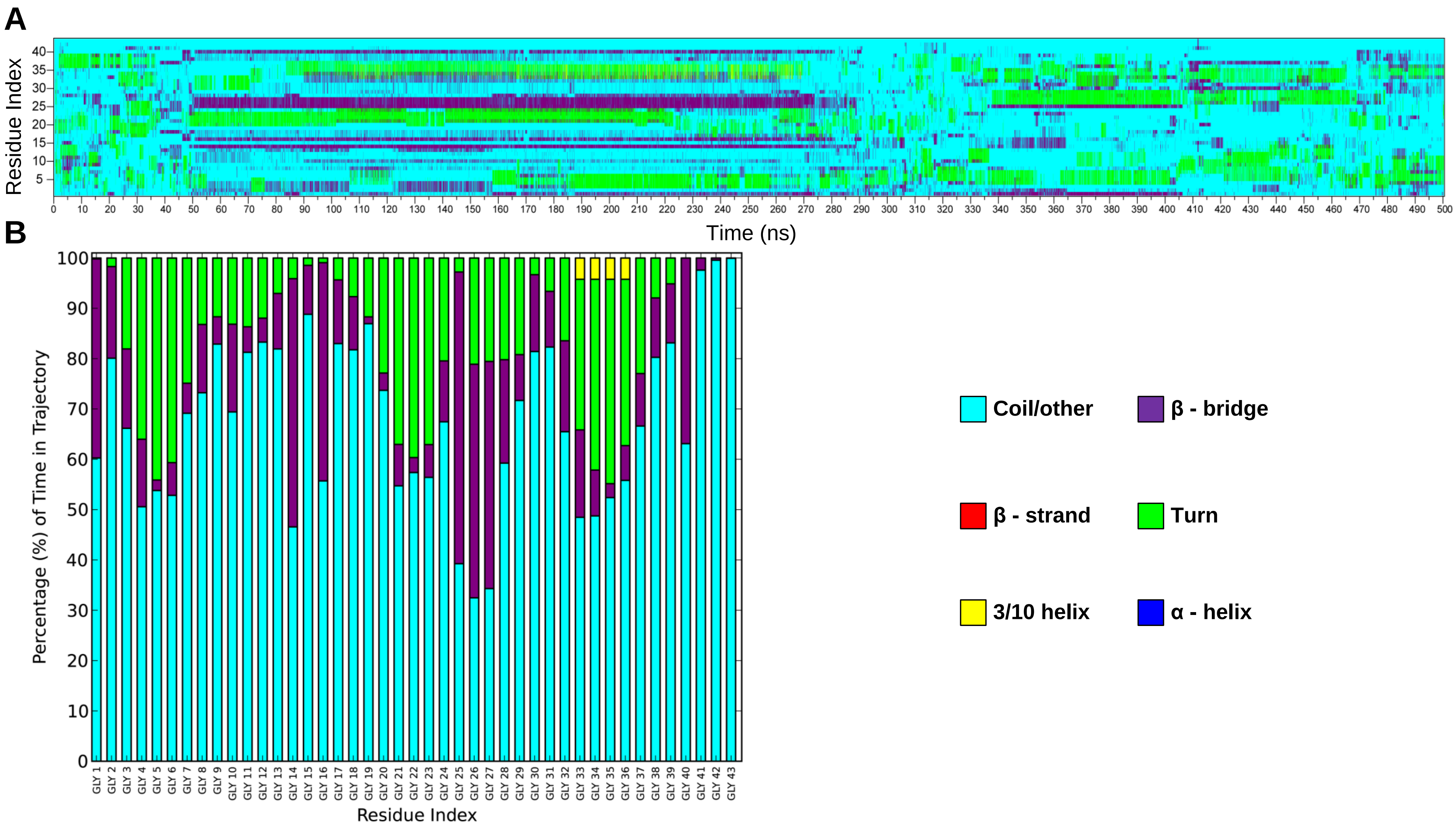


Figure 4.14 Secondary structure composition of the GLY_NEG_CTRL tail computed from a 500 ns explicit MD simulation using YASARA with a 250 ps sampling. The time evolution of the secondary structure for each residue during the simulation is shown in A. The percentage of time spent in a secondary structure element by each residue is shown in B.

(see Figure 4.14). The tail spent the majority of the time in a mixture of β – bridge, hydrogen – bonded turns and random coil elements. This was also expected and will be discussed below. The most stable structure was formed between ~50 ns and ~280 ns.

4.3.3 Hydrogen Bonding Analysis

In the section above we have presented data on the stability of various secondary structures assumed by the H3 tail peptides in an unmodified state, and in the presence of various epigenetic modifications. In this section, the MD results will be analyzed in terms of hydrogen bonding that were present in the folded conformations.

Figure 4.15 shows the total number of hydrogen bonds formed within each of the 500 ns explicit MD simulations. In terms of the overall number of hydrogen bonds, all the H3 simulations were within a similar range. The ALA_POS_CTRL stood out and produced almost double the amount of hydrogen bonds compared to the other simulations, because of the co-axial hydrogen bonds present in an extended α-helix. Thus, the differences in the structures of tails with various epigenetic modifications were not caused simply by the number of stabilizing hydrogen bonds, but were most likely due to specific hydrogen bond patterns allowed by the epigenetic modifications. This will be investigated below. It is important that the hydrogen bonding pattern analysis is viewed in the context of the secondary structure development over time, as well as with the secondary structure histograms. This approach may allow the identification of residues that played the most important role in stabilizing the specific secondary structure elements observed during the MD run. Hydrogen bonds which occurred outside of secondary structure elements were considered, as they could have played an indirect role in stabilizing or destabilizing the secondary structure elements observed. Also note that instances referring to a hydrogen bond pair or simply a pair refer to the hydrogen bond formed between one residue and another residue

4.3.3.1 ALA_POS_CTRL tail

In the case of the ALA_POS_CTRL peptide, the entire tail was described as an α – helix in the secondary structure analysis. In the hydrogen bonding pattern (Figure 4.16.A) the α – helix was clearly observed, with hydrogen bonding pairs spaced 4 residues apart (The ideal spacing for an α– helix is 3.6 residues). Also, a correspondence was observed between the magenta pairs (15 000 – 20 000 occurrences) and the stability of the α - helices during the 500 ns MD run.

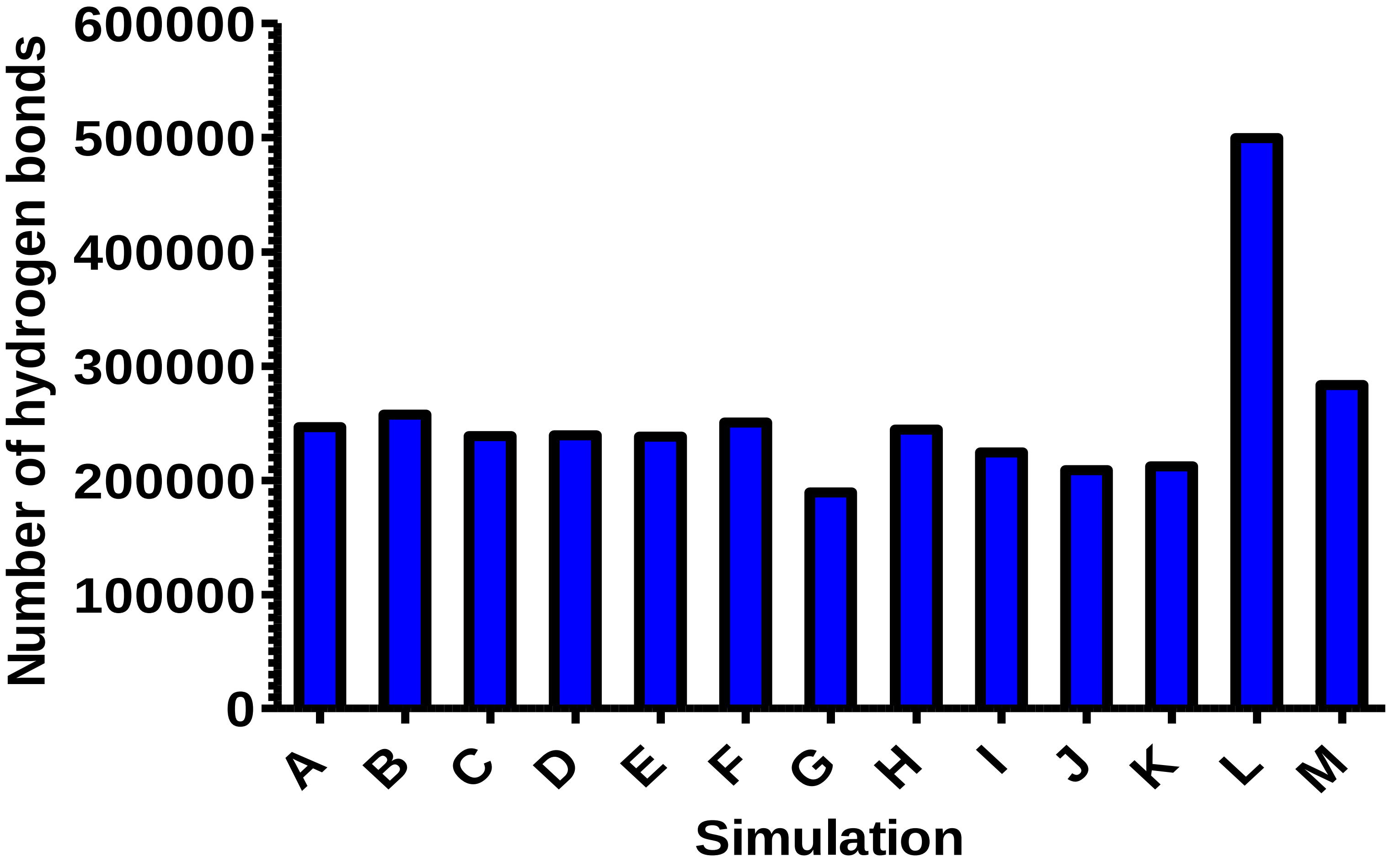


Figure 4.15 Total number of hydrogen bonds obtained from the WT (A), ACTIVE (B), INACTIVE (C), HYPER_ALY (D), K9ME1_S10PHO (E), K9ME2_S10PHO (F), K9ME3_S10PHO (G), K9ME1 (H), K9ME2 (I), K9ME3 (J), K9ACE_S10PHO (K), ALA_POS_CTRL (L) and GLY_NEG_CTRL (M) H3 tail from 500 ns explicit MD simulations

4.3.3.2 GLY_NEG_CTRL tail

The secondary structure analysis did not show any regular secondary structure pattern within the GLY_NEG_CTRL_tail (see Figure 4.14), a finding that is borne out by the hydrogen bonding pattern (Figure 4.16.B). The latter is characterized by an erratic pattern of hydrogen bonding. Only one hydrogen bond pair occurred between 10 000 and 15 000 times It is important to note this stark contrast between this erratic hydrogen bonding pattern and the highly regular pattern associated with an α – helix.

4.3.3.3 WT tail

The number of hydrogen bonds formed between a particular pair of residues did not exceed 15000 (See Figure 4.17) for any of the residue pairs during the MD run. The highest number of hydrogen bond pairs occurred between T3-T6, T3-A7, T6 – S10, A21 – A24 and A21 – A25. All of these bond pairs occurred between 10 000 and 15 000 times in the simulation and within the region of the two α – helices indentified in the secondary structure analysis: R2 – G12/G13 and L20 – A29. Three out of the five bond pairs belonged to the R-G12/G13 α – helix and only the last two bond pairs belonged to the L20 – A29 α – helix according to the spacing of the residues involved (3 or 4 residues). In the bond pairs identified between 5000 – 10 000 times, three bond pairs were associated to the R2 – G12/G13 α – helix, compared to only one bond pair associated to the L20 – A29 α – helix. The sets of hydrogen bond pairs flanking the L20 – A29 region represented the two flanking hydrogen – bonded turns observed in the time evolution of secondary structure plot This bond pair occurred between 5000 and 10 000 times. Hydrogen bond pairs between R17 and Q19, and R17 and A24 also occurred between 1000 and 5000 times during the simulation. This was quite interesting as these regions were close to and part of the L20 – A29 α – helix respectively. (Figure 4.2.A). Finally, an interesting long – range hydrogen bond pair was identified between R17 and T32.

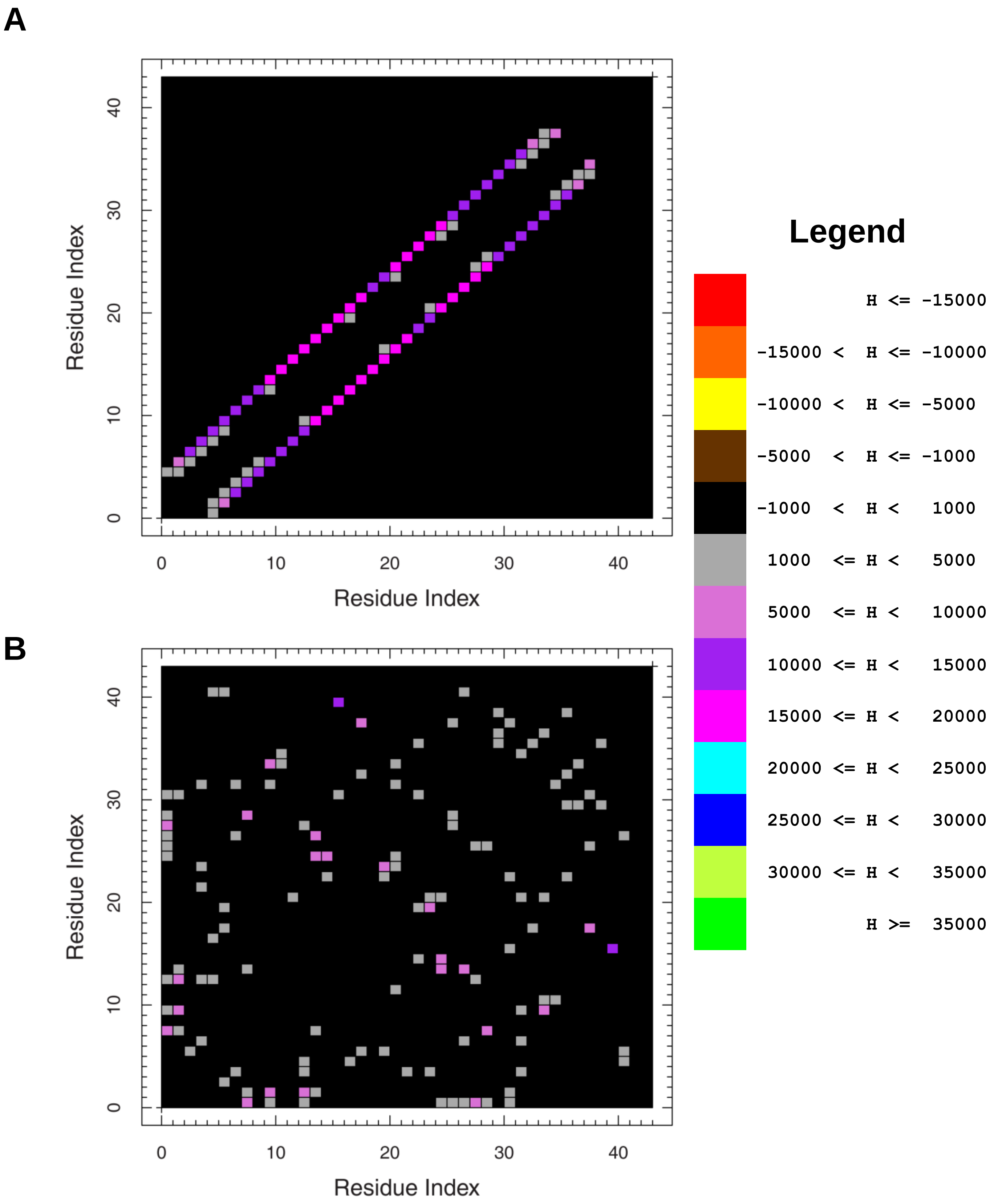


Figure 4.16 Hydrogen bonding of all residues in the ALA_POS_CTRL (A) and the GLY_NEG_CTRL (B) during 500 ns explicit MD simulations. The number of hydrogen bonds observed between two residues during the full 500 ns run is indicated by the colour, as shown in the legend.

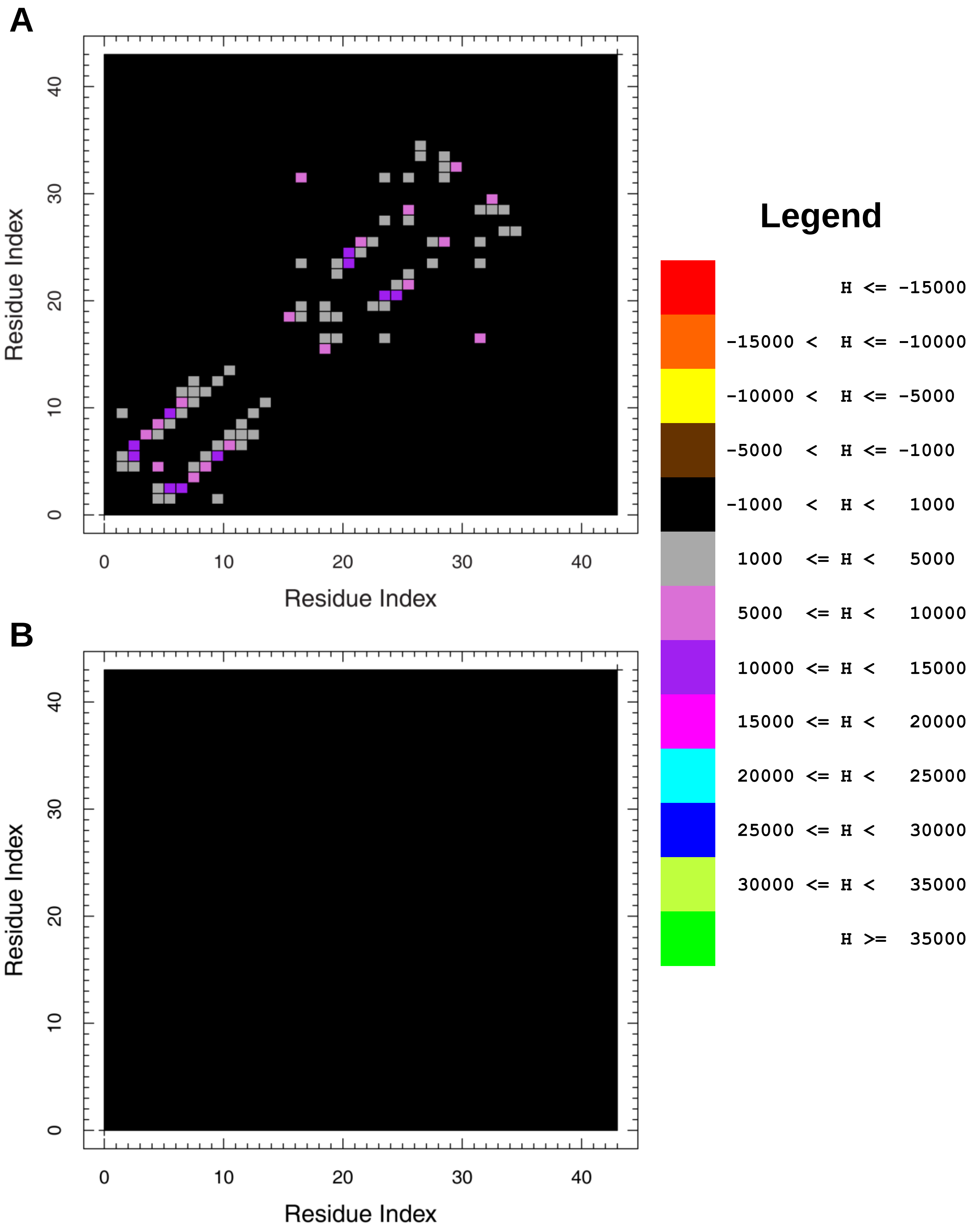


Figure 4.17 Hydrogen bonding of all residues of the WT tail during a 500 ns explicit MD simulation. The raw number of hydrogen bonds represented in A and the number of hydrogen bonds normalized (B) by subtraction of the raw number of hydrogen bonds in the WT tail (4.17.A). The range of the amount of hydrogen bonds (H) correspond to a colour indicated in the legend.

4.3.3.4 ACTIVE tail

We were interested in differences in the location and frequency of hydrogen bonding in the H3 tail peptide with various epigenetic modifications, and thus subtracted the hydrogen bond frequency matrix of the WT tail peptide from that of the matrices discussed below. Figure 4.17.B shows that subtraction of the WT frequency matrix from itself results in a null matrix, as is expected.

The absence of the α – helix at the N – terminal tip of tail in the ACTIVE tail was confirmed by the hydrogen bonding pattern in the brown, yellow and orange hydrogen bond pairs at the tip (See Figure 4.18.B). The active tail contained a hydrogen bond pair between R17 and K14 which occurred between 15 000 and 20 000 times. In the active tail K14 was acetylated and subsequently K14 was also found to form at hydrogen bond pair with K18 on between 10 000 and 15 000 occasions. Surprisingly, these two hydrogen bond pairs were the only ones associated with any of the modified residues. R17 was involved in three hydrogen bond pairs between 5000 and 10 000 occurrences: R17/Q19, R17/T32 and R17/G33, and one between 10 000 and 15 000 occurrences: R17 – A25. Q19 was also involved in quite a number of frequently occurring hydrogen bond pairs. The hydrogen bond pairs Q19 – 17 and Q19 – A23 occurred between 5000 and 10 000 times, and Q19 – R8, Q19 – A15 and Q19 – K23 occurred between 10 000 and 15 000 times. This was an interesting observation, since the α – helix formed in the simulation was between K14 and Q19.The occurrence of the β – bridge between T11 and G34 from 235 ns – 420 ns was also highlighted by the hydrogen bond pair that occurred between 10 000 and 15 000 times. In the WT tail α – helix between L20 and A29, the A21/A24, A21/A25 hydrogen bond pairs were identified as occurring at a high frequency, and thus crucial hydrogen bond pairs involved in the α – helix. However, in the ACTIVE tail three hydrogen bond pairs were found to be present at between 10 000 and 15 000 occurrences, and one hydrogen bond pair, Q19/A24, was present between 5000 and 10 000 occurrences. Thus, this may very well be the reason for the lack of a WT – like α – helix in the ACTIVE tail.

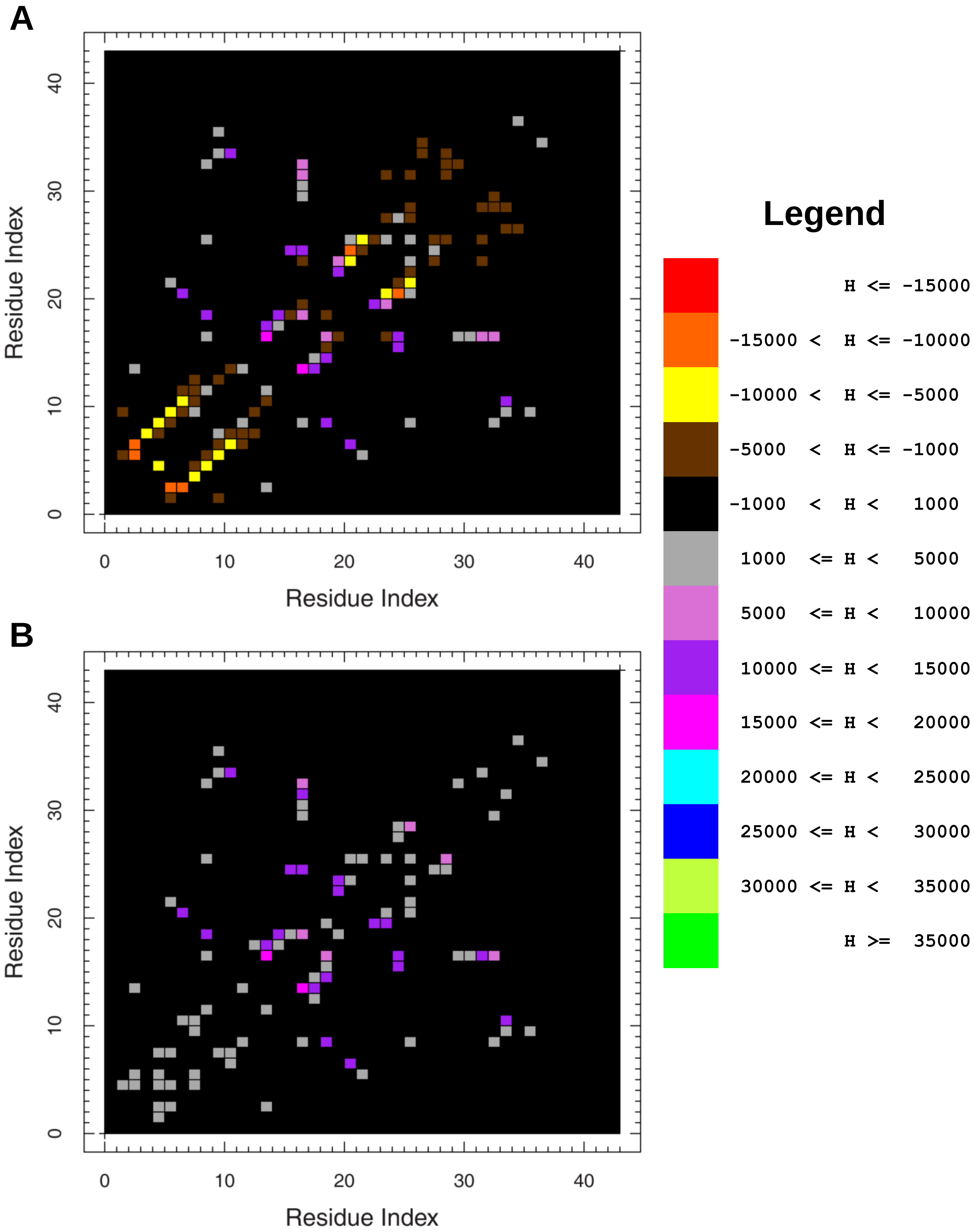


Figure 4.18 Hydrogen bonding of all residues of the ACTIVE tail during a 500 ns explicit MD simulation. The raw number of hydrogen bonds represented in A and the number of hydrogen bonds normalized (B) by subtraction of the raw number of hydrogen bonds in the WT tail (4.17.A). The range of the amount of hydrogen bonds (H) correspond to a colour indicated in the legend.

4.3.3.5 Inactive tail

The secondary structure analysis identified a highly stable α – helix between L20 and K27. A hydrogen bond pair was identified between S28 and R26 which occurred between 30 000 and 35 000 times during the simulation (See Figure 4.19). A hydrogen bond pair between A24 and L20 was also identified, which occurred between 10 000 and 15 000 times. Finally a hydrogen bond pair between K23 and K27 was identified, and occurred between 5000 and 10 000 times. The previous two hydrogen bond pairs had a residue spacing of four residues, suggesting that they were part of an α – helix. Note that S28 was phophorylated and K27 was di – methylated, thus modified residues was directly involved in a different secondary structure not found in either the WT tail or in the ACTIVE tail. Hydrogen bonding in this region which was found in the WT tail, was the hydrogen bond pair, A21/A24, which occurred between 10 000 and 15 000 times and the hydrogen bond pair, R26/A29, which occurred between 5000 and 10 000 times. The phosphorylated S10 was involved in two hydrogen bond pairs, with R8 and R17, which occurred between 15 000 and 20 000 times. Interestingly this was the only hydrogen bond pair of higher occurrence that R17 was a part of. This is in stark contrast to the WT tail and ACTIVE tail where R17 was involved in hydrogen bond pairs with residues more towards the C-terminus of the peptide. The core hydrogen bond pair in the hydrogen bonded turn between P30 and G33 was A29/G33 and occurred between 5000 and 10 000 times. Finally the hydrogen bond pair, Q19/G34, also occurred between 5000 and 10 000 times.

4.3.3.6 HYPER-ALY tail

Secondary structure analysis showed the absence of secondary structure in the HYPER-ALY tail. Likewise the hydrogen bonding pattern (See Figure 4.20) showed the absence of the hydrogen bond pairs associated with the α – helices in the WT tail. The overall pattern was very similar to the GLY_NEG_CTRL, showing some general hydrogen bonding. The highest range of hydrogen bond pair occurrences were in the range of 5000 – 10 000 and all the hydrogen bond pairs included acetylated lysines. These were K9/G12, K14/R17, K4/A31, K27/A29, and both K9/G12

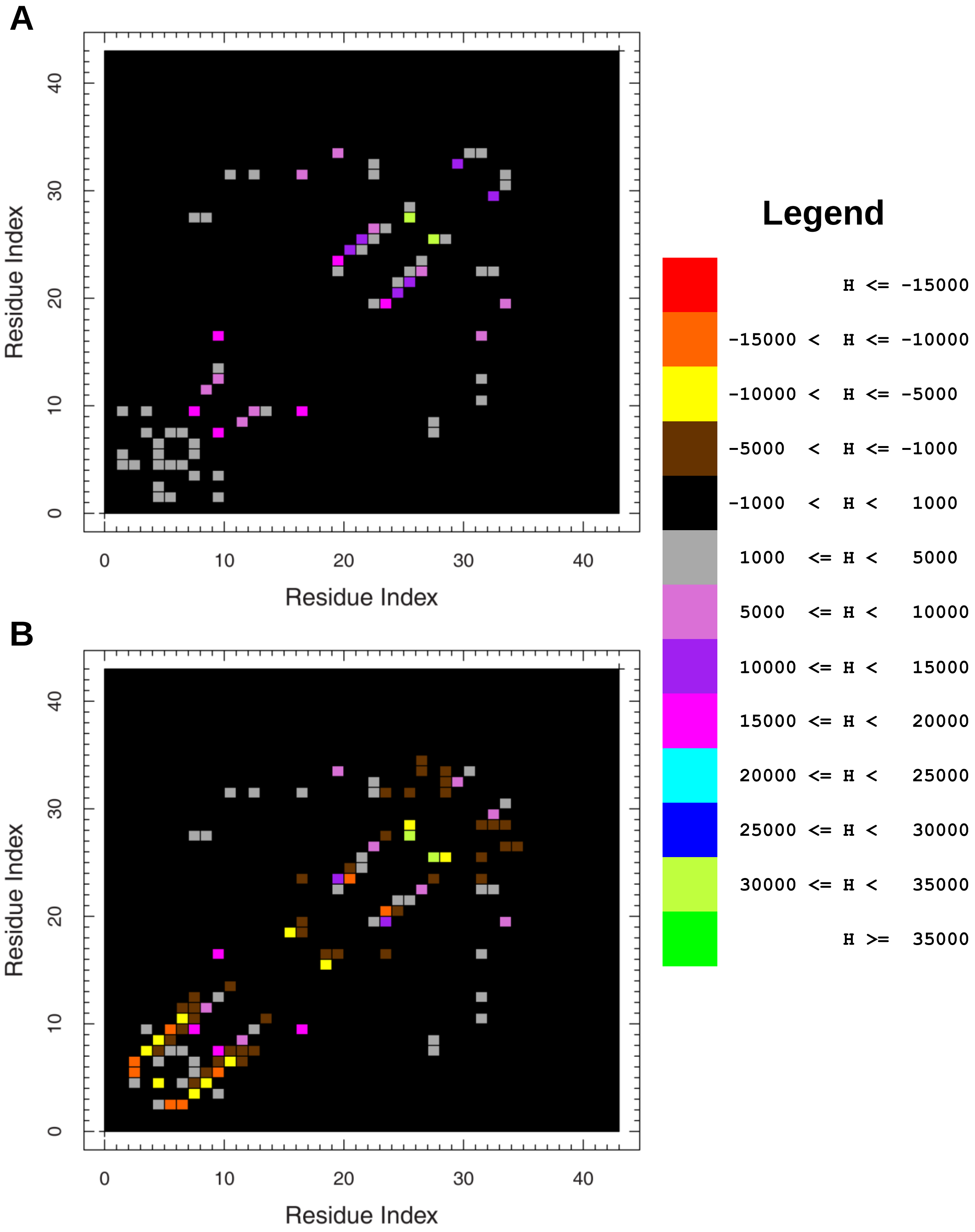


Figure 4.19 Hydrogen bonding of all residues of the INACTIVE tail during a 500 ns explicit MD simulation. The raw number of hydrogen bonds represented in A and the number of hydrogen bonds normalized (B) by subtraction of the raw number of hydrogen bonds in the WT tail (4.17.A). The range of the amount of hydrogen bonds (H) correspond to a colour indicated in the legend.

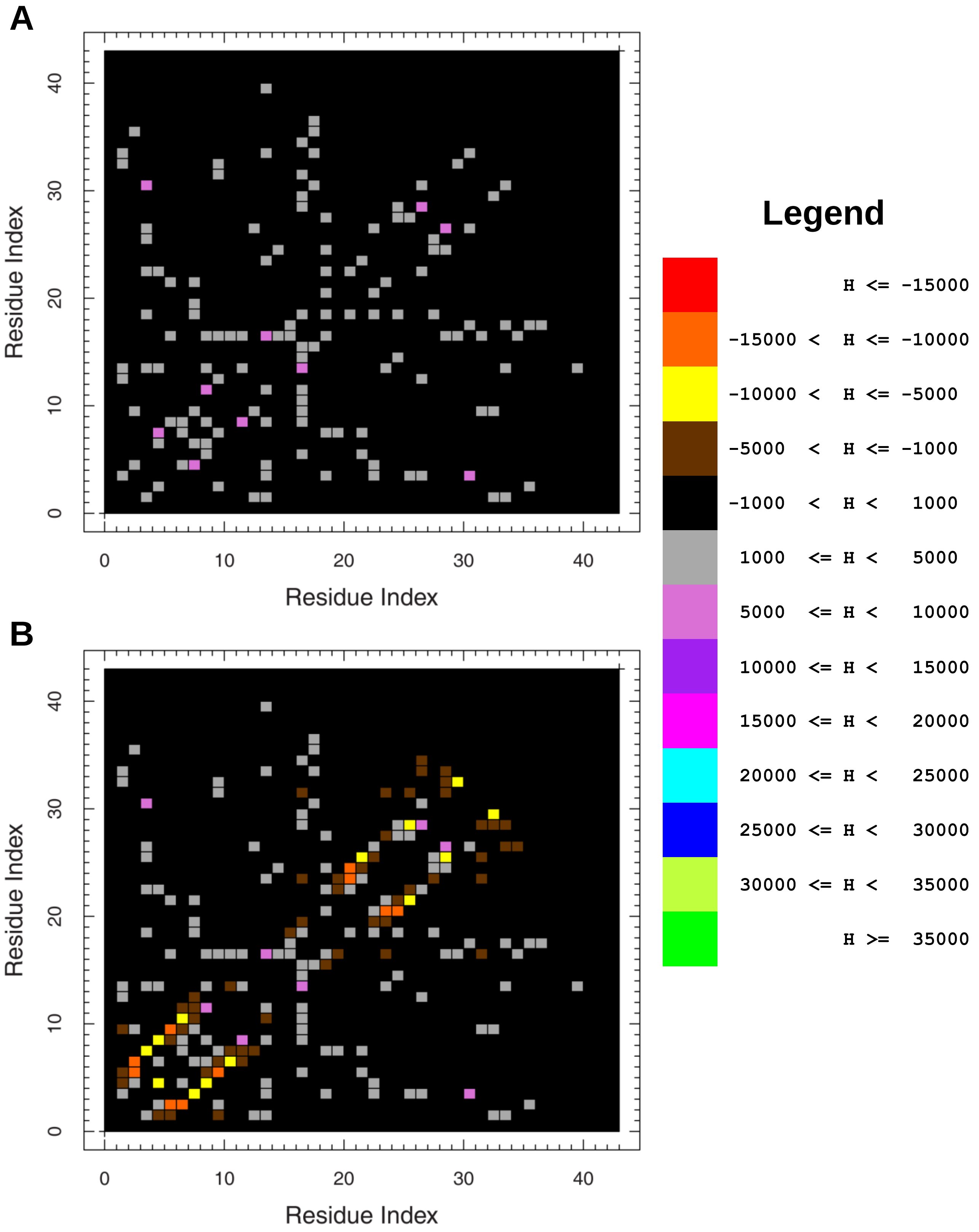


Figure 4.20 Hydrogen bonding of all residues of the HYPER-ALY tail during a 500 ns explicit MD simulation. The raw number of hydrogen bonds represented in A and the number of hydrogen bonds normalized (B) by subtraction of the raw number of hydrogen bonds in the WT tail (4.17.A). The range of the amount of hydrogen bonds (H) correspond to a colour indicated in the legend.

and K14/R17 were also found in the INACTIVE tail, which might indicate that these hydrogen bond pairs occurred regardless of modification state, and thus that the amino acid backbone atoms were involved in these hydrogen bonds.

4.3.3.7 K9ME1_S10PHO tail

The secondary structure analysis of the K9ME1_S10PHO tail showed the formation of both an α – helix, T22 – A29, and a β – bridge hairpin, A7/R8 – G12/G13. The hydrogen bonding pattern (See Figure 4.21) showed the formation of a hydrogen bond pair between S10 and R8 which occurred between 25 000 and 30 000 times. This hydrogen bond pair formed part of the hydrogen bonded turn of the β – bridge hairpin. Additionally, the hydrogen bond pair between R8 and G13 occurred between 10 000 and 15 000 times, while the hydrogen bond pairs A7/G13 and R8/G12 occurred between 5000 and 10 000 times. The hydrogen bond pairs T22/A25, T22/R26 and A24/S28 were correctly spaced to be in an α – helix and occurred between 5000 and 10 000 times. Two additional hydrogen bond pairs which were not directly involved in the secondary structure elements were also identified: First S10/R17 and secondly A21/R26, both of which occurred between 15 000 and 20 000 times.

4.3.3.8 K9ME2_S10PHO tail

Secondary structure analysis identified two α – helices: T3 – R8 and L20 – A25, and three prominent hydrogen bonded turns: K9 – G12, R26 – P30 and A31 – G34. This combination of secondary structure elements overlapped with the WT tail and subsequently, because the WT tail had more stable α – helices and hydrogen bonded turns, the hydrogen bond pairs associated with the secondary structures were not visible, because of a large number of occurrences in the hydrogen bonding matrix (See Figure 4.22). However, a number of hydrogen bond pairs were identified that did not belong to any of the secondary structure elements. The most frequently occurring hydrogen bond pair was R8/K9 and occurred between 15 000 and 20 000 times. This was followed by S10/V35 (carboxyl oxygen of S10 hydrogen bonded to the backbone amide of V35) which occurred between 10 000 and 15 000 times. Finally, K9/T11, which occurred between

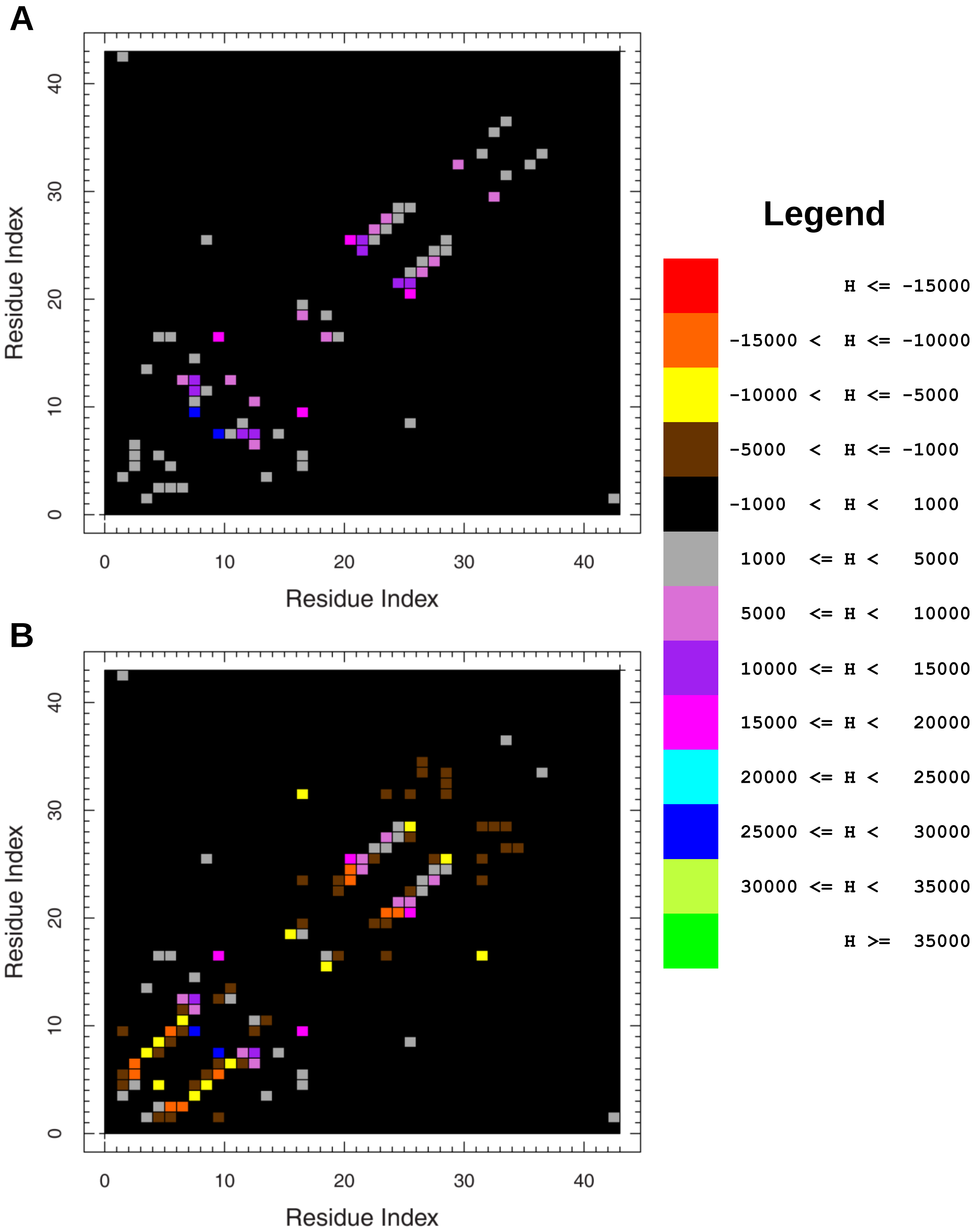


Figure 4.21 Hydrogen bonding of all residues of the K9ME1_S10PHO tail during a 500 ns explicit MD simulation. The raw number of hydrogen bonds represented in A and the number of hydrogen bonds normalized (B) by subtraction of the raw number of hydrogen bonds in the WT tail (4.17.A). The range of the amount of hydrogen bonds (H) correspond to a colour indicated in the legend.

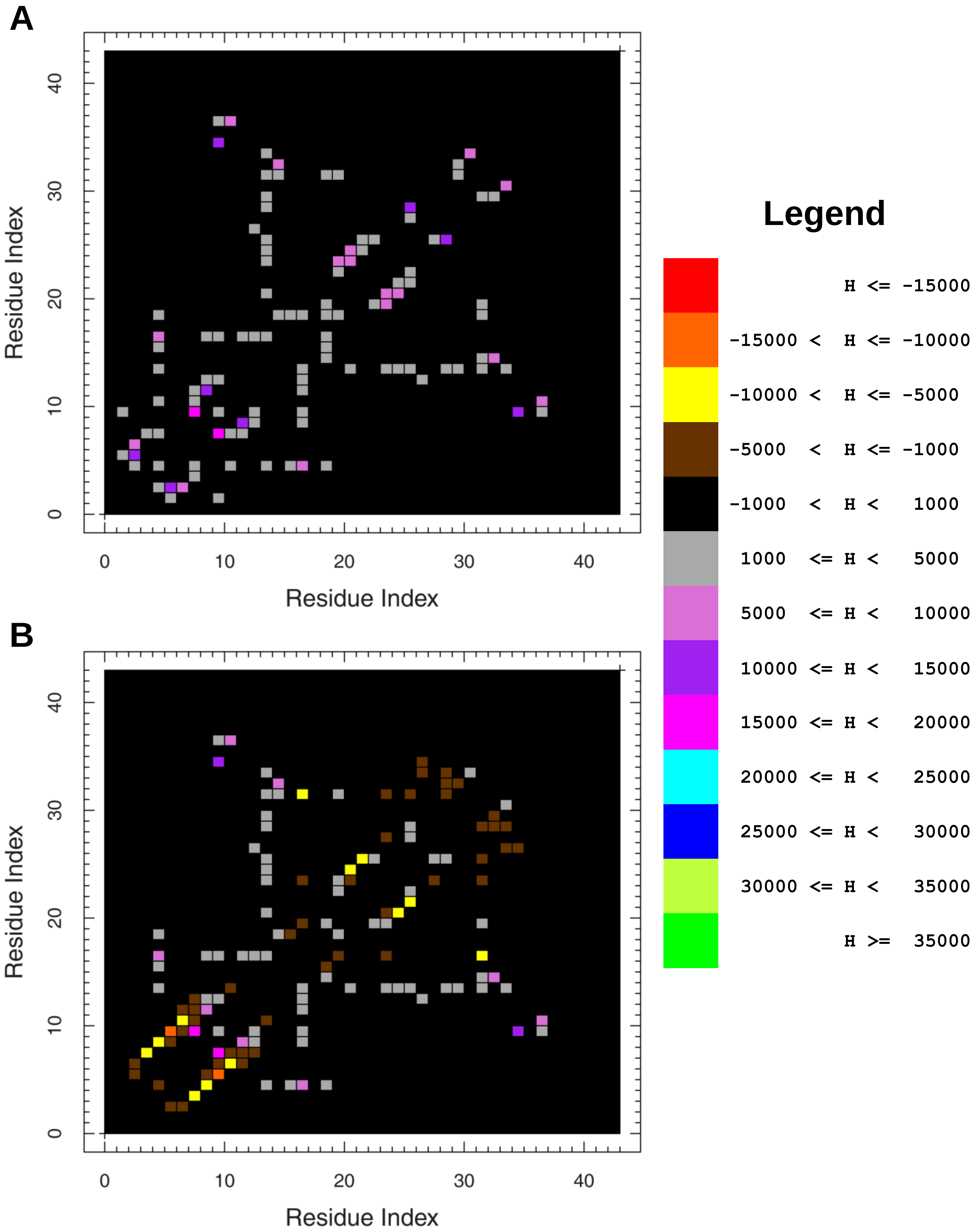


Figure 4.22 Hydrogen bonding of all residues of the K9ME2_S10PHO tail during a 500 ns explicit MD simulation. The raw number of hydrogen bonds represented in A and the number of hydrogen bonds normalized (B) by subtraction of the raw number of hydrogen bonds in the WT tail (4.17.A). The range of the amount of hydrogen bonds (H) correspond to a colour indicated in the legend.

5000 and 10 000 times. The rest of the hydrogen bond pairs occurred between 5000 and 10 000 times, and included Q5/R17, G33/A15 and T11/K37.

4.3.3.9 K9ME3_S10PHO tail

The secondary structure analysis showed that the K9ME3_S10PHO tail never really had any particular secondary structure elements for very long periods of time. The hydrogen bonding analysis underscored this (See Figure 4.23), with only one hydrogen bond pair occurring more than 5000 times. It was found between the phosphorylated S10 and R8. However, some of the secondary structure elements that could be identified, for example the β – bridge occurring between R1 and R26 from ~153 ns - ~205 ns, could be observed as a hydrogen bond pair occurring between 1000 and 5000 times. Two hydrogen bond pairs involved in the α – helix between A1 and A7 could also be identified. These were the hydrogen bond pairs: A1/Q5 and R2/R6, and both occurred between 1000 and 5000 times.

4.3.3.10 K9ME1 tail

The secondary structure analysis of the K9ME1 tail trajectory yielded four α – helices: A1 – R8, A1 – G12, P16 – A21 and T22 – A29. All except one hydrogen bond pair fell within these regions (see Figure 4.24). This hydrogen bond pair, Q19/T22, occurred between 10 000 and 15 000 times. The first two helices overlapped, with the second α – helix being an extended form of the first. Thus they shared hydrogen bond pairs and these were R2/T6 and T3/A7, which occurred between 10 000 and 15 000 times. However, T3/A7 was not observed in the normalized matrix, indicating that it was also observed in the WT tail with a similar frequency. Additionally, the hydrogen bond pair R2/Q5 was observed between 5000 and 10 000 times. The hydrogen bond pair, T6/S10 was only observed in the raw matrix and also occurred between 5000 and 10 000 times. The α – helix between P16 and A21, which was well represented by the hydrogen bond pairs: P16/Q19, P16/L20 and R17/A21, occurred between 5000 and 10 000 times. Interestingly P16/Q19 occurred in both the WT tail and K9ME1 tail, however it did so to a greater extend in the K9ME1 tail. The

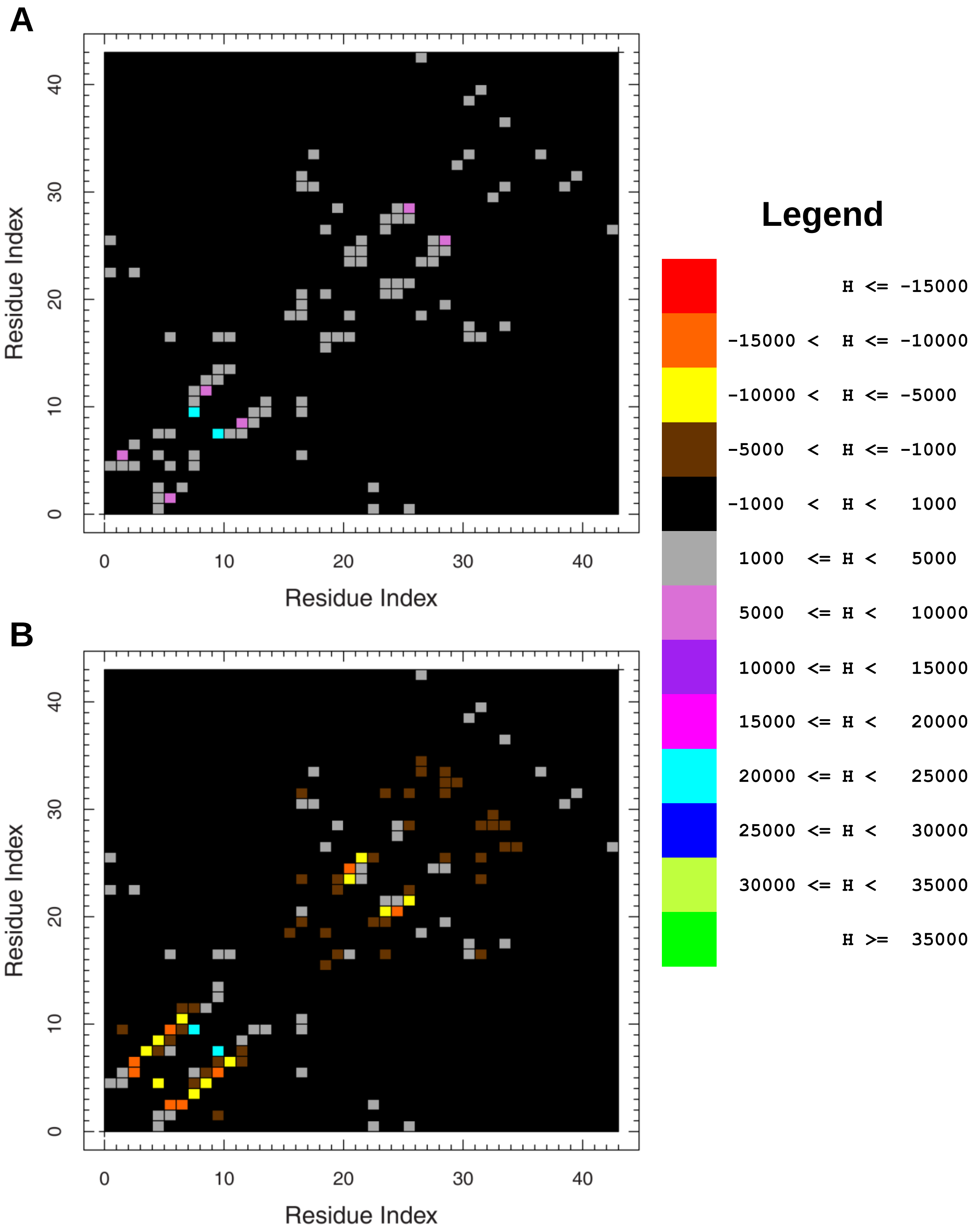


Figure 4.23 Hydrogen bonding of all residues of the K9ME3_S10PHO tail during a 500 ns explicit MD simulation. The raw number of hydrogen bonds represented in A and the number of hydrogen bonds normalized (B) by subtraction of the raw number of hydrogen bonds in the WT tail (4.17.A). The range of the amount of hydrogen bonds (H) correspond to a colour indicated in the legend.

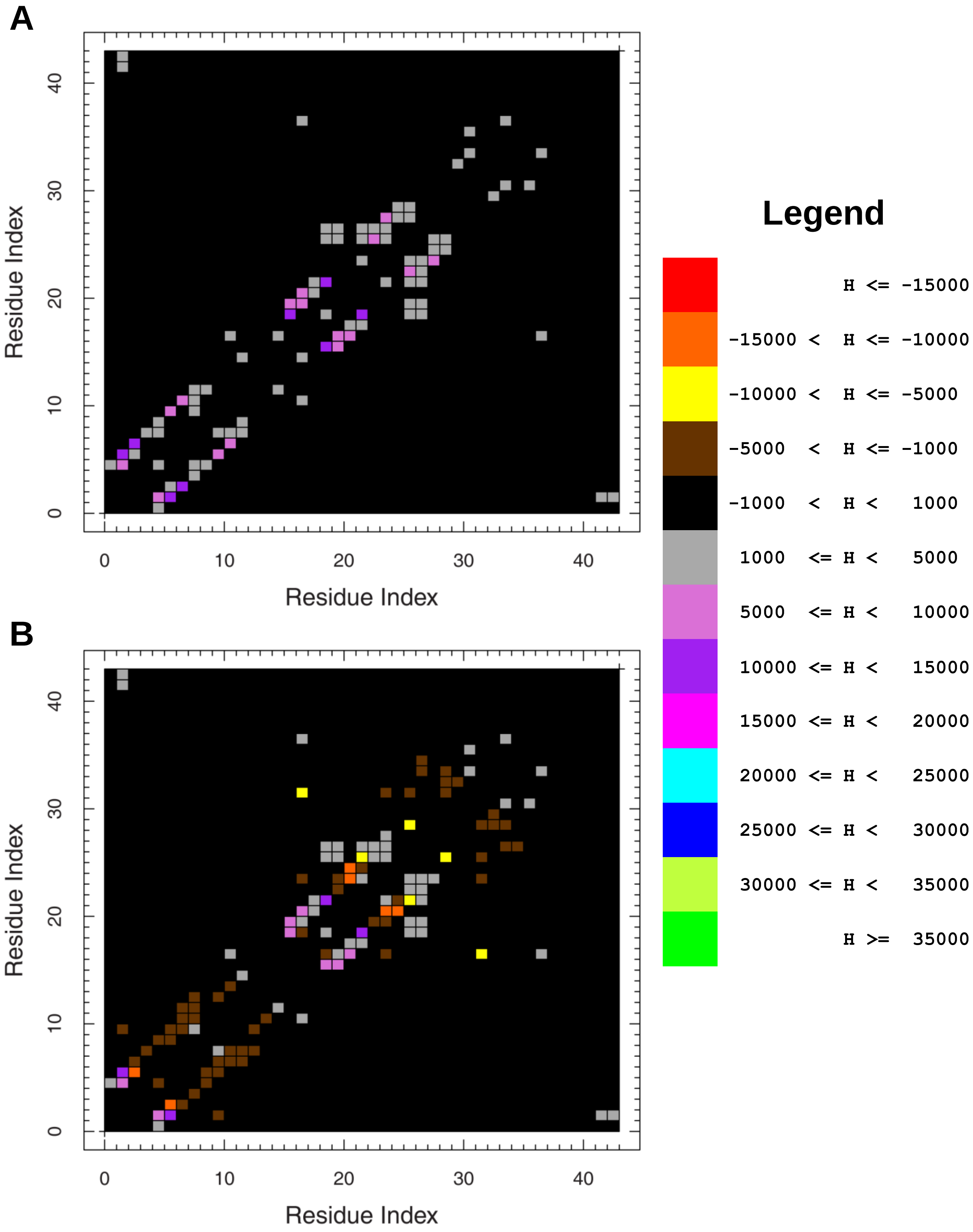


Figure 4.24 Hydrogen bonding of all residues of the K9ME1 tail during a 500 ns explicit MD simulation. The raw number of hydrogen bonds represented in A and the number of hydrogen bonds normalized (B) by subtraction of the raw number of hydrogen bonds in the WT tail (4.17.A). The range of the amount of hydrogen bonds (H) correspond to a colour indicated in the legend.

number of occurrences were none the less decreased from between 10 000 and 15 000 in the raw matrix to between 5000 and 10 000 times in the normalized matrix. The α – helix between T22 and A29 was represented in the raw matrix, where the hydrogen bond pairs: K23/R26 and A24/S28, were present between 5000 and 10 000 times. Thus, these hydrogen bond pairs were mostly shared with the WT tail.

4.3.3.11 K9ME2 tail

Secondary structure analysis identified two α – helices: A1 – G12 and L20 – A29, and one region of hydrogen bonded turns: P30 – G33. Again, when comparing the raw hydrogen bonding matrix (see Figure 4.25.A) with the normalized matrix (see Figure 4.25.B), it was evident that the K9ME2 tail shared a significant amount of hydrogen bond pairs with the WT tail. Thus, the most frequently occurring hydrogen bond pair, R2/T6, occurred between 5000 and 15 000 times. This pair was in the region of the first α – helix. From the raw matrix the T3/A7 bond pair occurred between 10 000 and 15 000 times, and the hydrogen bond pairs: R2/Q5, R2/T6, T3/T6, K4/R8, Q5/R8, Q5/K9, T6/S10 and A7/T11, occurred between 5000 and 10 000 times, also within the region of the first α – helix. Two hydrogen bond pairs, A21/A25 and T22/R26 fell within the region of the α – helix between L20 and A29, and occurred between 10 000 and 15 000 times. Three hydrogen bond pairs: A24/S28, R26/A29 and A29/G33, occurred between 5000 and 10 000 times and was also within the region of the L20 – A29 α – helix. The hydrogen bonded turn region was not identified by the normalized matrix, and the raw matrix only showed the hydrogen bond pair, A31/G34, which occurred between 1000 and 5000 times within the region of the turn.

4.3.3.12 K9ME3 tail

The secondary structure analysis of the K9ME3 tail revealed the formation of β – bridges between K14/A15/R17 and A25/R26/A29/P30, an α – helix between P16 and A21 and regions of hydrogen bonded turns: R2 – T6, K9 – G12 and P30 – G34. The hydrogen bond pairs, R17/R26 and A15/A29, were shown to be involved in β – bridges and occurred between 10 000 and 15 000

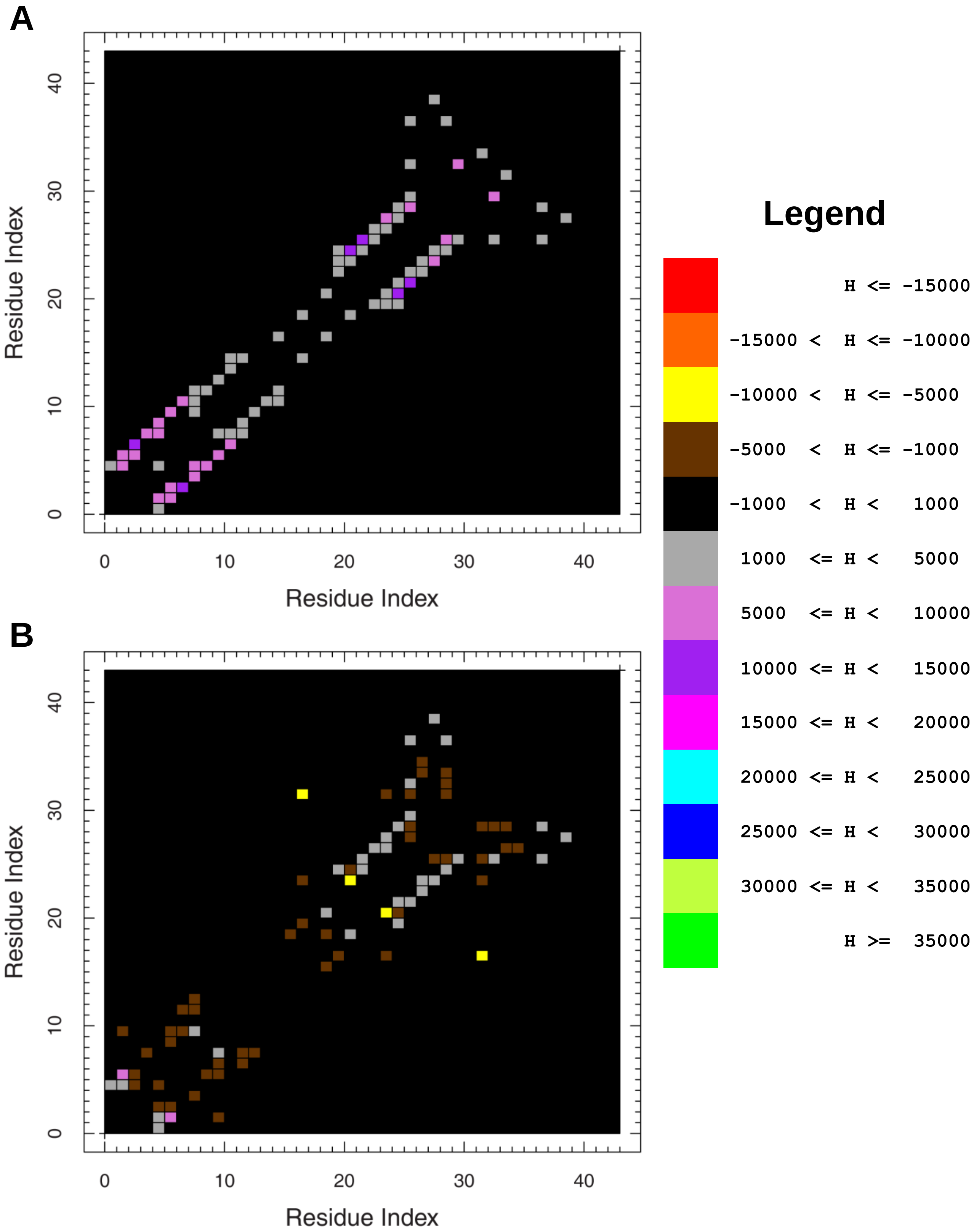


Figure 4.25 Hydrogen bonding of all residues of the K9ME2 tail during a 500 ns explicit MD simulation. The raw number of hydrogen bonds represented in A and the number of hydrogen bonds normalized (B) by subtraction of the raw number of hydrogen bonds in the WT tail (4.17.A). The range of the amount of hydrogen bonds (H) correspond to a colour indicated in the legend.

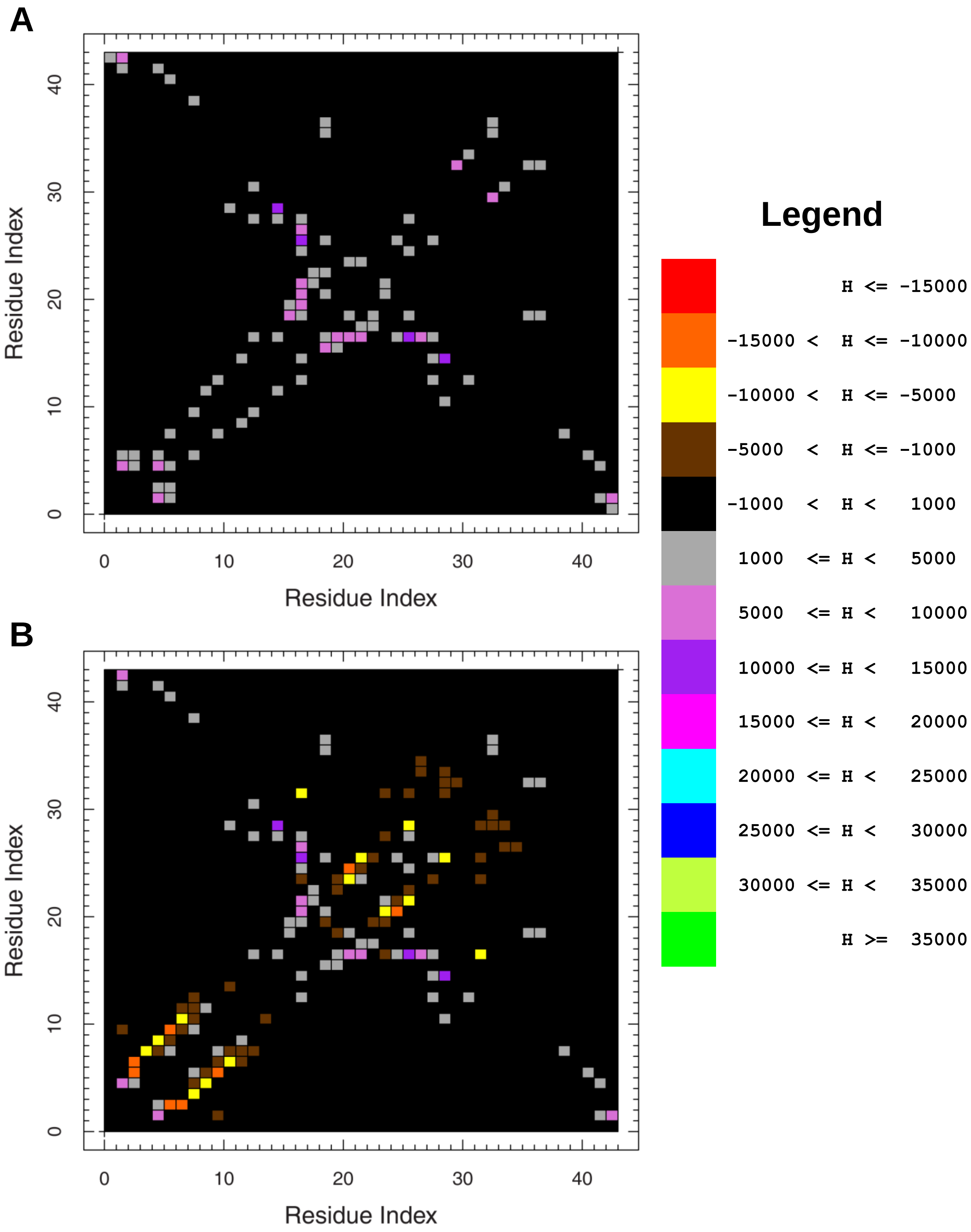


Figure 4.26 Hydrogen bonding of all residues of the K9ME3 tail during a 500 ns explicit MD simulation. The raw number of hydrogen bonds represented in A and the number of hydrogen bonds normalized (B) by subtraction of the raw number of hydrogen bonds in the WT tail (4.17.A). The range of the amount of hydrogen bonds (H) correspond to a colour indicated in the legend.

times (see Figure 4.26). The hydrogen bond pair, R17/A2, was within the region of the α – helix and occurred between 5000 and 10 000 times. The hydrogen bond pairs, R2/Q5, R17/T22 and R17/K27, were in the regions of hydrogen bonded turns and occurred between 5000 and 10 000 times. Interestingly R17 was involved in all the bonds occurring between 5000 and 15 000 times, with the exception of three bond pairs. Finally, the hydrogen bond pair, R2/P43, occurred between 5000 and 10 000 times, indicating that the N-terminal tip was interacting with the C – terminal tip.

4.3.3.13 K9ACE_S10PHO tail

Secondary structure analysis revealed that the K9ACE_S10PHO tail was rich in hydrogen bonded turns and also contained two α – helices, T3 – R8 and L20 – A25, in relatively low abundance. The normalized matrix (see Figure 4.27.B) did not show any hydrogen bonds involved in the α- helical regions. Three hydrogen bonds involving the phosphorylated S10: R8, T22, R26, was indentified and was shown to occur between 5000 and 10 000 times. Three hydrogen bond pairs involving Q19, A21, K23 and A24, was also shown to occur between 5000 and 10 000 times. Lastly a hydrogen bond pair, K14/R17 was shown to be present between 5000 and 10 000 times. Using the raw matrix (see Figure 4.27.A), hydrogen bond pairs in the region of the α – helices were found. Only one hydrogen bond pair, T3/T6 was identified for the α – helix between T3 and R8 and occurred between 10 000 and 15 000 times. For the α – helix between L20 and A25 three hydrogen bond pairs, A21/A24, K23/A25 and A24/A25, were found to occur between 5000 and 10 000 times. However the spacing of the hydrogen bond pairs suggests that they were more likely within the hydrogen bonded turn identified in the same region.

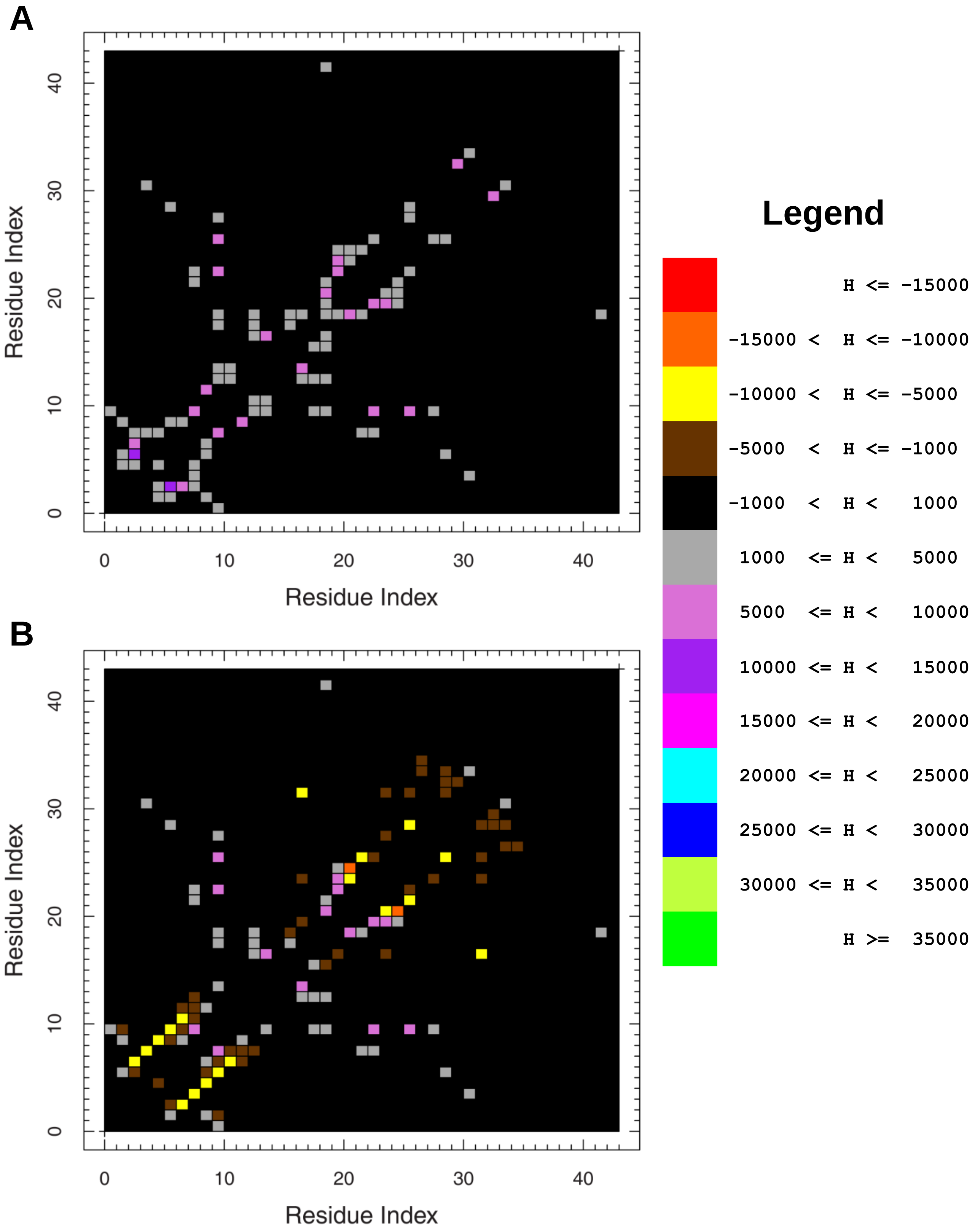


Figure 4.27 Hydrogen bonding of all residues of the K9ACE_S10PHO tail during a 500 ns explicit MD simulation. The raw number of hydrogen bonds represented in A and the number of hydrogen bonds normalized (B) by subtraction of the raw number of hydrogen bonds in the WT tail (4.17.A). The range of the amount of hydrogen bonds (H) correspond to a colour indicated in the legend.

4.3.4 Clustering of 500 ns explicit MD trajectories

As this point we have looked at the presence of various secondary structures in the H3 N-terminal peptide over a 500 ns MD simulation run, and identified clear differences between the peptides that carry different epigenetic modification patterns. We have also analyzed the frequency of appearance of hydrogen bonds between the various residues over the simulation period, and provided some understanding to the molecular mechanisms involved in the stabilization of some of the secondary structures. We have not yet considered any salt bridges or hydrophobic interactions. To this end, we studied the tertiary structures of the most prominent conformations present in the 500 ns MD run to gain insight into both the preferred shape of peptide in space, as well as allowing an analysis of the forces that stabilize such conformations.

4.3.4.1 WT tail

The WT tail showed three top clusters (see Figure 4.28.A.i - iii). The N-terminal α – helix was shown to be the more stable of the two α – helices observed, and this observation was confirmed by the clustered structures. In all the clusters the N-terminal α – helix was present, whereas the central α – helix was only observed in the third top cluster. In terms of time progression, the orders of the clusters were: iii, i, ii. Viewing the clusters in that order revealed that the central α – helix gradually unfolded during the course of the simulation.

4.3.4.2 ACTIVE tail

The ACTIVE tail showed only one top cluster (see Figure 4.28.B.i) and was represented by the structure of the tail at 236.325 ns in the trajectory. The structure revealed a bulge formed from the region K36 - A7 and no secondary structure elements were found in the N-terminal tip of the peptide. This bulge was maintained by three main hydrogen turns in the following regions: K14 – R17, L20 – R23 and P30 – G33, as observed throughout the majority of the trajectory (Figure 4.3.B). Interestingly, the side-chains of all the modified lysine residues pointed outwards, away from this structure.

4.3.4.3 INACTIVE tail

Two major clusters were found in the INACTIVE tail trajectory (see Figure 4.28.C.i - ii). These clusters were, however, very similar, and both were represented by time points near the end of the trajectory. Both contained the very stable α – helix between L20 and K27 and the hydrogen bonded turn between P30 and G33. The structures only differed by the 3-10 helix stabilized in the most predominant structure C.i, and a hydrogen bonded turn in the C.ii structure, between K9 and G13. This could be explained by the observation that the di – methylated side chain of K9 was pointing towards the α – helix in C.i as opposed to in C.ii, where the side chain was pointing in an opposite direction towards the N-terminal end of the peptide.

4.3.4.4 HYPER – ALY tail

The three top clusters were found from the trajectory of the HYPER – ALY tail (see Figure 4.28.D.i - iii). The representative structure of the clusters showed minimal secondary structure elements, yet showed a similar bulge motif as found in the ACTIVE tail. However these bulges seemed more comprehensive, and when looking at the representative structures in temporal order in the trajectory, it was observed that the bulges compressed the peptide increasingly with time. This led to the most compressed structure in the top cluster which contained a β – bridge between A25/R26 and A29 as well as a hydrogen bonded turn, K9 – G12.

4.3.4.5 K9ME1_S10PHO tail

Only one cluster was found in the trajectory of the K9ME1_S10PHO tail (see Figure 4.28.E.i). The structure was sampled near the end of the simulation, and showed an α – helix between T22 and A29. Interestingly the α – helix sat on the outside of a loop formed in the peptide with a hydrogen bonded turn, between R8 and T11, positioned towards the inside the loop, opposite the α – helix.

4.3.4.6 K9ME2_S10PHO tail

The K9ME2_S10PHO tail trajectory produced three clusters (see Figure 4.28.F.i - iii). The representative structures of the top cluster and the third cluster were very similar, with both showing the formation of a loop similar to the one found in K9ME1_S10P tail. However the α – helix on the outside of the loop in K9ME1_S10P tail was replaced by three hydrogen bonded turns and no hydrogen bonded turn that faced towards the inside of the loop. The only difference between the two structures was the presence of an N-terminal α – helix between T3 and R8 in the top ranked cluster structure. The second structure showed a bulge – like motif in the same region similar to the ACTIVE tail and also contained no folding in the N-terminal tip of the peptide.

4.3.4.7 K9ME3_S10PHO tail

Clustering of the K9ME3_S10PHO trajectory also yielded three clusters (see Figure 4.28.G.i - iii). The representative structures showed the formation of the same loop observed for both the K9ME1_S10PHO and K9ME2_S10PHO tails, yet the loop in this instance was narrower and resembled a hairpin – motif rather than a loop. The structures of the top and second ranked clusters were virtually identical, except for a hydrogen bonded turn between T22 and A25 in the second ranked structure. In the third ranked cluster the structure exhibited the formation of a β – bridge between A1 and R26 which served as the stem for a loop formed with the modified residues in a 3-10 – helix on the outside of the loop, resembling the α – helix found in the K9ME1_S10PHO tail’s top ranked cluster.

4.3.4.8 K9ME1 tail

Three major clusters were found in the K9ME1 tail trajectory (see Figure 4.28.H.i - iii). The structure for the top ranked cluster showed an α – helix between P16 and A21. The α – helix and a hydrogen bonded turn in the region G34 – K37 formed the stem of a small loop, while the N-terminal tip of the peptide looped over and formed a closed loop with the C-terminal end of the peptide. The structure from the second ranked cluster also contained the α – helix between P16 and A21, however it lacked the hydrogen bonded turn between G34 and K37. It did, however, contain a hydrogen bonded turn in the region A25 – K27, which allowed the peptide to adopt a more hairpin – like structure. In addition, another α – helix was formed between R2 and A7. The third ranked cluster revealed a loop – like structure with three α – helices. The α – helix found in the other two cluster between P16 and A21 was found on the outside of the loop flanked by two α – helices which together formed the stem of the loop. On the N-terminal side an α – helix was seen between T3 and T11 and on the C – terminal side was the α – helix was observed between A24 and A29.

4.3.4.9 K9ME2 tail

The K9ME2 tail trajectory produced two major clusters (see Figure 4.28.I.i - ii). The top ranked cluster showed an N-terminal tip α – helix between R2 and R8. This α – helix was connected to a bulge – like structure in the region of T11 – P30 which contained two hydrogen bonded turns: T11 – K14 and R26 – A29. The structure from the second cluster showed a series of hydrogen bonded turns where the α – helix was in the top cluster, as well as a small loop between A21 and G33 instead of the bulge – like structure, connected by almost a straight chain of residues.

4.3.4.10 K9ME3 tail

Three major clusters were found in the K9ME3 tail trajectory (see Figure 4.28.J.i - iii). The representative structure of the top cluster formed a long hairpin – like loop similar to the K9ME1 top cluster structure, yet without any α – helical content. There was, however, a β – bridge formed between R17 and A25/R26 and R2 made contact with the C-terminal tip. The second most populated cluster's structure showed a similar conformation; however no β – bridge was present because R17 was part of a 3-10 helix in the region P16 – Q19. The N-terminal tip of the tail also did not interact with the C-terminal end of the peptide. The third most populated cluster showed a structure with a well defined bulge – like conformation. Firstly, a loop was formed between A15 and A29, with a β – bridge formed between these residues, forming the stem of the loop. Within this loop a hydrogen bonded turn was formed between P16 and Q19 which a hydrogen bond between K36 and K37. This basically folded the tip of the loop over, forming the bulge.

4.3.4.11 K9ACE_S10PHO tail

The trajectory of the K9ACE_S10PHO tail produced two clusters (see Figure 4.28.K.i - ii). The most populated cluster showed a bulge/loop structure flanked by two hydrogen bonded turns in G5

– R8 and L20 – K23. The structure from the second most populated cluster showed two bulges: the first between R8 and L20, as seen in the most populated structure, and the second between K23 and G34. Overall the structure was rich in hydrogen bonded turns with four occurring within the structure at: T3 – T6, K9 – G12, L20 – K23 and P30 – G33.

4.3.4.12 GLY_NEG_CTRL tail

The GLY_NEG_CTRL tail produced two clusters (see Figure 4.28.L.i - ii). Both clusters were predictably folded into itself, in a bulge, with both being rich in hydrogren bonded turns and β – bridges.

4.3.4.13 ALA_POS_CTRL tail

The ALA_POS_CTRL tail delivered only one major cluster (see Figure 4.28.M.i). The structure of this top ranked cluster was that of one long straight α – helix with only the very tip unfolded, as is expected from terminal instability.

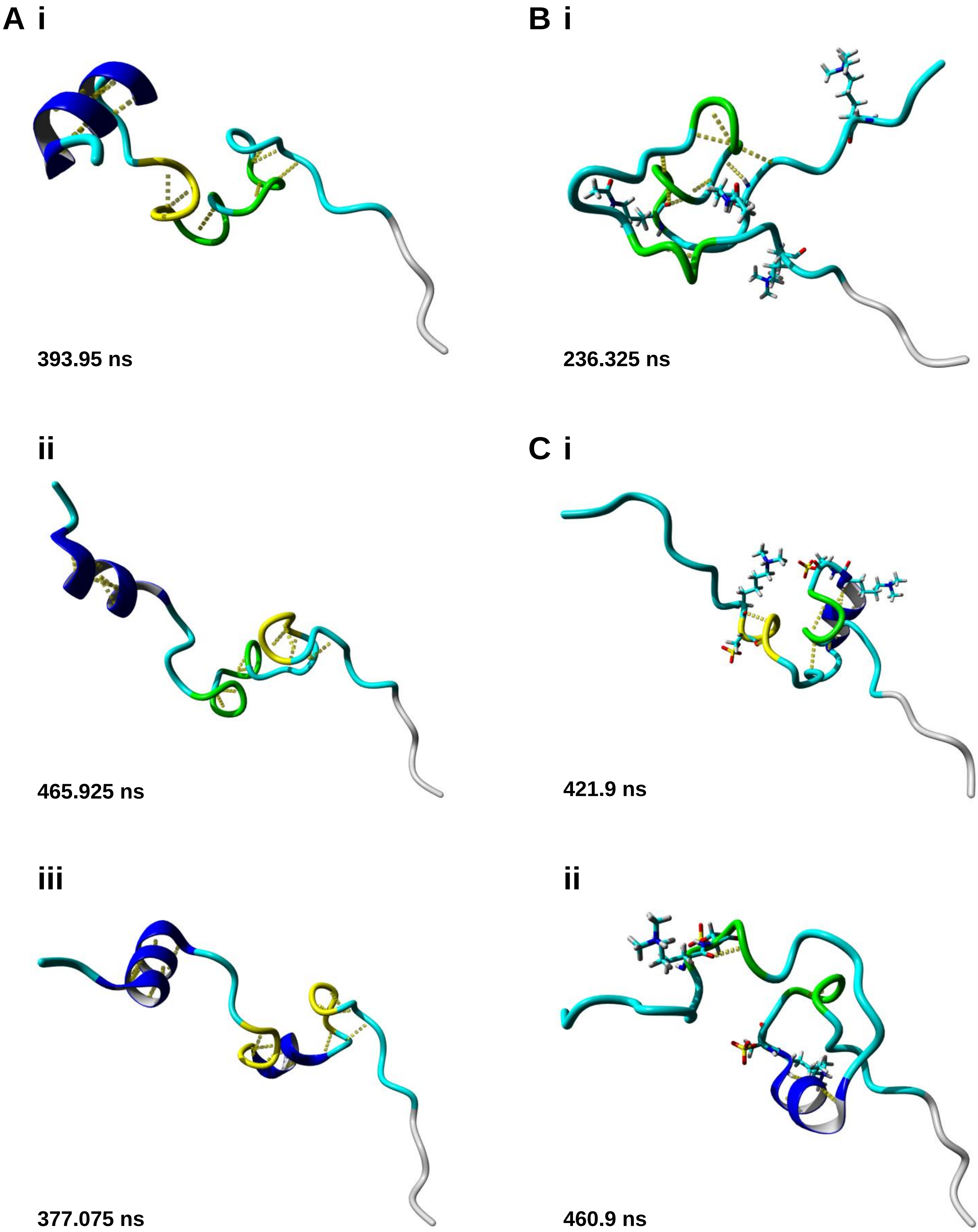


Figure 4.28 Representative structures of the most populated clusters for the WT (A), ACTIVE (B) and INACTIVE (C) tail obtained from 500 ns - explicit MD simulations. Numbers next to the structure indicate the cluster rank. The time point from which the structure was sampled is shown at the bottom of each structure. Molecular graphics created with YASARA (www.yasara.org) and POVRay (www.povray.org)

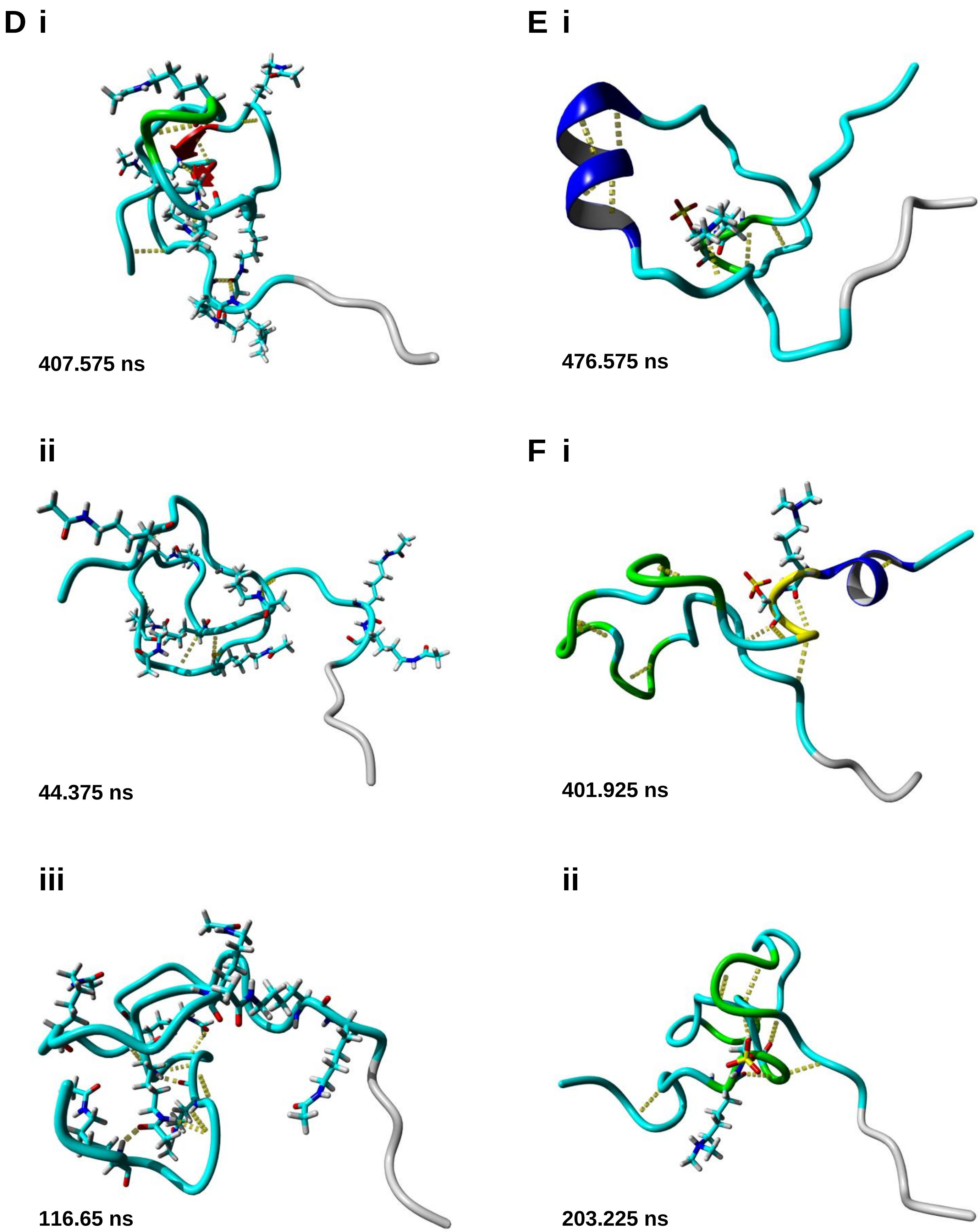


Figure 4.28 (cont.) Representative structures of the most populated clusters for the HYPER-ALY (D), K9ME1_S10PHO (E) and K9ME2_S10PHO (F) tail obtained from 500 ns - explicit MD simulations.

Molecular graphics created with YASARA (www.yasara.org) and POVRay (www.povray.org)

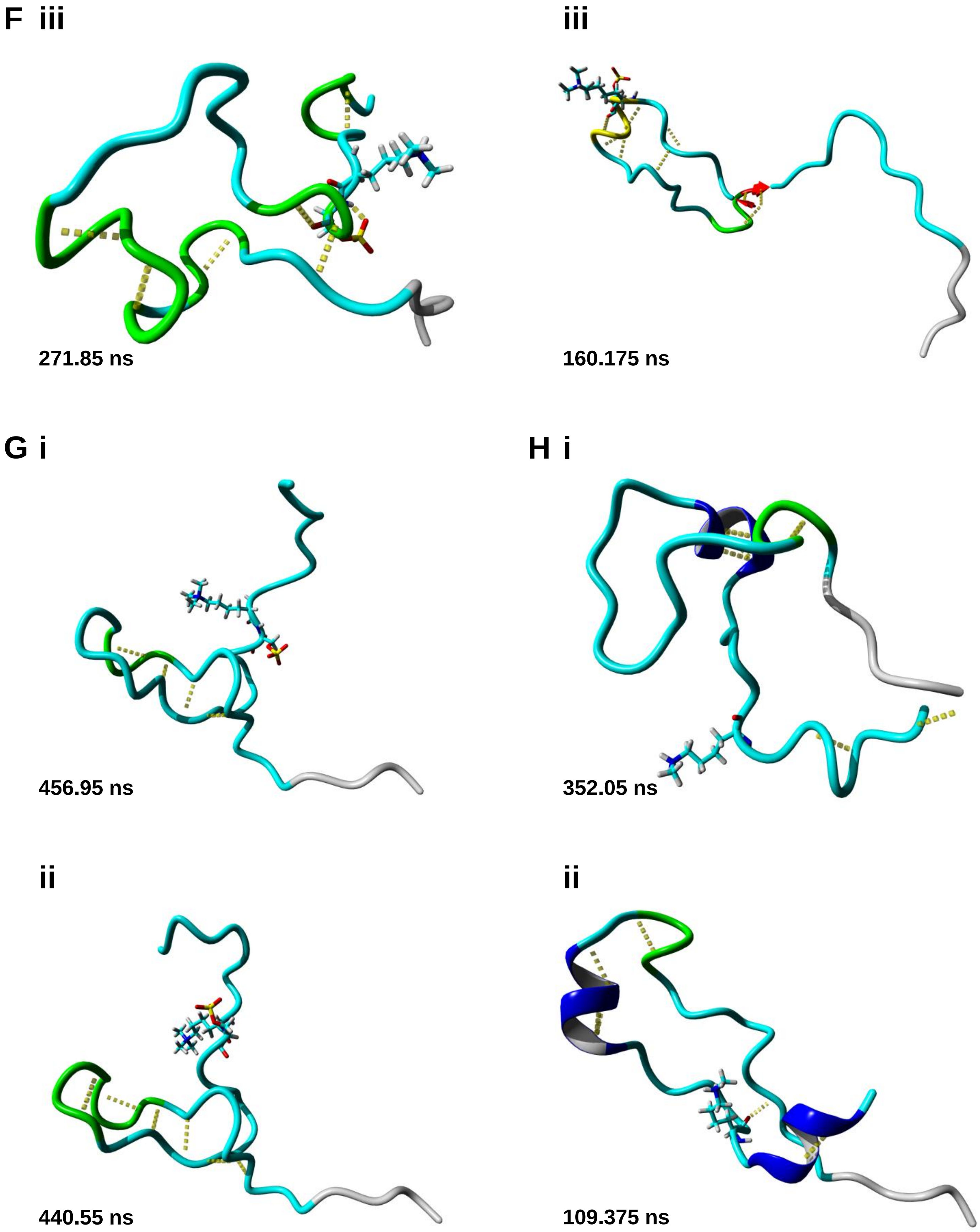


Figure 4.28 (cont.) Representative structures of the most populated clusters for the K9ME2_S10PHO (F), K9ME3_S10PHO (G) and K9ME1 (H) tail obtained from 500 ns - explicit MD simulations.

Molecular graphics created with YASARA (www.yasara.org) and POVRay (www.povray.org)

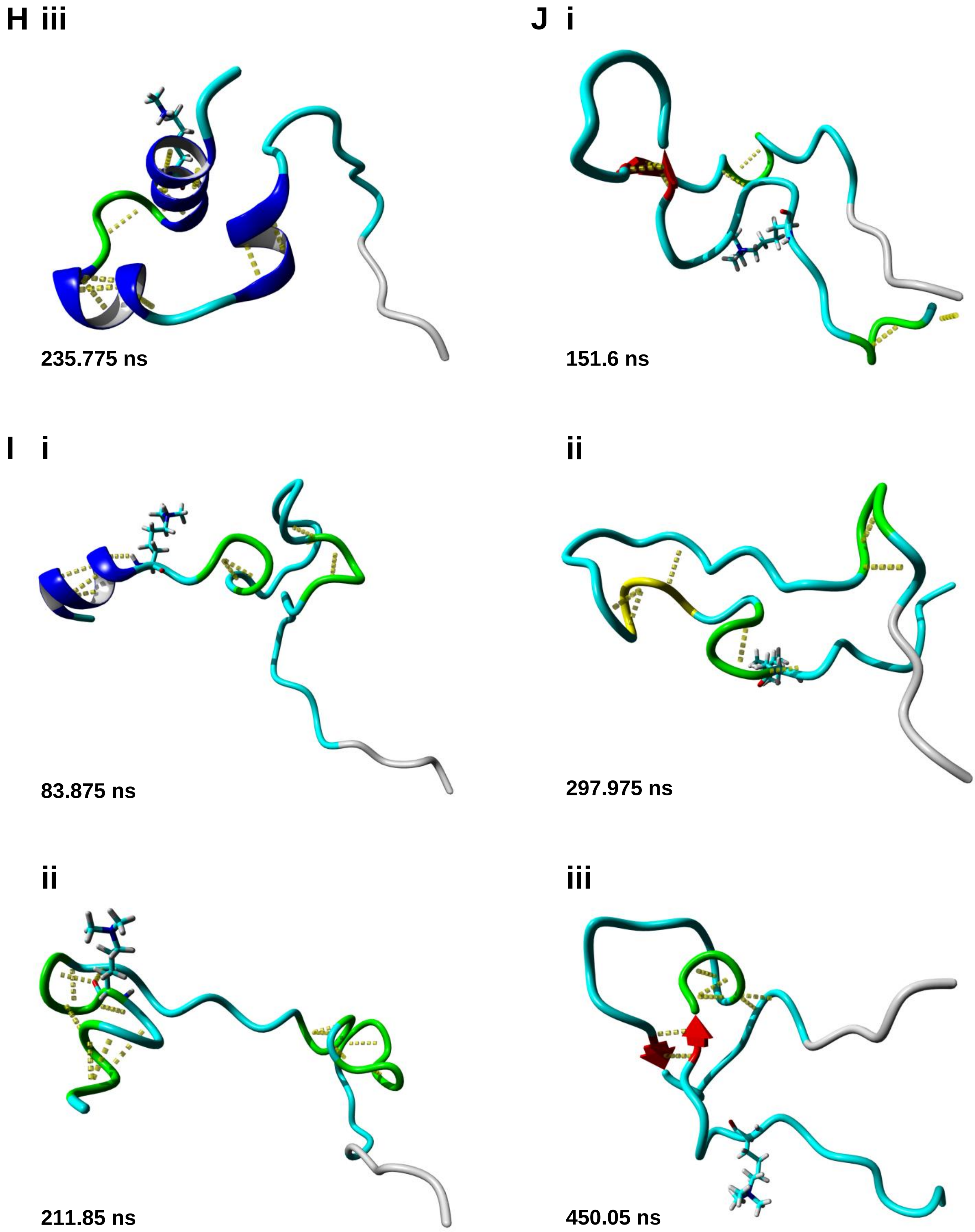


Figure 4.28 (cont.) Representative structures of the most populated clusters for the K9ME1 (H), K9ME2 (I) and K9ME3 (J) tail obtained from 500 ns - explicit MD simulations.

Molecular graphics created with YASARA (www.yasara.org) and POVRay (www.povray.org)

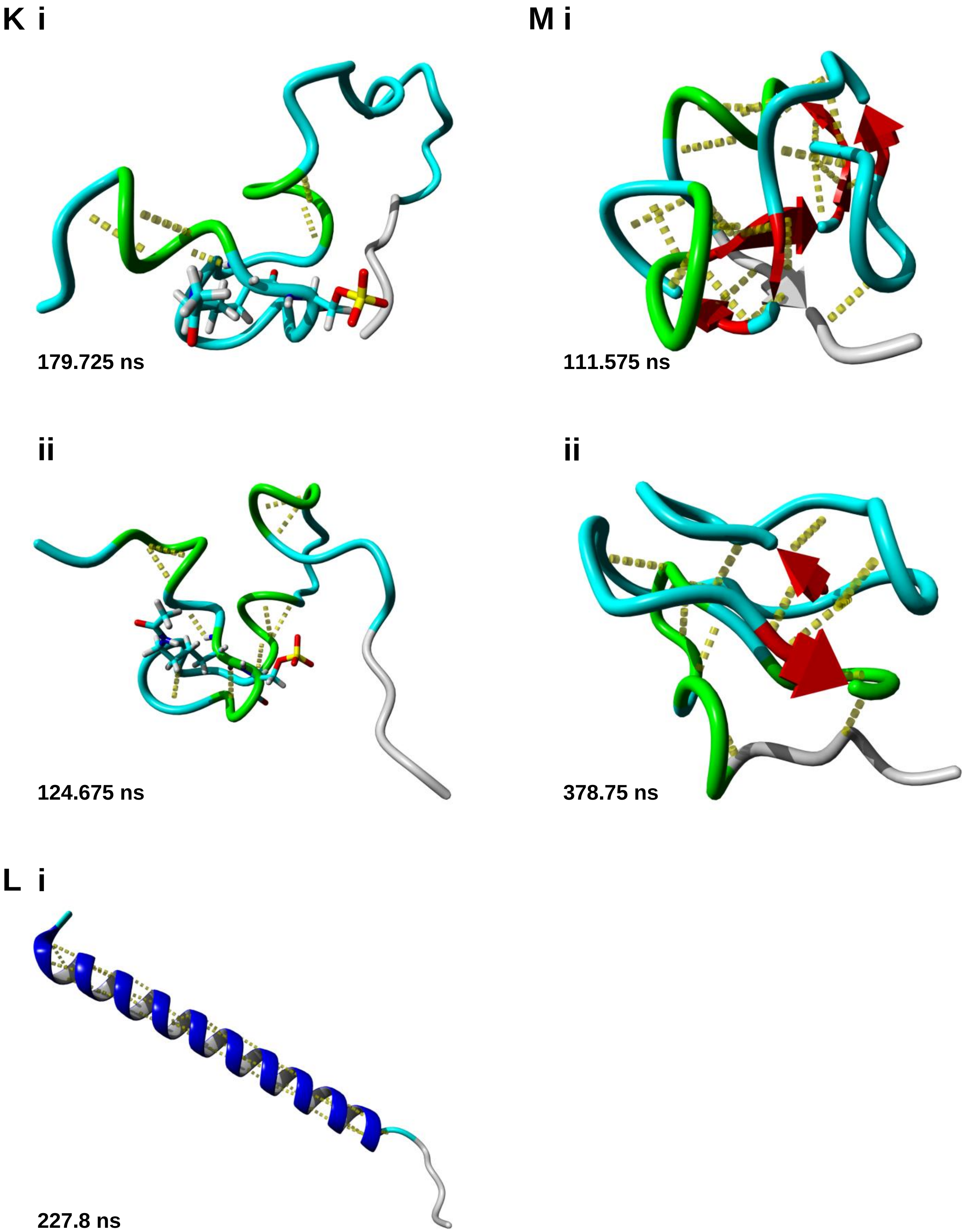


Figure 4.28 (cont.) Representative structures of the most populated clusters for the K9ACE_S10PHO (K), ALA_POS_CTRL (L) and GLY_NEG_CTRL (M) tail obtained from 500 ns - explicit MD simulations.

Molecular graphics created with YASARA (www.yasara.org) and POVRay (www.povray.org)

4.4 DISCUSSION

4.4.1 The unmodified tail shows the formation of two distinct α – helical regions

The histone tails are generally regarded as being unstructured coils [49]. However, the 500 ns explicit MD simulation of the unmodified histone H3 tail showed the formation of two distinct α – helices. The first and more abundant α – helix was at the N-terminal tip of the tail between T3 and G12, while the second and less abundant α – helix was towards the middle of the peptide between L20 and R26. These results were in agreement with both the secondary structure prediction and previous MD studies on the histone H3 N-terminal tail [50, 51]. While limited experimental information is available on the structure of the H3, circular dichroism results obtained by Banéres and co – workers showed that the H3 and H4 tails were bound to the nucleosome core particle (NCP) with half of their residues combined in an α – helical conformation [52]. Again this hinted at the α – helical character in the H3 tail, which was shown in this study.

4.4.2 Active tail versus Inactive tail – a structural difference

The active tail showed a lack of α – helical content; preferring hydrogen bonded turns and produced a bulge like structure as seen in Figure 4.28.B. In contrast, the inactive showed the stabilization of an α – helix between L20 and K27, a hydrogen bonded turn / 3-10 helix between K9 and G13 and a hydrogen bonded turn between P30 and G33 (see Figure 4.28.C). The phosphorylated serine residues played an important role in the stabilization of the structure observed. Firstly the most abundant hydrogen bond pair observed was R26/S28 which occurred between 30 000 and 35 000 times and the pairs R8/S10 and S10/R17 which occurred between 15 000 and 20 000 times. Examining the representative structure of the top cluster (Figure 4.28. C.i), it is seen that one of the phosphor oxygen atoms acts as hydrogen bond acceptor to two of the side chain amides of R26 (Figure 4.29.A). The same observation is made for S10 (Figure 4.29.B), where the phosphor oxygen atoms was involved in hydrogen bonds with the side chains of R8 and R17. Moving to the N-terminus of the α – helix (Figure 4.29.C), G34 was positioned in close proximity to L20 for its carboxyl oxygen to act as a hydrogen bond acceptor for the backbone amide from L20. Additionally, the side chain of K23 was involved in a hydrogen bond with the carboxyl oxygen from G33. The hydrogen bonded turn between G30 and G33 was stabilized by a hydrogen bond between the two residues (Figure 4.29.D), while the turn was positioned in such a way that the R17 amide was involved in a hydrogen bond with the carboxyl oxygen of T32. The side chain OH group of T32 was involved in a hydrogen bond with the carboxyl oxygen of T11. The interacting residue pairs for the hydrogen bonds mentioned were detected in the hydrogen bonding count matrix (Figure 4.19). The di – methylated lysines subsequently did not make any noticeable interactions in the formation of hydrogen bonds, with K9 only being involved in a hydrogen bond with G12, and not forming hydrogen bonds with Q5 and T6, as was observed in the unmodified tail. These hydrogen bond pairs occurred between 5000 and 10 000 times. Also, the backbone of K27 was part of the α – helix by hydrogen bonding with K23. This bond pair also occurred between 5000 and 10 000 times. Another noticeable difference displayed by the inactive tail was the absence of secondary structure in the N-terminal end of the tail between A1 and K9. The unique structure stabilized also supported experimental evidence available. Eberlin and co-workers were able to identify a unique structure associated with the inactive modifications exclusively between diplotene and metaphase stages in the cell cycle of spermatocytes and oocytes [6]. They were able to detect the unique structure with an antibody, and speculated that the antibody recognized a hydrogen bond between the phosphorylated S10 and R8. However, the K9ME2_S10PHO tail showed a similar abundance of the R8/S10 hydrogen bond pair. Yet, experimentally, the antibody did not bind to tails with the K9ME2_S10PHO modification. However there was a difference in the abundance of the R17/S10 hydrogen bond pair between the inactive and the K9ME2_S10PHO tails, with the pair occurring between 15 000 and 20 000 times in the inactive tail versus 5000 to 10 000 in the K9ME2_S10PHO tail. Thus the antibody most likely recognized either only the S10/R17 hydrogen bond or both the S10/R17 and S10/R8 hydrogen bonds.

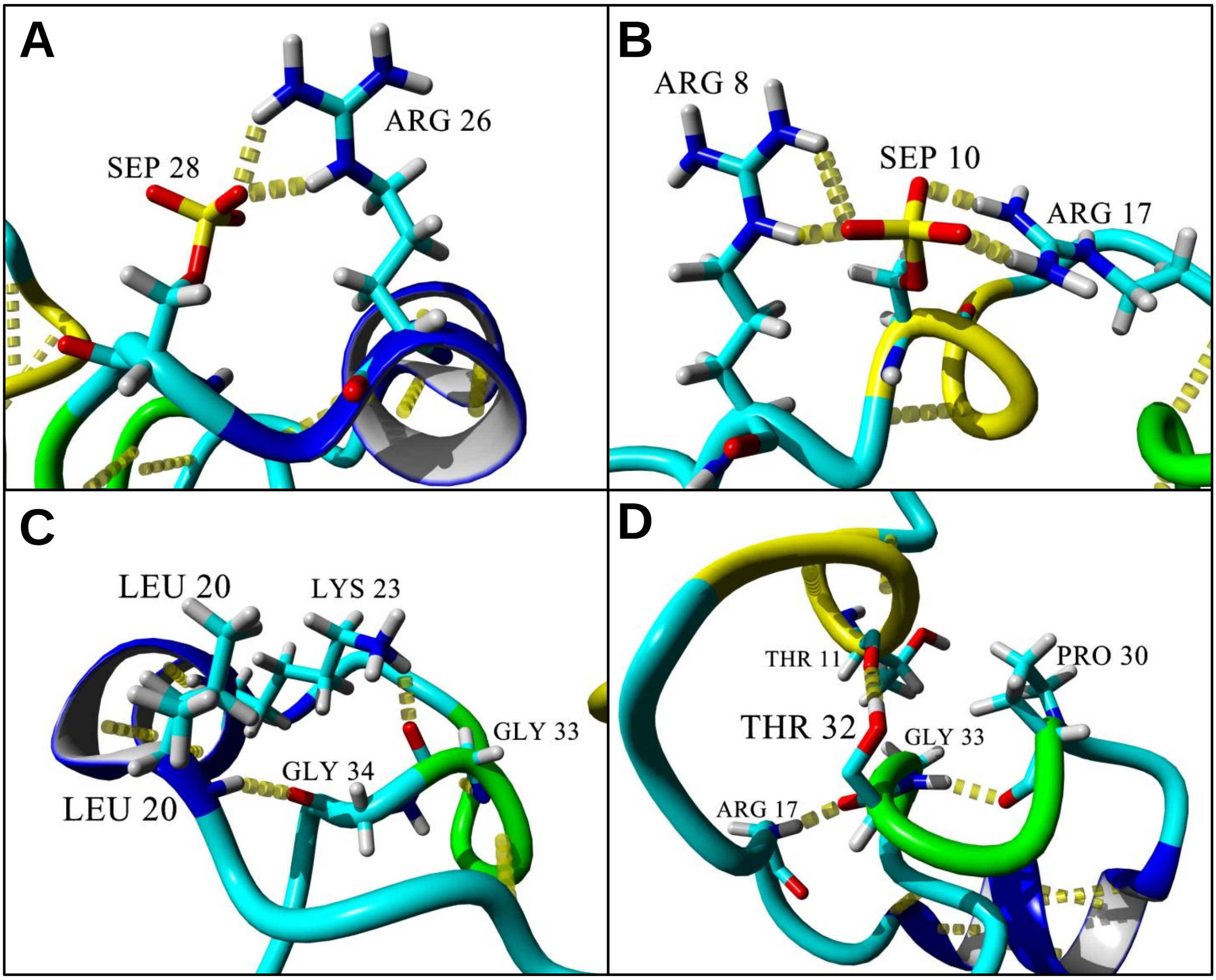


Figure 4.29 Stabilization of the unique structure found in the inactive tail. Panel A shows the S28 hydrogen bonding to the R26 side chain. Panel B shows S10 hydrogen bonding the side chains of R8 and R17. Panel C shows the stabilization of the N-terminal end of the α – helix through an extensive hydrogen bonding network. Panel D shows how the hydrogen bonded turn is maintained in position by hydrogen bonding.

Molecular graphics created with YASARA (www.yasara.org) and POVRay (www.povray.org)

The actively modified H3 tail showed a different structure from the structure of the inactive tail. From the hydrogen bonding analysis, the hydrogen bond pair K14/R17 was identified as the highest occurring pair, occurring between 15 000 and 20 000 times. Studying the most representative structure for the most populated cluster (Figure 4.28.B) revealed that R17 played a key role in the structure formed. In Figure 4.30.A the K14/R17 bond is visualized together with other contacts made by the R17 backbone. The backbone amide of R17 formed a hydrogen bond with the carboxyl oxygen of K14. Incidentally, the acetylated side chain of K14 was not involved in any hydrogen bonds. The carboxyl oxygen of R17, in turn, was involved in a hydrogen bond with the backbone amide of Q19. The hydrogen bond pair R17/Q19 was shown to occur between 5000 and 10 000 times during the simulation. Two of the side chain amides of R17 were involved in two hydrogen bonds with the carboxyl oxygen of G33, while the other side chain amide was involved in a hydrogen bond with the carboxyl oxygen of A25 (see Figure 4.30.B). The R17/G33 hydrogen bond pair was shown to occur between 5000 and 10 000 times, while the R17/A25 pair occurred between 10 000 and 15 000 times. The side chain of R17 was positioned in the middle of the loop between A25 and G33, which subsequently stabilized the loop. The overall structure of the bulge in Figure 4.28.B was actually two loops stacked on top of each other, with the loop stabilized by R17 in the top loop.

Figure 4.30.C shows how the loops were stabilized on top of each other. The carboxyl oxygen of P16, located on the bottom loop, was involved in a hydrogen bond with the backbone amide of A25, located on the top loop. The second hydrogen bond was formed between the side chain hydroxyl oxygen of T11, located on the bottom loop, and the carboxyl oxygen of G34, located on the top loop. Both the hydrogen bond pair occurred between 10 000 and 15 000 times during the simulation. In contrast, with the stabilization of the top loop, the bottom loop was primarily stabilized at the stem of the loop by two hydrogen bonds (see Figure 4.30.D). The first hydrogen bond was between the carboxyl oxygen of Q19 and the backbone amide of K9, and the second hydrogen bond was between the carboxyl oxygen of A7 and the backbone amide of A21. Hydrogen bonding analysis showed both hydrogen bond pairs occurring between 10 000 and 15 000 times. Neither the side chains of the acetylated lysines, K9 and K14, nor the tri – methylated lysine, K4 and K36, were involved in any significant hydrogen bonding. Interestingly, the acetylated lysines were located directly opposite each other on the bottom loop (Figure 4.30.E), with both side chains pointing upwards, and the two methylated lysines were also located directly across from each other on either end of the bulge, with their side chains pointing outwards (Figure 4.30.F).

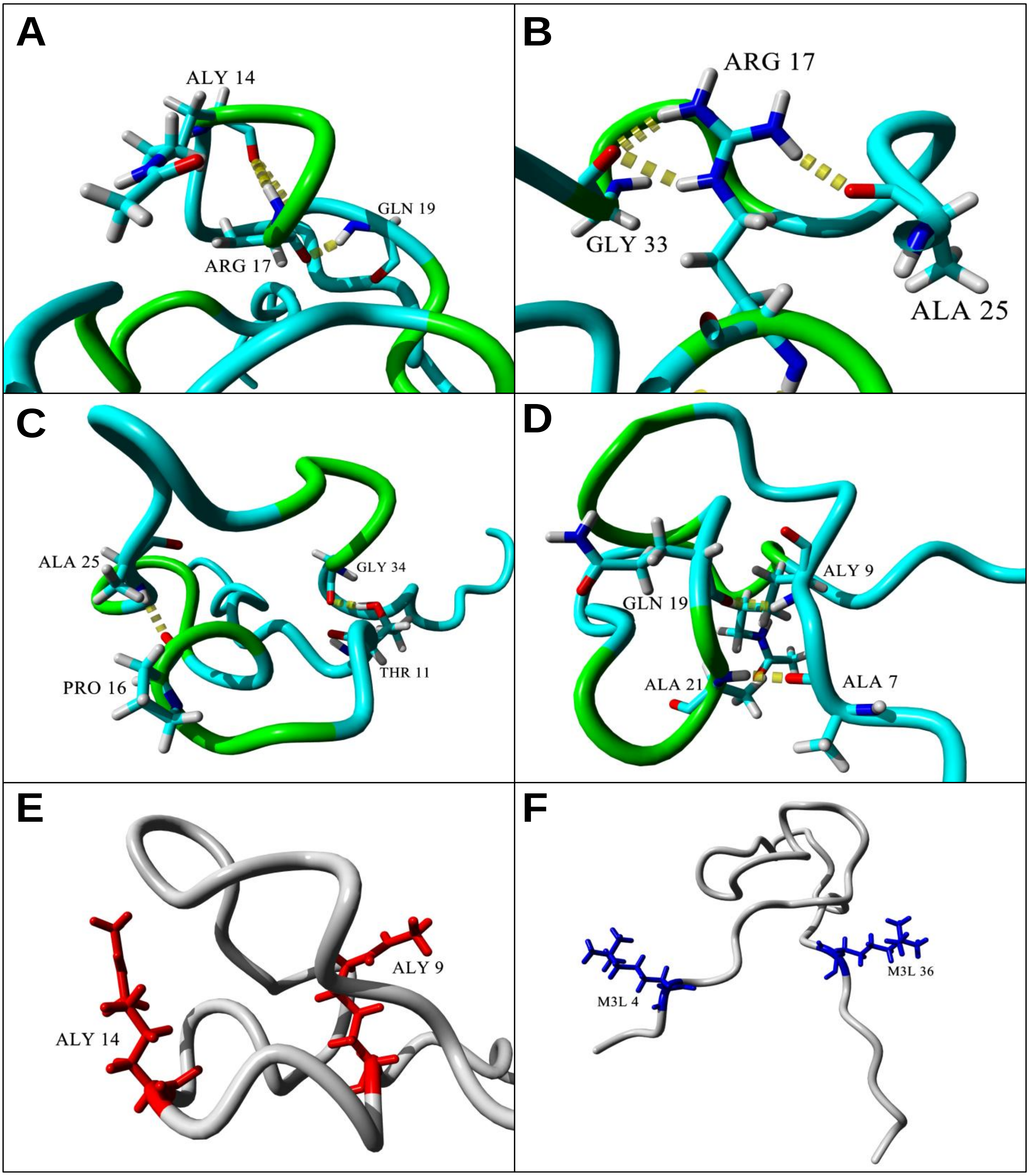


Figure 4.30 The stabilization of the active structure. Panel A shows the interactions of the R17 backbone. Panel B shows the interactions of the R17 side chain. Panel C shows how the two loops are stabilized on top of each other. Panel D shows the stabilization of the bottom loop. Panel E shows the position and orientation of the acetylated K9 and K14 (indicated in red). Panel F shows the position and orientation of the tri – methylated K4 and K36 (indicated in blue).

Molecular graphics created with YASARA (www.yasara.org) and POVRay (www.povray.org)

It is clear from the results that there is a significant transition in conformation from the actively to the inactively modified H3 tails. The role of serine phosphorylation was clearly demonstrated, as its phosphor oxygens acted as hydrogen bond acceptors primarily to the side chains of arginine residues. Comparing the active and inactive tail clearly showed the disruptive effect of particularly R17 to the formation and stabilization of an α – helix in the region of L20 – K27. The two phosphorylation sites in the inactive tail showed two separate mechanisms which additively allowed the formation and stabilization of the α – helix between L20 – K27. Firstly, S10 worked indirectly by "pulling" the disruptive R17 away from the N-terminal of the α – helix. This prevented the R17 side chain from hydrogen bonding with G33 and with A25, as seen in the active tail. Thus A25 was free to form part of the α – helix through hydrogen bonding and the carboxyl of G33 was free to hydrogen bond with the side chain amide of K23. This shifted the hydrogen bonded turn between P30 and G33 into position where T32 could participate in two hydrogen bonds: firstly the carboxyl oxygen with the amide backbone of R17 and secondly the hydroxyl side chain oxygen with the carboxyl oxygen of T11.This, in turn, allowed G34 to serve as the final hydrogen acceptor to the amide of L20, effectively capping/stabilizing the α – helix (G34 hydrogen bonded with T11 in the active tail).

In contrast to the elaborate mechanism proposed for S10, S28 directly stabilized the α – helix by hydrogen bonding with the side chain of R26, and thus preventing its side chain from making destabilizing interactions.

Oddly enough, the side chains of the modified lysines did not feature in any significant hydrogen bonding within the structure of the peptide in the context of the active and inactive modifications. The side chains were often found pointing outwardly, and would thus be more accessible. This makes sense in light of the histone code hypothesis [53], where modifications on the side chains serve as molecular beacons for effector proteins which bind to them.

4.4.3 Hyper acetylation induces to a loss of secondary structure in the H3 tail

Hyper acetylation of the H3 tail led to the loss of any helical content found in the unmodified tail. Hydrogen bonding analysis showed more hydrogen bond pairs occurring between 1000 and 5000 times and the overall pattern resembled the pattern observed for the glycine control. One might expect the side chains of the acetylated lysines to be involved in extensive hydrogen bonding, though the representative structures of the most populated clusters showed few hydrogen bonds involving the side chains of acetylated lysines. These structures resembled the structures obtained for the active tail, however in the hyper - acetylated tail the N – terminal tip is more tightly bound to the bulge structure obtained. Experimental evidence have shown that hyper – acetylation of the H3 and H4 tails decreased the initial melting temperature [9] and decreased the linking number [10] of the linker DNA, indicating that the entering – and exiting DNA were more mobile. Given that the tails were associated with the nucleosome at low salt concentrations [11], and that the H3 tail exits the nucleosome between the two DNA gyres at the entry/exit point of the linker DNA, it is likely that the H3 tail binds the entry/exiting DNA, immobilizing it. Hyper-acetylation eliminates the positive charge on 24 lysine residues in the H3 tail, and as observed in this study, changes the structure of the tail significantly. Thus, it is proposed that the H3 tail binds to the entering/exiting linker DNA and the N-terminal tip of the H3 tail binds to the nucleosomal surface, effectively “strapping” the linker DNA to the NCP. Upon hyper - acetylation the tail dissociates from the DNA because the positive charge on the lysine residues is eliminated, and thus their binding to the negatively charged DNA is weakened. The tail subsequently folds into a compressed structure and retracts it N – terminal tip from the nucleosomal surface, allowing the entering and exiting linker DNA to become mobile.

4.4.4 Methylation of K9 – Invisible hand guiding tail folding?

The K9ME1 tail produced a structure which bore some resemblance to the unmodified tail. Comparing the secondary structure histograms of the WT tail (Figure 4.2.B) and the K9ME1 tail (Figure 4.8.B) shows that both contained the N – terminal tip α – helix at a similar frequency, though in K9ME1 the N-terminal tip was extended and included A1 - R2. However, the K9ME1 showed two α – helices: P16 – A21 and K23 – S28, instead of the one α – helix, L20 – R26, in the unmodified tail. Figure 4.28.H.iii showed that the three α- helices occurred at the same time. The hydrogen bonding analysis (Figure 4.24 B) only showed that the Q5/K9 and the T6/K9 hydrogen bond pairs were missing in the K9ME1 tail, and occurred between 5000 and 10 000 times more in the WT tail. These hydrogen bonds were identified as backbone hydrogen bonds by using the structures in Figure 4.28.A. All the differences in the other hydrogen bond pair occurrences were part of the different α – helices in the two simulations.

The di – methylated side chain of K9 had almost no effect on the secondary structure or hydrogen bonding pattern of the unmodified tail. Both simulations produced almost identical secondary structure histograms (see Figure 4.2.B and 4.9.A). One difference was an extended tip α – helix towards the N-terminus, including the residues A1 and R2. This extension was picked up by the normalized hydrogen bond matrix as the R2/T6 hydrogen pair which occurred between 5000 and 10 000 times. A second difference was the presence of a hydrogen bonded turn / 3-10 helix region between P16 and L20 in the unmodified tail which was absent in the K9ME2 tail.

Whereas K9ME1 showed a structure resembling the unmodified structure and K9ME2 came close to reproducing it, K9ME3 produced a completely different structure (Figure 4.28.J). The K9ME3 tail showed little α – helical content during the simulation (Figure 4.10.A) and instead showed the formation of β – bridges, R17/R26 and A15/A29. The hydrogen bonding analysis showed that these hydrogen bond pairs occurred between 10 000 and 15 000 times. The representative cluster structures (see Figure 4.28 J) and the time wise evolution of secondary structure plot (see Figure 4.11 A) showed that the β- bridges were mutually exclusive. Again, no significant hydrogen bond pairs were identified involving K9. However, R17 again proved an instrumental agent in changing the structure. It formed hydrogen bond pairs with A21, T22 and K27 which occurred between 5000 and 10 000 times and was also directly involved in one of the β – bridges, observed as a hydrogen bond pair with R26 and occurring between 10 000 and 15 000 times in the simulation.

Methylated lysines have always been thought of as molecular beacons for effector proteins in the context of chromatin biology for two reasons. Firstly, the discovery of a vast family of lysine methylases and lysine di-methylases specific to different methylated states and residues within the various histone tails [54-56] and, secondly, because methylation leaves the charge state of the lysine residue intact, unlike lysine acetylation which naturalizes the positive charge on the side chain. However, methylation caps the side chain of the lysine residue with bulky and hydrophobic methyl groups, which is thought to prevent the side chain from forming hydrogen bonds with other residues [57]. However, according to the results obtained, the side chain of K9, methylated, acetylated or unmodified, were involved in minimal hydrogen bonding. Yet different structures were obtained with the different degrees of K9 methylation. A possible explanation could be that hydrophobic side chain of K9 subtlety guides the folding of the peptide, though hydrophobic interactions [58, 59], which places other residues in specific positions where they can then make specific interactions. A perfect example of this was R17, where the position of the side chain essentially dictated the structure formed. Also, one would not expect the side chain of K9 to be directly involved in hydrogen bonding or direct interaction, because effector proteins would still need to be able to access and bind to it.

Mono – methylated H3K9 was found to be enriched in the body of active genes, while di – methylated and tri – methylated H3K9 was found to be depleted in active genes and enriched in non-genic regions [60]. Also, while similar levels of di – methylated H3K9 and tri – methylated H3K9 was found in gene deserts and telomeric regions, the levels of tri – methylated H3K9 was almost double that of di – methylated H3K9 in peri-centromeric regions. It has also been shown that tri – methylated H3K9 was present at active gene regions [61]. Thus, the structural variation observed with different levels of methylation of K9 may in fact hint at the idea that methylation may play a structural role in defining separate regions in the nuclear architecture.

4.4.5 Serine 10 phophorylation – Epigenetic Mercenary?

The tail with mono-methylated K9 and phosphorylated S10 displayed a different structure from that of either the WT tail or the K9ME1 tail. The structure contained β – bridges between A7/R8 and G12/G13 and an α – helix between T22 and A29. From the hydrogen bonding analysis, the hydrogen bond pair R8/S10 occurred between 25 000 and 30 000 times. Examining Figure 4.28.E revealed that the side chain of R8 hydrogen bonded with the phosphor oxygens of the phosphorylated S10. Subsequently, the hydrogen bonding analysis also showed the hydrogen bond pair, S10/R17, occurred between 15 000 and 20 000 times. However this hydrogen bond pair was not observed in Figure 4.28.E. An interesting observation was the hydrogen bond pair T22/A25, which occurred between 5000 and 10 000 times. Figure 4.28.E showed that the side chain hydroxyl oxygen hydrogen bonded with the backbone amide of A25. This phenomenon is known as N-terminal α – helix capping [62].

K9ME2_S10PHO showed a diminished ability to form α – helices, only showing a small percentage of α – helix in two regions, T3 – R8 and L20 – A25. The most abundant hydrogen bond pairs involved S10 with V35 and R8 and occurred between 10 000 and 15 000, and between 15 000 and 20 000 times, respectively. Examining the representative structures, from the most populated clusters showed that S10 carboxyl oxygen hydrogen bonded with the backbone amide of V35, forming the stem of the large loop between S10 and V35.

Similarly, the K9ME3_S10PHO tail also showed very little α- helical content. Only one hydrogen bond pair occurred more than 1000 times, R8/S10, which occurred between 20 000 and 25 000.

Given that any α – helical character in the region of the N-terminal tip was lost with the phosphorylation, suggests that R8 might play an important role in the formation of an α – helix at the N-terminal tip of the H3 tail. Comparing Figure 4.28.G and Figure 4.28.J, revealed that the subset of structures obtained for K9ME_S10PHO were different from the ones obtained for the K9ME3 tail.

The K9ACE_S10PHO tail showed a similar α – helical profile as K9ME2_S10PHO; however the structures obtained from the most populated cluster were different (Figure 4.28.K and Figure 4.28.F). The K9ACE_S10PHO structures only showed hydrogen bonded turns in terms of secondary structure elements. The hydrogen bonding analysis revealed that the hydrogen bond pairs, R8/S10, K23/S10 and R26/S10, occurred between 5000 and 10 000 times during the simulation. The only hydrogen bond visualized (Figure 4.28.K.ii), was the K23/S10 pair which showed the side chain of K23 hydrogen bonding with the side chain of S10.

Phosphorylated serine was shown to perturb the structure of all the tails studied which contained this modification. It was also shown that the phosphorylated serine was capable of perturbing the structure by its phosphor oxygens acting as hydrogen bond acceptors to primarily the side chain amide groups of arginine side chains. Also the positioning of the phosphorylated serine residues was important. S10 is located close to R8 and within reach of R17. Both S10 and R8 were shown to be part of the N-terminal tip α – helix in the unmodified tail and all tails containing S10PHO was shown to have a significantly reduced α – helical content. Examining these tails' normalized hydrogen bonding, the missing hydrogen bond pair were usually T6/S10 and K4/R8 at an occurrence level of at least 1000 - 5000 times. Also S10PHO was shown to prevent R17 from interacting near the N-terminal end of the α – helix found in the region L20 and onwards.

Phosphorylated S10 was shown to be present at chromatin condensation during mitosis and meiosis [6, 63], however S10 was also found to be associated with active transcription [64, 65]. This correlates with the results obtained in this study to some degree, because different structures were obtained with each of the S10PHO tails and these structures were also different from the structures obtained by their S10 counterparts. Thus, if the structure of the tail was a relevant determinant in the compaction state of chromatin and thus the level of transcription, S10PHO could then be observed in both compressed and decompressed regions. Joeng and co – workers made the interesting observation that S10PHO prevented their antibodies from binding to K9ME2, and thus prevented detection of the modification [66]. Even more interesting was the observation that this blockage did not occur with K9ME3 antibodies or with the adjacent lysine next to phosphorylated S28 at any level of methylation. They subsequently speculated that the phosphor moiety directly interfered with antibody. However, results obtained in this study might lead one to rather speculate that the change in the structure around K9 and S10 affected antibody binding. At this time it is proposed that the modified side chain of phosphorylated S10 will hydrogen bond to available hydrogen bond donors, especially with arginine and lysine side chains, in its immediate environment, while it is the modification state of K9 that guides the folding of tails in such a manner as to place certain hydrogen bond donors within reach of the phosphorylated S10, and in this way generate the structural variety observed in this study.

4.5 CONCLUSION

In this chapter 500 ns explicit MD simulations showed that the unmodified tail of histone H3 contained two α – helices as predicted by secondary structure prediction algorithms. By introducing post translation modifications into the tail, it was subsequently shown that these modifications induced unique structural changes in the H3 tail. Among these were the formation of stable α – helices in the inactive tail and further analysis showed that the phosphorylated serine residues at position 10 and 28 were directly involved in the stabilization of the structure. The active tail showed a completely different structure. It was established that the side chains of the modified lysines did not directly change the structure through non-bonded interactions, however, they did influence the structures obtained, particularly those with lysine methylation; the structural changes were likely due to a hydrophobic effect. Serine 10 phosphorylation was shown to induce structural change by its side chain oxygen group’s hydrogen bonding with the amide groups of especially arginine side chains. However, the phosphate oxygens would hydrogen bond any suitable donor within its proximity and thus the overall structural effect of the modification relied on the adjacent K9 to position the appropriate hydrogen bond donor within reach, and thus produce a unique structure within the tail. This would explain why S10 phosphorylation is found in active as well as in silent chromatin regions. In conclusion, it was shown that the histone H3 tail does have defined structures and that these structures were interchangeable by the addition of post translational modification on certain residues. This observation adds a new structural dimension to chromatin biology and provides a new avenue for future research.

4.6 REFERENCES

1. Grauffel, C., Stote, R. H. & Dejaegere, A. (2010). Force field parameters for the simulation of modified histone tails. *J. Comput. Chem.* **31**, 2434-2451.
2. Zhao, G. J. & Cheng, C. L. Molecular dynamics simulation exploration of unfolding and refolding of a ten-amino acid miniprotein. *Amino Acids* 1-9.
3. Niggli, D. A., Ebert, M. O., Lin, Z., Seebach, D. & van Gunsteren, W. F. (2012). Helical Content of a β^3- Octapeptide in Methanol: Molecular Dynamics Simulations Explain a Seeming Discrepancy between Conclusions Derived from CD and NMR Data. *Chem. Eur. J.* **18**, 586-593.
4. Bhaumik, S. R., Smith, E. & Shilatifard, A. (2007). Covalent modifications of histones during development and disease pathogenesis. *Nat Struct Mol Biol* **14**, 1008-1016.
5. Guenther, M. G., Levine, S. S., Boyer, L. A., Jaenisch, R. & Young, R. A. (2007). A Chromatin Landmark and Transcription Initiation at Most Promoters in Human Cells., pp. 77-88.
6. Eberlin, A., Grauffel, C., Oulad-Abdelghani, M., Robert, F., Torres-Padilla, M. E., Lambrot, R., Spehner, D., Ponce-Perez, L., Würtz, J. M., Stote, R. H., Kimmins, S., Schultz, P., Dejaegere, A. & Tora, L. (2008). Histone H3 Tails Containing Dimethylated Lysine and Adjacent Phosphorylated Serine Modifications Adopt a Specific Conformation during Mitosis and Meiosis. *Molecular and Cellular Biology* **28**, 1739-1754.
7. Mateescu, B., England, P., Halgand, F., Yaniv, M. & Muchardt, C. (2004). Tethering of HP1 proteins to chromatin is relieved by phosphoacetylation of histone H3. *EMBO Rep* **5**, 490-496.

8. Edmondson, D. G., Davie, J. K., Zhou, J., Mirnikjoo, B., Tatchell, K. & Dent, S. Y. R. (2002). Site-specific Loss of Acetylation upon Phosphorylation of Histone H3. *Journal of Biological Chemistry* **277**, 29496-29502.

9. BODE, J., HENCO, K. & WINGENDER, E. (1980). Modulation of the Nucleosome Structure by Histone Acetylation. *European Journal of Biochemistry* **110**, 143-152.

10. Norton, V. G., Marvin, K. W., Yau, P. & Bradbury, E. M. (1990). Nucleosome linking number change controlled by acetylation of histones H3 and H4. *Journal of Biological Chemistry* **265**, 19848-19852.

11. Wang, X. & Hayes, J. J. (2006). Physical methods used to study core histone tail structures and interactions in solution. *Biochemistry and Cell Biology* **84**, 578-588.

12. Marqusee, S., Robbins, V. H. & Baldwin, R. L. (1989). Unusually stable helix formation in short alanine-based peptides. *Proceedings of the National Academy of Sciences* **86**, 5286-5290.

13. Go, M., Go, N. & Scheraga, H. A. (1970). Molecular Theory of the Helix-Coil Transition in Polyamino Acids. II. Numerical Evaluation of s and sigma for Polyglycine and Poly-l-alanine in the Absence (for s and sigma) and Presence (for sigma) of Solvent. *The Journal of Chemical Physics* **52**, 2060-2079.

14. Shental-Bechor, D., Kirca, S., Ben-Tal, N. & Haliloglu, T. (2005). Monte Carlo Studies of Folding, Dynamics, and Stability in α - Helices., pp. 2391-2402.

15. Krieger, E., Koraimann, G. & Vriend, G. (2002). Increasing the precision of comparative models with YASARA NOVA – a self-parameterizing force field. *Proteins* **47**, 393-402.

16. Davey, C. A., Sargent, D. F., Luger, K., Maeder, A. W. & Richmond, T. J. (2002). Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 Å resolution. *J. Mol. Biol.* **319**, 1097-1113.

17. Wu, H., Min, J., Lunin, V. V., Antoshenko, T., Dombrovski, L., Zeng, H., Allali-Hassani, A., Campagna-Slater, V. r., Vedadi, M., Arrowsmith, C. H., Plotnikov, A. N. & Schapira, M. (2010). Structural Biology of Human H3K9 Methyltransferases. *PLoS ONE* **5**, e8570.

18. Yap, K. L., Li, S., Muñoz-Cabello, A. M., Raguz, S., Zeng, L., Mujtaba, S., Gil, J., Walsh, M. J. & Zhou, M. M. (2010). Molecular Interplay of the Noncoding RNA ANRIL and Methylated Histone H3 Lysine 27 by Polycomb CBX7 in Transcriptional Silencing of INK4a., pp. 662-674.

19. Kaustov, L., Ouyang, H., Amaya, M., Lemak, A., Nady, N., Duan, S., Wasney, G. A., Li, Z., Vedadi, M., Schapira, M., Min, J. & Arrowsmith, C. H. (2011). Recognition and Specificity Determinants of the Human Cbx Chromodomains. *Journal of Biological Chemistry* **286**, 521-529.

20. Zeng, L., Zhang, Q., Li, S., Plotnikov, A. N., Walsh, M. J. & Zhou, M. M. (2010). Mechanism and regulation of acetylated histone binding by the tandem PHD finger of DPF3b. *Nature* **466**, 258-262.

21. Macdonald, N., Welburn, J. P. I., Noble, M. E. M., Nguyen, A., Yaffe, M. B., Clynes, D., Moggs, J. G., Orphanides, G., Thomson, S., Edmunds, J. W., Clayton, A. L., Endicott, J. A. & Mahadevan, L. C. (2005). Molecular Basis for the Recognition of Phosphorylated and Phosphoacetylated Histone H3 by 14-3-3., pp. 199-211.

22. Krieger, E., Darden, T., Nabuurs, S. B., Finkelstein, A. & Vriend, G. (2004). Making optimal use of empirical energy functions: Force-field parameterization in crystal space. *Proteins* **57**, 678-683.

23. Duan, Y., Wu, C., Chowdhury, S., Lee, M. C., Xiong, G., Zhang, W., Yang, R., Cieplak, P., Luo, R., Lee, T., Caldwell, J., Wang, J. & Kollman, P. (2003). A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *Journal of Computational Chemistry* **24**, 1999-2012.

24. Essmann, U., Perera, L., Berkowitz, M., Darden, T., Lee, H. & Pedersen, L. (1995). A smooth particle mesh Ewald method. *The Journal of Chemical Physics* **103**, 8577-8593.

25. Krieger, E., Nielsen, J. E., Spronk, C. A. E. M. & Vriend, G. (2006). Fast empirical pKa prediction by Ewald summation. *Journal of Molecular Graphics and Modelling* **25**, 481-486.

26. Jakalian, A., Jack, D. B. & Bayly, C. I. (2002). Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *Journal of Computational Chemistry* **23**, 1623-1641.

27. Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. (2004). Development and testing of a general amber force field. *Journal of Computational Chemistry* **25**, 1157-1174.

28. Stewart, J. J. P. (1990). MOPAC: A semiempirical molecular orbital program. *Journal of Computer-Aided Molecular Design* **4**, 1-103.

29. Klamt, A. (1995). Conductor-like Screening Model for Real Solvents: A New Approach to the Quantitative Calculation of Solvation Phenomena. *J. Phys. Chem.* **99**, 2224-2235.

30. Chou, P. Y. & Fasman, G. D. (1974). Prediction of protein conformation. *Biochemistry* **13**, 222-245.

31. Cole, C., Barber, J. D. & Barton, G. J. (2008). The Jpred 3 secondary structure prediction server. *Nucleic Acids Research* **36**, W197-W201.

32. Meiler, J., Müller, M., Zeidler, A. & Schmäñschke, F. (2001). Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks. *Journal of Molecular Modeling* **7**, 360-369.

33. Petersen, B., Petersen, T., Andersen, P., Nielsen, M. & Lundegaard, C. (2009). A generic method for assignment of reliability scores applied to solvent accessibility predictions. *BMC Structural Biology* **9**, 51.

34. Mooney, C. & Pollastri, G. (2009). Beyond the Twilight Zone: Automated prediction of structural properties of proteins by recursive neural networks and remote homology information. *Proteins* **77**, 181-190.

35. Pollastri, G., Martin, A., Mooney, C. & Vullo, A. (2007). Accurate prediction of protein secondary structure and solvent accessibility by consensus combiners of sequence and structure information. *BMC Bioinformatics* **8**, 201.

36. Pollastri, G. & McLysaght, A. (2004). Porter: a new, accurate server for protein secondary structure prediction. *Bioinformatics* **21**, 1719-1720.

37. Rost, B. (1996). [31] PHD: Predicting one-dimensional protein structure by profile-based neural networks. In *Methods in Enzymology Computer Methods for Macromolecular Sequence Analysis* (Russell, F. D., ed), pp. 525-539, Academic Press.

38. Rost, B., Yachdav, G. & Liu, J. (2004). The PredictProtein server. *Nucleic Acids Research* **32**, W321-W326.

39. Ouali, M. & King, R. D. (2000). Cascaded multiple classifiers for secondary structure prediction. *Protein Science* **9**, 1162-1176.

40. Cheng, J., Randall, A. Z., Sweredoski, M. J. & Baldi, P. (2005). SCRATCH: a protein structure and structural feature prediction server. *Nucleic Acids Research* **33**, W72-W76.

41. Pollastri, G., Przybylski, D., Rost, B. & Baldi, P. (2002). Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins* **47**, 228-235.

42. Combet, C., Blanchet, C., Geourjon, C. & Deléage, G. (2000). NPS@: Network Protein Sequence Analysis. *Trends in Biochemical Sciences* **25**, 147-150.

43. McGuffin, L. J., Bryson, K. & Jones, D. T. (2000). The PSIPRED protein structure prediction server. *Bioinformatics.* **16**, 404-405.

44. Hess, B., Kutzner, C., van der Spoel, D. & Lindahl, E. (2008). GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory Comput.* **4**, 435-447.

45. Budesinsky, M., Sebestik, J., Bednarova, L., Baumruk, V., Safarik, M. & Bour, P. (2008). Conformational Properties of the Pro-Gly Motif in the d-Ala-l-Pro-Gly-d-Ala Model Peptide Explored by a Statistical Analysis of the NMR, Raman, and Raman Optical Activity Spectra. *J. Org. Chem.* **73**, 1481-1489.

46. Forneris, F., Binda, C., Vanoni, M. A., Battaglioli, E. & Mattevi, A. (2005). Human histone demethylase LSD1 reads the histone code. *J. Biol. Chem.* **280**, 41360-41365.

47. Bannister, A. J., Zegerman, P., Partridge, J. F., Miska, E. A., Thomas, J. O., Allshire, R. C. & Kouzarides, T. (2001). Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. *Nature* **410**, 120-124.

48. Shareef, M. M., King, C., Damaj, M., Badagu, R., Huang, D. W. & Kellum, R. (2001). Drosophila Heterochromatin Protein 1 (HP1)/Origin Recognition Complex (ORC) Protein Is Associated with HP1 and ORC and Functions in Heterochromatin-induced Silencing. *Molecular Biology of the Cell* **12**, 1671-1685.

49. Luger, K., Mader, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. (1997). Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* **389**, 251-260.

50. Potoyan, D. A. & Papoian, G. A. (2011). Energy Landscape Analyses of Disordered Histone Tails Reveal Special Organization of Their Conformational Dynamics. *J. Am. Chem. Soc.* **133**, 7405-7415.

51. Liu, H. & Duan, Y. (2008). Effects of post-translational modifications on the structure and dynamics of histone H3 N-terminal peptide. *Biophys. J.* **94**, 4579-4585.

52. Banères, J. L., Martin, A. & Parello, J. (1997). The N tails of histones H3 and H4 adopt a highly structured conformation in the nucleosome. *J. Mol. Biol.* **273**, 503-508.

53. Strahl, B. D. & Allis, C. D. (2000). The language of covalent histone modifications. *Nature* **403**, 41-45.

54. Zhang, X., Wen, H. & Shi, X. (2012). Lysine methylation: beyond histones. *Acta Biochimica et Biophysica Sinica* **44**, 14-27.

55. Yap, K. L. & Zhou, M. M. (2011). Structure and Mechanisms of Lysine Methylation Recognition by the Chromodomain in Gene Transcription. *Biochemistry* **50**, 1966-1980.

56. Ng, S., Yue, W., Oppermann, U. & Klose, R. (2009). Dynamic protein methylation in chromatin biology. *Cellular and Molecular Life Sciences* **66**, 407-422.

57. Hansen, J. C., Lu, X., Ross, E. D. & Woody, R. W. (2006). Intrinsic Protein Disorder, Amino Acid Composition, and Histone Terminal Domains. *Journal of Biological Chemistry* **281**, 1853-1856.

58. Lins, L. & Brasseur, R. (1995). The hydrophobic effect in protein folding. *The FASEB Journal* **9**, 535-540.

59. Spolar, R. S., Ha, J. H. & Record, M. T. (1989). Hydrophobic effect in protein folding and other noncovalent processes involving proteins. *Proceedings of the National Academy of Sciences* **86**, 8382-8385.

60. Rosenfeld, J., Wang, Z., Schones, D., Zhao, K., DeSalle, R. & Zhang, M. (2009). Determination of enriched histone modifications in non-genic portions of the human genome. *BMC Genomics* **10**, 143.

61. Barski, A., Cuddapah, S., Cui, K., Roh, T. Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I. & Zhao, K. (2007). High-Resolution Profiling of Histone Methylations in the Human Genome., pp. 823-837.

62. Presta, L. G. & Rose, G. D. (1988). Helix signals in proteins. *Science* **240**, 1632-1641.

63. Wei, Y., Mizzen, C. A., Cook, R. G., Gorovsky, M. A. & Allis, C. D. (1998). Phosphorylation of histone H3 at serine 10 is correlated with chromosome condensation during mitosis and meiosis in *Tetrahymena*. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 7480-7484.

64. Labrador, M. & Corces, V. G. (2003). Phosphorylation of histone H3 during transcriptional activation depends on promoter structure. *Genes & Development* **17**, 43-48.

65. Nowak, S. J. & Corces, V. G. (2000). Phosphorylation of histone H3 correlates with transcriptionally active loci. *Genes & Development* **14**, 3003-3013.

66. Jeong, Y. S., Cho, S., Park, J. S., Ko, Y. & Kang, Y. K. (2010). Phosphorylation of serine-10 of histone H3 shields modified lysine-9 selectively during mitosis. *Genes to Cells* **15**, 181-192.

CHAPTER 5

Docking of the H3 N-terminal tail to the nucleosome core

5.1 INTRODUCTION

In chapter 3 it was shown how molecular docking with AutoDock[1] implemented in YASARA[2] could successfully predict the binding site of a basic peptide on the surface of the nucleosome. In this chapter the methods and parameters designed and derived respectively in chapter 3 will be used to evaluate the docking of the first 15 residues from the histone H3 N-terminal tip, using structures obtained from the MD trajectories generated in Chapter 4. It is known that the histone tails are bound to the nucleosome at NaCl concentrations below 2 M[3], and the H4 tail was shown to be bound to the acidic patch of the nucleosome in co – crystal structures[4]. Also, given that our test peptide in chapter 3, KSHV - LANA, was also bound to the nucleosome surface[5], we thought it possible that the H3 tail could, similarly, be bound somewhere on the nucleosome. The aim in this chapter, thus, was to find a binding site for the H3 N-terminal tail on the nucleosome surface, and to describe the effect of post translation modifications in the H3 tail on this binding.

5.2 METHODS

5.2.1 Clustering of the N-terminal tip structures from MD trajectories

Trajectories generated in Chapter 4 were clustered with the GROMACS[6] program g_cluster, using the single – linkage clustering algorithm with a 13 Å RMSD. RMSD values were calculated between the first 15 α – carbons of each structure in each trajectory.

5.2.2 Selection of structures for molecular docking

Three structures were selected from each trajectory generated in Chapter 4, excluding the control peptides. The trajectories are listed in Table 5.1. In cases where clustering yielded less than three substantially populated clusters, additional structures in the most populated cluster were chosen to be temporally spaced , i.e. a structure at the middle and the extremities of the cluster. The additional structures were chosen at random with regards to the temporal location of the RMSD - middle structure. For example, where the middle structure was found at 100 ns, the two additional structures would be sampled at 350 ns and 500 ns, in the case where the most populated cluster contained structures from 100 ns to 500 ns. The structures chosen from each trajectory is shown in Figure 5.1.

Table 5.1 Systems simulated in the MD experiments.

Label	System Title	Modifications
A	WT	None
B	ACTIVE	K4 + K36 Me3, K9 + K14 Ace
C	INACTIVE	K9 + K27 Me2, S10 + S28 Pho
D	HYPER_ALY	All K Ace
E	K9ME1_S10PHO	K9 Me1 + S10 Pho
F	K9ME2_S10PHO	K9 Me2 + S10 Pho
G	K9ME3_S10PHO	K9 Me3 + S10 Pho
H	K9ME1	K9 Me1
I	K9ME2	K9 Me2
J	K9ME3	K9 Me3
K	K9ACE_S10PHO	K9 Ace + S10 Pho

*** Ace – acetylation, Me1 – mono – methylation, Me2 – Di – methylation, Me3 – tri – methylation, Pho - phophorylation**

5.2.3 Molecular docking of N-terminal tip structures

Molecular docking of the 15 –residue histone H3 N-terminal tips were performed using Autodock in YASARA according to the method of rigid, grid – based docking developed and described in the

Methods (section 3.2) of Chapter 3. The partially over-lapping grid boxes were labeled A to C for the top to the bottom row, and 1 to 3 for the left to the right columns, viewed along the supercoil axis of the nucleosome, as is shown in Figure 3.4 of Chapter 3.

5.2.4 Analysis of docking results

Hydrogen bonding -and contact analysis were performed with YASARA and python scripts to evaluate and further define how the H3 tail contacts the nucleosome. A contact would generally be defined as atoms within a cut-off region of 4 Å from another atom, but not closer than the length of a hydrogen bond. The program would then iteratively search for the atoms in the receptor that satisfy this criterion for each atom in the ligand. These scripts are included in the directory titled "Chapter_5" on the included disk.

5.3 RESULTS

5.3.1 Clustering of the N-terminal tip structures from MD trajectories

In Chapter 3 we showed that flexible ligand docking could not correctly reproduce a known binding site of a ligand co-crystallized with the nucleosome core. However, rigid ligand docking was able to correctly place our test ligand, given the bound structure of the ligand. We thus decided to use a rigid, grid – based docking method to dock the H3 tail to the nucleosome. However, we needed a collection of structures which had the highest probability of being the bound structure of the tip of the histone H3 N-terminal tail, or very close to it. We thus used the MD trajectories obtained in Chapter 4, and clustered them based on the 15 N-terminal residues of the tail to obtain the N-terminal tip structures of the tail that could be docked to the nucleosome.

Clustering of the MD trajectories yielded 17 structures, of which 16 structures were selected from the highest scoring clusters (see Figure 5.1). Most structures did not have any secondary structure, such as seen in Figure 5.1.A.III, while other displayed minimal secondary structure elements, such as hydrogen bonded turns (Figure 5.1.A.II), and 3-10 helices (Figure 5.1.E.I). However, the K9ME1_S10PHO tip contained a β – bridge (Figure 5.1.E.II) and both K9ME1

(Figure 5.1.H.I and II) and K9ME2 (Figure 5.1.I and II) showed 2 structures containing α – helical content.

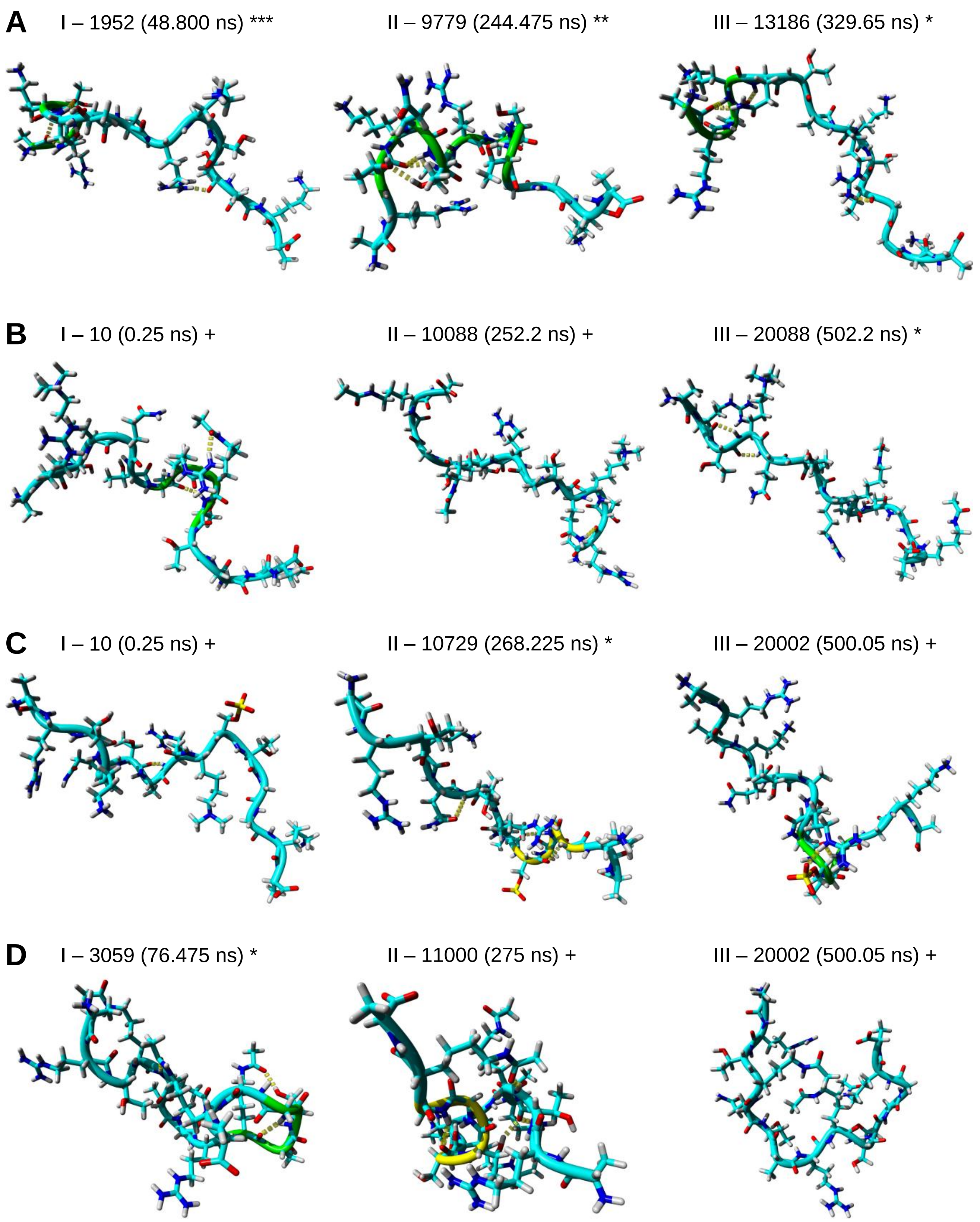


Figure 5.1 Structures obtained from clustering for the WT (A), ACTIVE (B), INACTIVE (C) and HYPER-ALY tail tips. The first numerical value indicates the trajectory frame number and the value in brackets indicates the corresponding time point in the trajectory. Structures obtained from clustering are indicated by an asterisk (*) and the rank of the cluster is indicated by the number of asterisks, thus * represents the top cluster. Chosen structures are labeled with a plus (+).

Molecular graphics created with YASARA (www.yasara.org) and POVRay (www.povray.org)

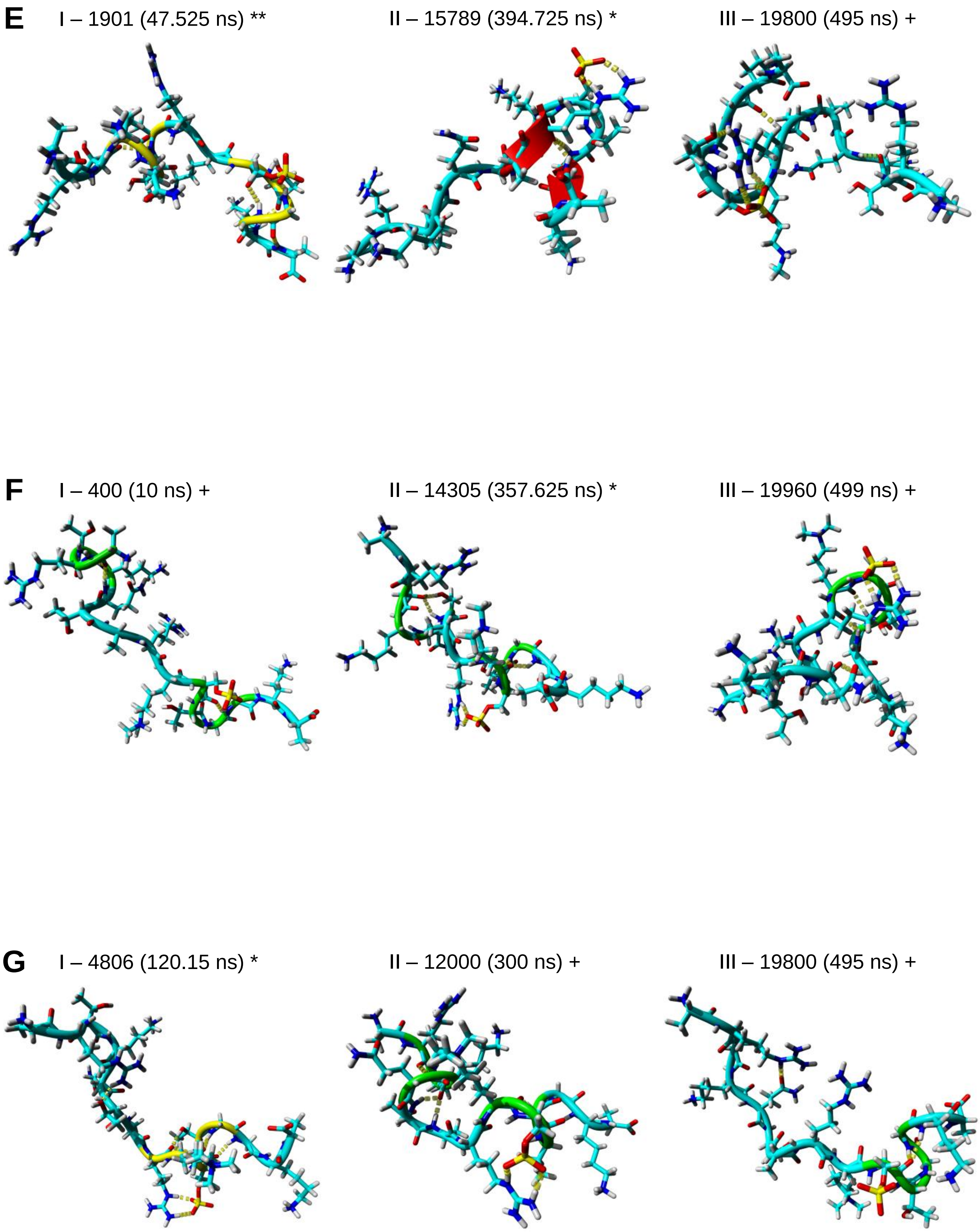


Figure 5.1 (cont.) Structures obtained from clustering for the K9ME1_S10PHO (E), K9ME2_S10PHO (F) and K9ME3_S10PHO (G) tail tips. The first numerical value indicates the trajectory frame number and the value in brackets indicates the corresponding time point in the trajectory. Structures obtained from clustering are indicated by an asterisk (*) and the rank of the cluster is indicated by the number of asterisks, thus * represents the top cluster. Chosen structures are labeled with a plus (+).

Molecular graphics created with YASARA (www.yasara.org) and POVRay (www.povray.org)

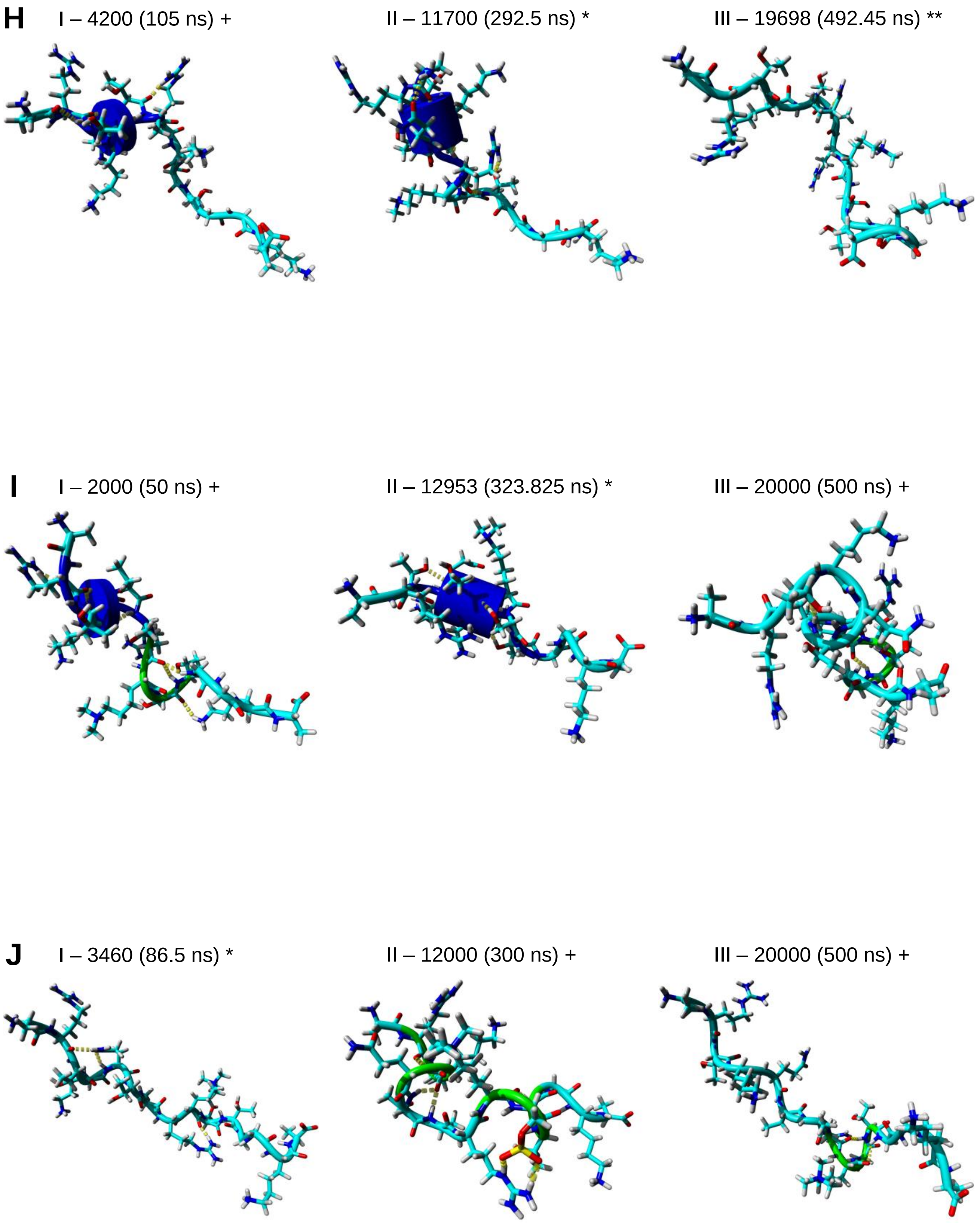


Figure 5.1 (cont.) Structures obtained from clustering for the K9ME1 (H), K9ME2 (I) and K9ME3 (J) tail tips. The first numerical value indicates the trajectory frame number and the value in brackets indicates the corresponding time point in the trajectory. Structures obtained from clustering are indicated by an asterisk (*) and the rank of the cluster is indicated by the number of asterisks, thus * represents the top cluster. Chosen structures are labeled with a plus (+).

Molecular graphics created with YASARA (www.yasara.org) and POVRay (www.povray.org)

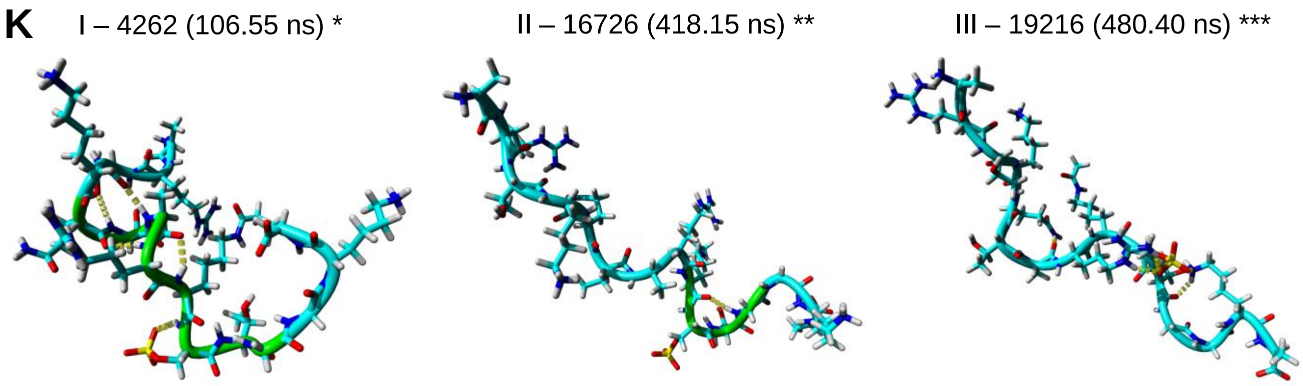


Figure 5.1 (cont.) Structures obtained from clustering for the K9ACE_S10PHO (K) tail tip. The first numerical value indicates the trajectory frame number and the value in brackets indicates the corresponding time point in the trajectory. Structures obtained from clustering are indicated by an asterisk (*) and the rank of the cluster is indicated by the number of asterisks, thus * represents the top cluster. Chosen structures are labeled with a plus (+).

Molecular graphics created with YASARA (www.yasara.org) and POVRay (www.povray.org)

5.3.2 Molecular docking of the N-terminal tip structures to the NCP

It is important to note again that in contrast to the stand-alone version of AutoDock and many other programs, YASARA reports its free energy values as positive instead of negative values. Thus when we talk about a high binding energy, that means that a high amount of energy was released upon binding, thus leaving the system at a lower total energy.

The indicated clustered structures (see Figure 5.1) were selected and docked to the nucleosome using the grid – based docking method, developed and used in Chapter 3. Using the trajectories from Chapter 4 also meant that we could observe the effect of the post – translational modifications present in the structures, on the binding (if any existed) of the H3 tail to the nucleosome. The majority of the structures yielded low to negative binding energy values (see Table 5.2). However the WT, K9ME1, K9ME2 and the K9ME3 tip showed binding above 8.00 kcal/mol, and was thus selected as the top ranking docking poses. Interestingly, these structures were the only structures that did not contain acetylated lysine or phosphorylated serine residues. In Chapter 3 it was observed that specific binding of the test peptide to the nucleosome surface produced binding energies higher than 8 kcal/mol, while non – specific binding to the DNA produced binding energy values in the region between 5 and 8 kcal/mol. We therefore decided to use 8 kcal/mol as a minimum cut-off value for the binding energy of poses that exhibit specific binding to the nucleosome. The top ranked docking poses consisted of 11 poses (see Table 5.3), with the 3 structures from the WT tip, 2 structures from the K9ME3 tip, and 1 structure each from the K9ME2 and K9ME1 tips. The one structure from the K9ME1 tip produced only one pose, which was the pose with the highest binding energy of 11.95 kcal/mol. The top ranking docking poses were all found in the cells A3, B3 and C3 of the grid (Figure 5.2.A), and all the poses were found to be horizontally sandwiched between the DNA gyres (Figure 5.2.B), making contact with both the nucleosome and the DNA. Vertically the poses were located between the exit points of the H3 and H2B N-terminal tails. The other structures showed extremely low energy docked poses, and we thus decided to superimpose the best docked poses in all the B3 and C3 cells onto the top ranked poses (see Figure 5.3). Surprisingly, there was visible overlap between the binding sites of the top ranked poses and the other poses. The only reason for these lower binding energy values could be the different modifications present on the other structures.

Table 5.2 Binding energy (kcal/mol) of the top docked poses obtained during rigid grid docking. Highlighted cells indicate top poses across all poses.

Structure	Binding Energy (kcal/mol) in grid cells								
	A1	**A2**	**A3**	**B1**	**B2**	**B3**	**C1**	**C2**	**C3**
WT_1952	3.32	3.34	4.8	2.7	3.93	7.74	5	4.99	8.49
WT_9779	3.97	4.63	7.02	3.38	3.88	8.1	5.62	4.41	6.78
WT_13186	4.35	3.86	9.33	3.81	4.05	9.27	4.36	3.75	6.91
ACTIVE_10	0.24	0.07	0.35	-0.93	-1.82	1.42	0.53	0.19	1.48
ACTIVE_10088	-0.66	0.32	1.42	0.07	1.37	1.77	-0.05	-0.57	0.52
ACTIVE_20088	0.89	0.97	1.32	-1.43	-0.79	1.93	-0.26	-0.12	1.54
INACTIVE_10	-0.49	-0.73	-0.01	-0.88	-1.18	0.33	0.21	0.3	1.53
INACTIVE_10729	-0.7	-0.55	0.75	-0.85	-1	0.69	1.48	-0.81	-0.2
INACTIVE_20002	0.14	1.76	0.68	-0.74	-0.65	0.61	0.44	0.64	3.63
HYPER_ALY_3059	-2.55	-2.21	-0.24	-3.51	-3.88	-1.02	-2.8	-3.28	0.28
HYPER_ALY_11000	-2.86	-3.08	-2.44	-3.53	-2.5	0.77	-3.08	-2.58	-2.41
HYPER_ALY_20002	-3.13	-2.75	-1.65	-2.95	-3.02	-1.99	-3.48	-3.57	-1.08
K9ME1_S10PHO_1901	0.76	0.82	1.47	0.3	-1.18	4.04	0.21	-0.57	2.38
K9ME1_S10PHO_15789	0.32	1.23	1.23	-0.59	0.2	1.27	1.24	1.31	1.62
K9ME1_S10PHO_19800	-1.1	-1.13	0.41	-1.1	-0.99	1.5	-0.44	-1.14	2.57
K9ME2_S10PHO_400	-1.13	-0.35	1.35	-2.33	-0.06	2.52	2.81	-1.4	3.81
K9ME2_S10PHO_14305	-0.67	0.2	-0.47	-1.05	-2.08	2.04	-0.64	-0.64	2.01
K9ME2_S10PHO_19960	-0.07	0	3.28	0.87	-1.24	3.31	0.91	0	0.72
K9ME3_S10PHO_4806	-0.98	0.13	-0.32	-1.14	-1.19	2.65	0.14	-1.14	0.67
K9ME3_S10PHO_12000	-1.03	-0.14	0.37	-1.89	-1.48	1.93	1.99	-1.43	1.9
K9ME3_S10PHO_19800	-0.08	-0.9	3.32	-1.59	-0.5	2.47	0.26	-0.97	0.22
K9ME1_4200	4.35	4.87	6.31	4.45	3.65	11.95	5.64	5.76	5.3
K9ME1_11700	3.38	4.49	7.39	2.65	3.94	7.9	4.88	3.25	7.16
K9ME1_19698	6.23	4.12	5.4	4.07	3.32	3.95	4.81	3.11	7.37
K9ME2_2000	5.58	5.87	6.71	2.14	3.41	8.58	6.43	6.41	8.01
K9ME2_12953	4.56	4.55	4.61	4.74	2.68	6.05	5.3	5.16	4.33
K9ME2_20000	3.55	4.19	6.48	3.52	2.37	7.72	4.21	3.78	5.78
K9ME3_3460	4.95	3.55	9.28	3.55	3.07	9.13	6.15	3.26	9.28
K9ME3_12000	3.46	2.9	0.37	4.96	3.3	7.16	4.83	3.65	5.78
K9ME3_20000	3.93	4.03	5.16	2.6	2.33	9.79	7.58	4.15	6.37

Table 5.2 (cont.) Binding energy (kcal/mol) of the top docked poses obtained during rigid grid docking. Highlighted cells indicate top poses across all poses.

Structure	**Binding Energy (kcal/mol) in grid cells**								
	A1	**A2**	**A3**	**B1**	**B2**	**B3**	**C1**	**C2**	**C3**
K9ACE_S10PHO_4262	-1.77	-2.01	-1.61	-2.25	-3.71	-2.41	-0.77	-0.09	-0.45
K9ACE_S10PHO_16726	-1.07	-1.07	-1.15	-0.8	-2.78	-0.31	-0.1	-2.18	0.67
K9ACE_S10PHO_19216	-2.23	-2.25	-1.18	-2.4	-3.19	-1.98	-2.55	-2.54	-0.75

Table 5.3 The best docked poses in terms of binding energy, arranged from the highest to the lowest energy

	System and Position	**Binding Energy (kcal/mol)**
A	K9ME1_4200_B3	11.95
B	K9ME3_20000_B3	9.79
C	WT_13186_A3	9.33
D	K9ME3_3460_A3	9.28
E	K9ME3_3460_C3	9.28
F	WT_13186_B3	9.27
G	K9ME3_3460_B3	9.13
H	K9ME2_2000_B3	8.58
I	WT_1952_C3	8.49
J	WT_9779_B3	8.1
K	K9ME2_2000_C3	8.01

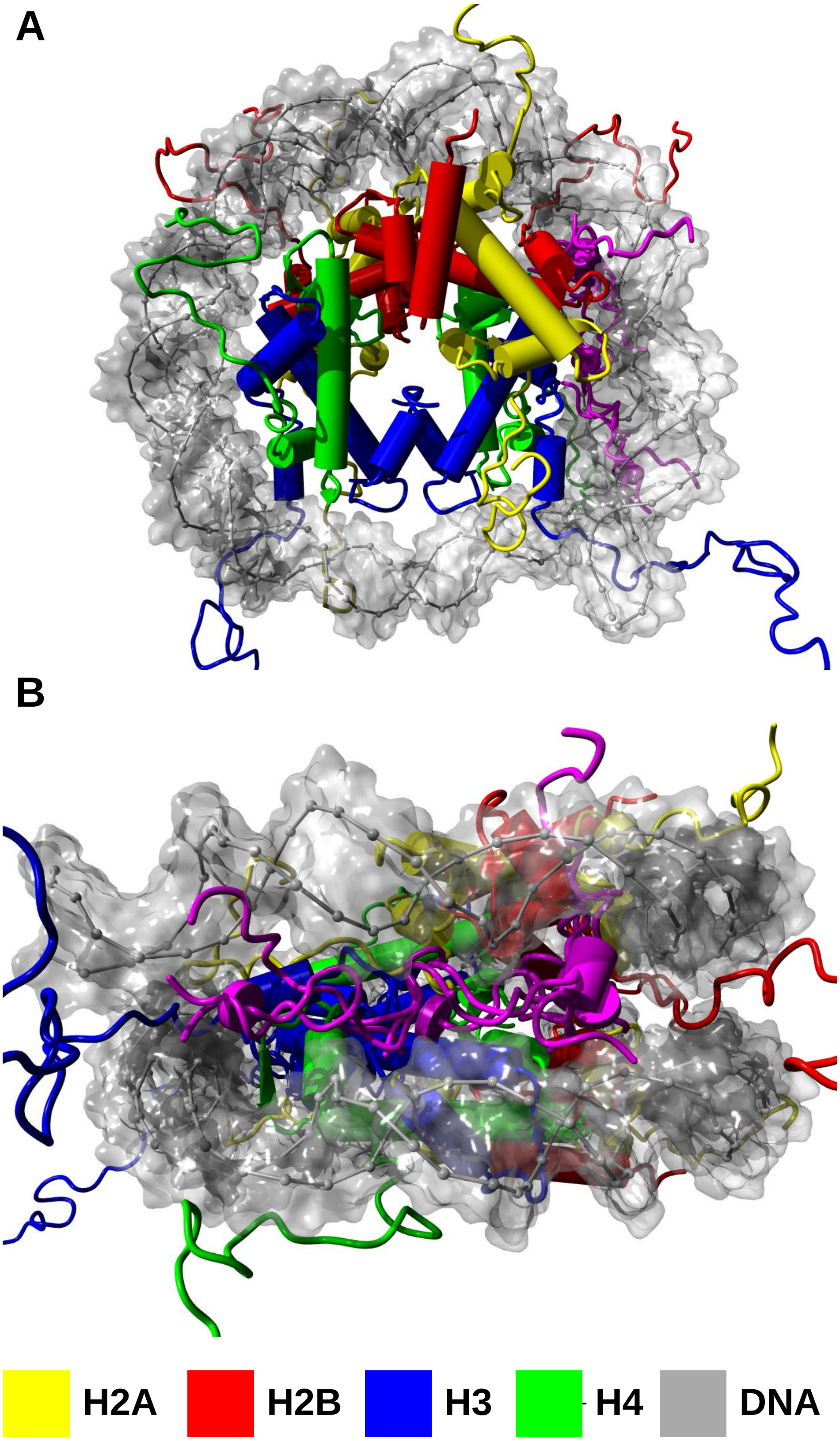


Figure 5.2 The best docked poses of the H3 N-terminal tip on the nucleosome (as shown in Table 5.3). Two views are presented: The view from the top of the nucleosome (A) and the view from the side of the nucleosome (B). Docked poses are shown in magenta.

Molecular graphics created with YASARA (www.yasara.org) and POVRay (www.povray.org)

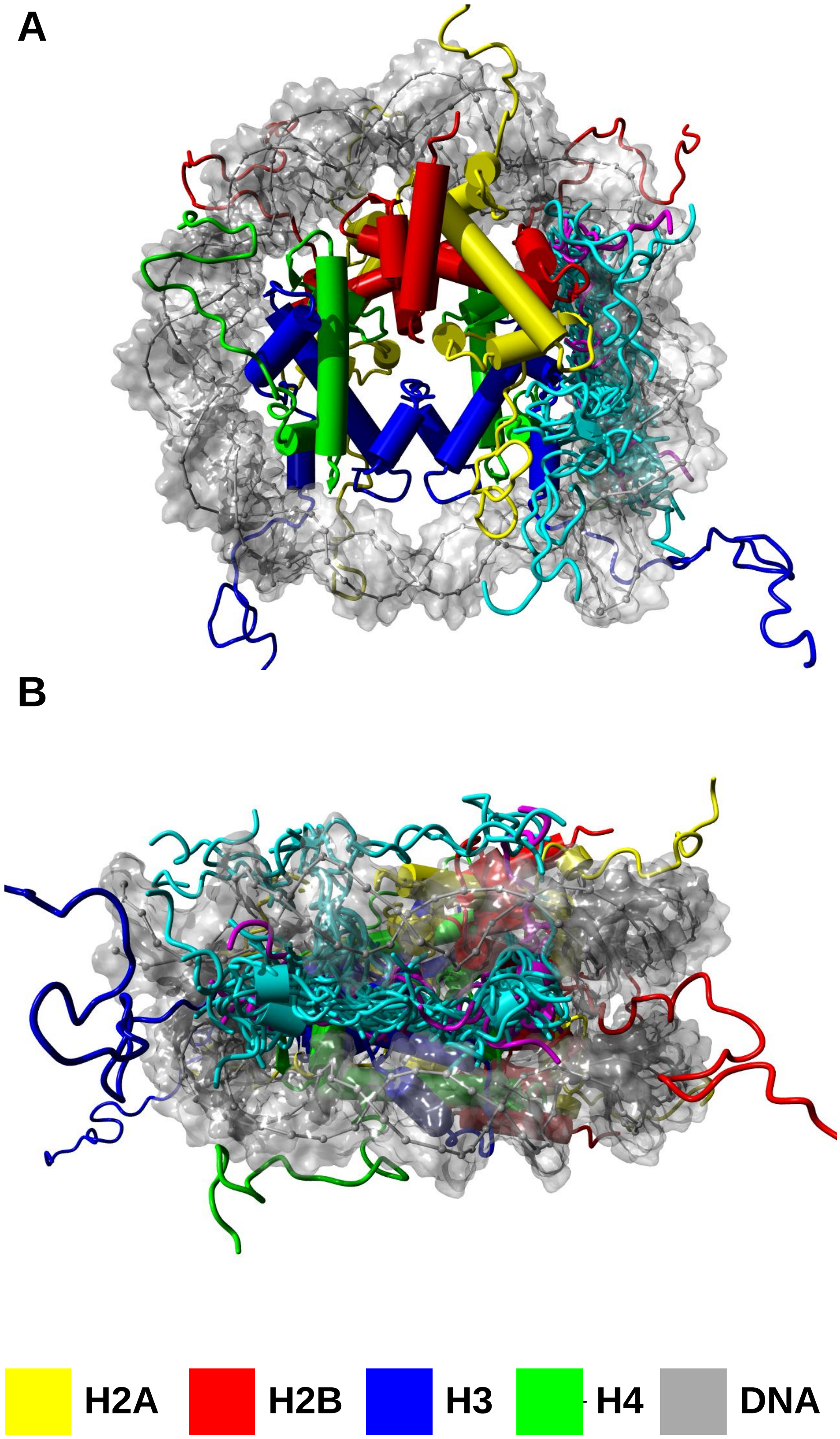


Figure 5.3 The best poses of the B3 and C3 cells for the rest of the tail structures superimposed onto the best docked poses of the H3 N-terminal tip on the nucleosome (as shown in Table 5.2). Two views are presented: The view from the top of the nucleosome (A) and the view from the side of the nucleosome (B). Top ranked docked poses are shown in magenta and the other docked poses are shown in cyan.

Molecular graphics created with YASARA (www.yasara.org) and POVRay (www.povray.org)

5.3.3 Contact of the histone H3 N-terminal tail with the nucleosome

In the molecular docking of the H3 tail we found that the tail was docked between the DNA and protein octamer of the nucleosome. We wanted more details on what contacts the tail was making to the DNA and the protein octamer and whether some specific contacts involved the modified residues. We first checked the hydrogen bonding between the top docking poses and the nucleosome using YASARA and a python script. The hydrogen bonds are listed in Table 5.4. Interestingly, in the top docking pose, the tail was only involved in 3 hydrogen bonds and all of these were to the DNA. The side chain of K9 was also involved in one of these hydrogen bonds. This trend was observed for most of the docking poses, with the majority of the hydrogen bonding occurring between the DNA, mostly acting as hydrogen bond acceptors, and the residues of the H3 tail acting as hydrogen bond donors. The side chains of K4, K9 and K14 were constantly involved in hydrogen bonds, while the side chain of S10 was also involved in some hydrogen bonds. The side chains of R2 and R8 was also noticeably involved in hydrogen bonding with the DNA.

To investigate other types of contacts, we performed a contact analysis with YASARA to obtain the amount of contacts that the H3 tail made with each of the molecules within the nucleosome within a distance of 4 Å. Figure 5.4 shows the number of contacts between the DNA and protein for each top docked pose. In all but one pose, the H3 tail made significantly more contacts with the DNA compared to the octamer. This observation is in line with a more traditional view of the H3 tail, i.e. where the H3 tail is an unstructured basic polypeptide that binds the DNA [4, 7]. We thus investigated the basic residues in the top docked pose and the elements in their immediate environment. Figure 5.5 shows a close up view of R2, K4, K9 and K14 and their immediate environment. R8 was not close to the DNA and was not included in this figure. It is clearly seen that the side chain of these residues were surrounded by the negatively charged phosphate groups of the DNA. Subsequently, we speculated that the docked structure might be structurally related to DNA binding domain of a protein in the PDB. To investigate this I searched the PDB database using the protein structure comparison service "Fold" at the European Bioinformatics Institute (http://www.ebi.ac.uk/msd-srv/ssm), authored by E. Krissinel and K. Henrick [8-12], with each the top docked poses used as separate queries. However, no hits were obtained with any DNA binding protein or domain in the PDB database (queried in January 2012).

Table 5.4 Hydrogen bonds formed between the top ligands and the NCP. The histones and DNA contacted are indicated in brackets after the receptor chain. The type refers to the role of the ligand residue in the hydrogen bond.

Pose	Ligand Residue	Atom	Receptor Residue	Atom	Receptor Chain	Type
A	GLN 5	N	DC -25	O2P	I	Donor
	M1L 9	NZ	DA -55	O2P	J	Donor
	GLY 13	N	DG -56	O1P	J	Donor
B	THR 3	N	DT 20	O2P	J	Donor
	LYS 4	NZ	DG -56	N7	J	Donor
	GLN 5	N	DG -56	O2P	J	Donor
	LYS 14	NZ	DT -54	O2P	J	Donor
C	ALA 1	N	DC -24	O1P	I	Donor
	THR 3	N	DC -25	O2P	I	Donor
	LYS 4	NZ	DG -56	O2P	J	Donor
	LYS 14	NZ	DG 49	O2P	I	Donor
D	SER 10	OG	PRO 47	O	H	Donor
	LYS 14	NZ	DC -52	O1P	J	Donor
E	ALA 1	N	DT -68	O2P	J	Donor
	ARG 2	NE	DT -68	O5*	J	Donor
	ARG 2	NH2	DT -68	O1P	J	Donor
	THR 11	OG1	LEU 60	O	A	Donor
F	ALA 1	N	DC -24	O1P	I	Donor
	THR 3	N	DC -25	O2P	I	Donor
	LYS 4	NZ	DG -56	O2P	J	Donor
	LYS 14	NZ	DG 49	O2P	I	Donor

Table 5.4 (cont.) Hydrogen bonds formed between the top ligands and the NCP. The histones and DNA contacted is indicated in brackets after the receptor chain. The type refers to the role of the ligand residue in the hydrogen bond.

Pose	Ligand Residue	Atom	Receptor Residue	Atom	Receptor Chain	Type
G	SER 10	OG	PRO 47	O	H	Donor
H	ALA 1	N	DC -14	O1P	I	Donor
	M2L 9	N	DG -56	O1P	J	Donor
	ALA 15	N	GLY 102	OT2	B	Donor
	ALA 15	OT1	GLN 85	NE2	A	Acceptor
I	ARG 2	NH1	DC -58	O4*	J	Donor
	THR 3	N	DG 61	O1P	I	Donor
	SER 10	OG	DA -67	O2P	J	Donor
	LYS 14	NZ	DT 67	O4	I	Donor
J	ALA 1	N	DT -26	O2P	I	Donor
	ARG 2	NE	DC -25	O2P	I	Donor
	ARG 2	NH2	DC -25	O2P	I	Donor
	LYS 4	NZ	DG 21	N7	J	Donor
	LYS 4	NZ	DG 21	O6	J	Donor
	ARG 8	NH1	DG -56	O2P	J	Donor
	LYS 9	NZ	GLY 99	O	B	Donor
	LYS 14	NZ	LYS 31	O	H	Donor
	LYS 14	NZ	DG 49	O5*	I	Donor
K	ARG 2	NE	DT -68	O2P	J	Donor
	ARG 2	NH2	DT -68	O2P	J	Donor
	LYS 4	N	DA -67	O2P	J	Donor
	LYS 4	NZ	DT -66	O4	J	Donor
	ARG 8	NH1	DT 63	O2P	I	Donor
	GLY 13	N	DC -14	O1P	I	Donor
	ALA 15	OT1	LYS 64	N	A	Acceptor

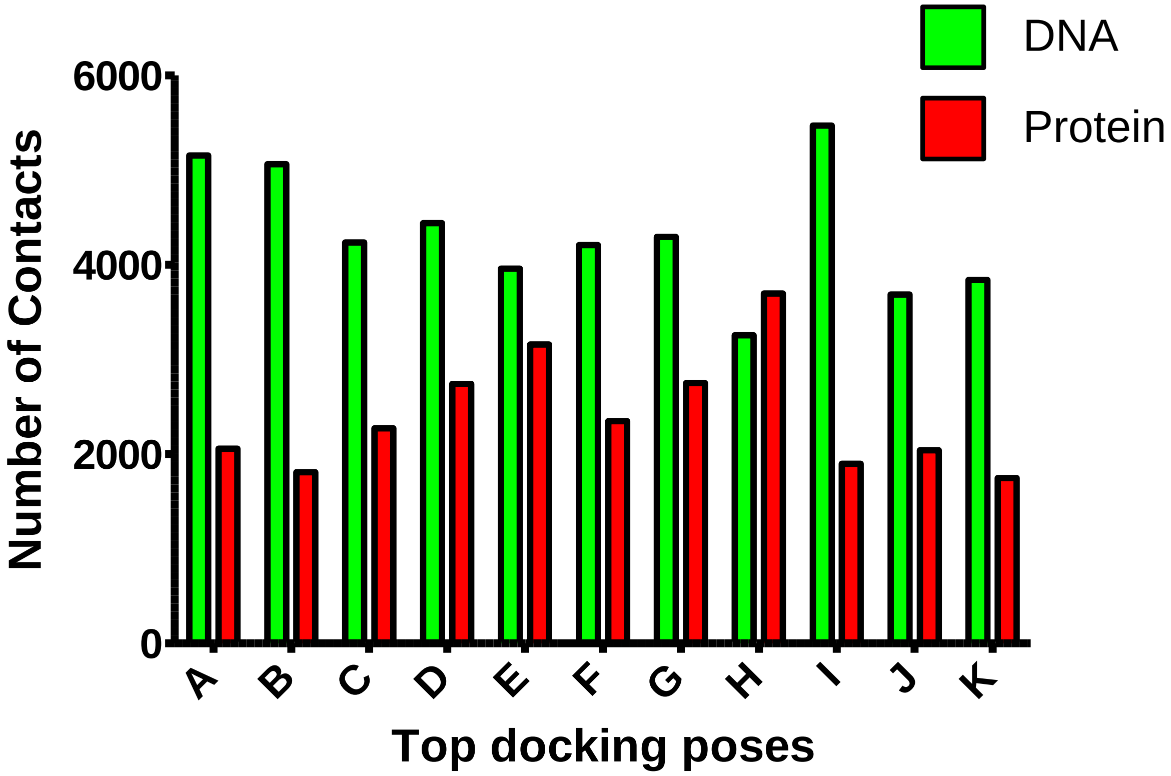


Figure 5.4 Total number of contacts (proximal atoms within 4Å) between the top docking poses (see Table 5.3) of the H3 tail and the nucleosome.

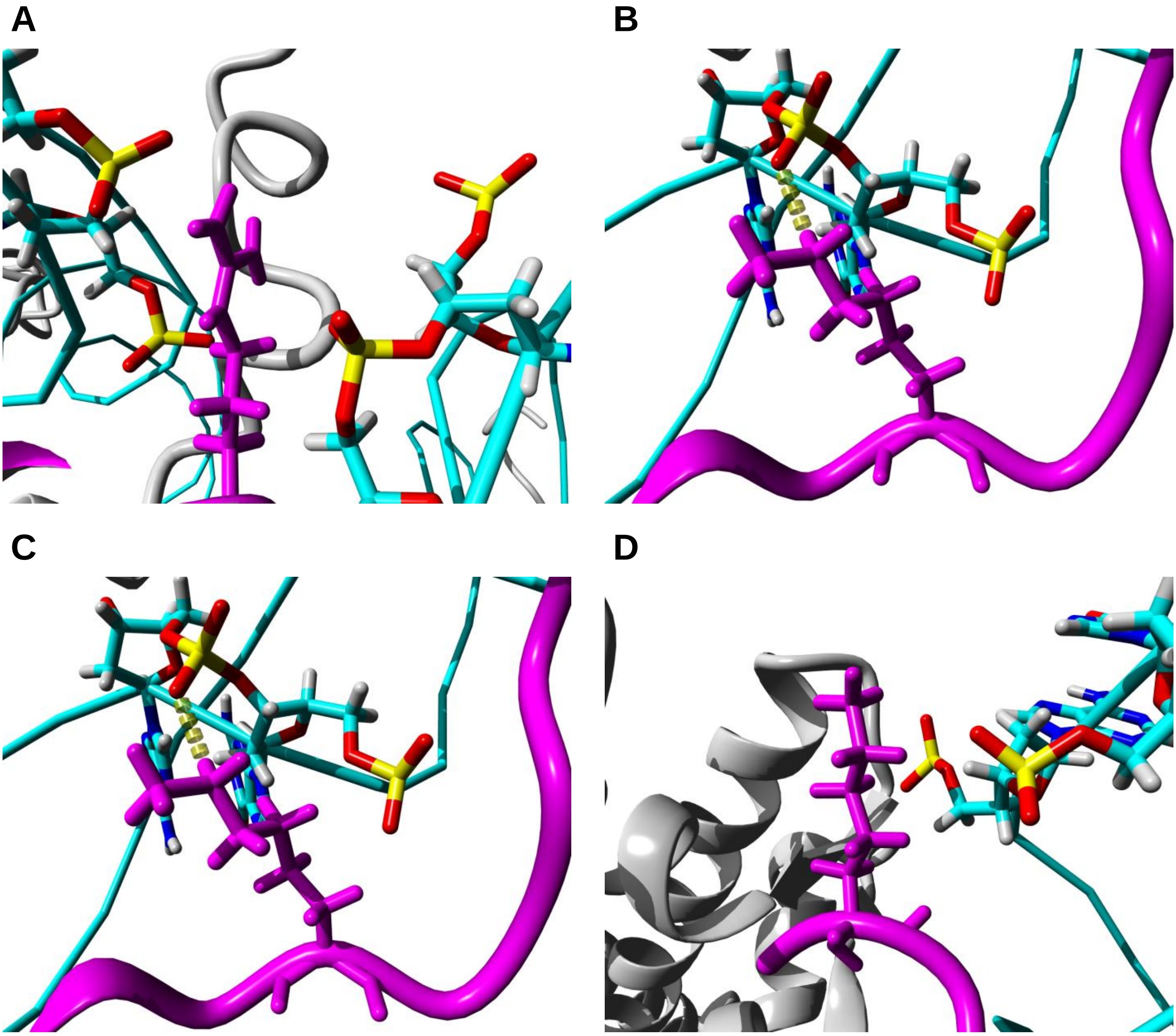


Figure 5.5 Interaction of the basic residues of the top docking pose with the nucleosomal DNA. R2 is shown in A, K4 is shown in B, K9 is shown in C and K14 in D. The ligand is colored in magenta and the DNA is colored according to the atomic elements, where carbon is cyan, oxygen is red, nitrogen is blue and phosphor is yellow.

Molecular graphics created with YASARA (www.yasara.org) and POVRay (www.povray.org)

5.4 DISCUSSION

5.4.1 Grid – based docking – a new idea using existing techniques

In this chapter we implemented of a novel technique to dock structures to a large receptor without losing accuracy while at the same time increasing the density of coverage of the receptor by using an existing program, AutoDock[1]. By splitting a large docking problem into a set of overlapping smaller docking problems, we were able to identify a binding "hotspot" for the 15 – residue H3 N-terminal tail on the nucleosome. To our knowledge this is the first time such an approach has been implemented in solving a molecular docking problem.

5.4.2 The H3 tail binds between the DNA and the side of the octamer

In chapter 3 we used rigid, grid - based docking to confirm the binding site of KSHV LANA to the acidic patch on the nucleosome surface, as was found in an elucidated crystal structure[5]. Because the 15 – residue H3 N-terminal tail shared the basicity characteristics of the KSHV LANA peptide, and was long enough to reach the binding site, the general expectation was that the H3 tail would also bind to the acidic patch. However, using rigid, grid – based docking we showed that energetically the most favorable binding position of the H3 N-terminal tail was on the lateral surface between the exit point of the H3 tail and the H2B tail, sandwiched between the two DNA gyres (see Figure 5.2).

Although we did not expect this specific binding position, it is actually well supported by experimental studies. Cary *et al.*[3] showed in NMR studies that the histone tails were bound to the nucleosome at NaCl concentrations below 2 M, and dissociated from the core particle at concentrations above 2 M. This suggested that the main interaction between the tails and the nucleosome was electrostatic in nature. Later Toland and co – workers[13] showed through mechanical pulling experiments with tailless H3/H4 nucleosomes that the entering and exiting DNA of the nucleosome were less tightly bound, being released at a lower pulling force than with the H3/H4 nucleosomes. Given that the H3 tail originates between the two DNA gyres where the linker DNA enters and exits the nucleosome [4, 7], and that the H4 tail was seen to bind to the acidic patch of a nucleosome in a co – crystal, it is likely that the H3 tail is the largest contributor to this missing force, and thus bound to the entering and exiting linker DNA. This was reinforced by the observation that the N-terminal tails of H3 and H4 blocked access to the DNA – binding, anticancer drugs mithramycin and daunomycin, shown by tail - deletion studies [14]. Finally, La Penna and co – workers[15] observed in short 25 ns MD simulations that a 25 – residue H3 N-terminal tail could bind with a 10 – bp DNA fragment in two ways. Firstly, with the tail docked perpendicularly to the DNA axis, into the major groove of the DNA, and also length wise with the DNA. We observed the second conformation in our docking experiments.

Examining the hydrogen bonding within the top ranked docking poses (see Table 5.4); we showed that the majority of the hydrogen bonding was made to the DNA. Furthermore, the majority of the residues involved in these hydrogen bonds acted as hydrogen bond donors. However, the number of hydrogen bonds was not a determinant of higher binding energies, although two out of the three hydrogen bonds found in the highest ranked structure were not found in any of the other poses. One of these included the side chain of the mono – methylated K9, indicating that this residue and its modification state might be critical to binding. We also performed a contact analysis, and, again, the majority of the top docked poses were seen to make significantly more contacts with the DNA compared to the histone proteins (see Figure 5.4). This was an intriguing finding, and we investigated the structure of the top docked structure in greater detail. We observed that all the basic residues in the H3 tail, except R8, were surrounded by DNA phosphate groups (see Figure 5.5). This confirmed that electrostatic interaction was mainly involved in the binding of the H3 tail to the nucleosome.

5.4.3 The effect of PTMs on the binding of the H3 tail to the nucleosome

Having sampled the docking structures from the MD trajectories, we also had a chance to look at differences in binding between differentially modified H3 tails. However, only the unmodified structures and the K9 methylated structures, without S10 phosphorylation, yielded binding energy values similar in magnitude to that observed for the specific binding of KSHV – LANA (see Chapter 3). The other structures yielded very low and some even negative binding energy values. We therefore decided to superimpose the rest of the tail structures' best poses in grid cells B3 and C3 onto the top docked poses (see Figure 5.5). It is seen in Figure 5.3 that there is a significant overlap between the binding positions of the low ranked structures and the binding positions of the top ranked docked poses. This is fascinating because the structures of the top ranked poses were different and although they were docked at roughly the same position, the sort of overlap seen with top poses for KSHV – LANA was, significantly, never observed. Yet, the binding energy values obtained were between 8 kcal/mol and 12 kcal/mol. Thus, the presence of the different modifications must have caused these low binding energy values. In the case of the tails with phosphorylated serine residues, the low binding energy is expected because the repulsion forces between the negatively charged phosphate group and the negatively charged DNA would be energetically unfavorable. In Chapter 4 it was shown that S10 in the inactive tail was hydrogen bonded to the side chain of R8 and R17, thus partially shielding the negative charge. Because the tails docked only included residue 1 -15 of the H3 N-terminal tail, the residue was not fully shielded, which could have led to low binding energy observed. The remaining structures all contained acetylated lysine residues in the structures. As mentioned before, if electrostatic interaction is the main contributor to the binding of the tail to the nucleosome, then eliminating the positive charges the lysine residues would be energetically undesirable. This is especially true in the case of K9, where it was shown to form a unique hydrogen bond with the DNA in the top ranked structure, which differed by almost 2 kcal/mol from the next best structure.

5.5 CONCLUSION

I have shown in this chapter that the 25 – residue H3 N-terminal tail could bind to the nucleosome by using a grid – based docking approach with AutoDock, implemented in YASARA, using rigid ligand structures that were derived from MD trajectories. We identified a single docking position for the H3 tail between the exit points of the H3 and H2B N-terminal tails, with the tail docked laterally between the nucleosome and the two DNA gyres. It was also shown that the tail made contacts predominantly with the DNA, leading us to conclude that an electrostatic interaction was the main contributor to binding. By using differently modified tails in the docking, we observed that specific modifications such as serine phosphorylation and lysine acetylation yielded similar docking positions, but with significantly lower binding energies compared to the unmodified tail.

5.6 REFERENCES

1. Morris, G. M., Huey, R., Lindstrom, W., Sanner, M. F., Belew, R. K., Goodsell, D. S. & Olson, A. J. (2009). AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *Journal of Computational Chemistry* **30**, 2785-2791.
2. Krieger, E., Koraimann, G. & Vriend, G. (2002). Increasing the precision of comparative models with YASARA NOVA – a self-parameterizing force field. *Proteins* **47**, 393-402.
3. Cary, P. D., Moss, T. & Bradbury, E. M. (1978). High-Resolution Proton-Magnetic-Resonance Studies of Chromatin Core Particles. *European Journal of Biochemistry* **89**, 475-482.
4. Luger, K., Mader, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. (1997). Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* **389**, 251-260.
5. Barbera, A. J., Chodaparambil, J. V., Kelley-Clarke, B., Joukov, V., Walter, J. C., Luger, K. & Kaye, K. M. (2006). The nucleosomal surface as a docking station for Kaposi's sarcoma herpesvirus LANA. *Science* **311**, 856-861.
6. Hess, B., Kutzner, C., van der Spoel, D. & Lindahl, E. (2008). GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory Comput.* **4**, 435-447.
7. Davey, C. A., Sargent, D. F., Luger, K., Maeder, A. W. & Richmond, T. J. (2002). Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 Å resolution. *J. Mol. Biol.* **319**, 1097-1113.

8. Krissinel, E. & Henrick, K. (2004). Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallographica Section D* **60**, 2256-2268.
9. Krissinel, E. & Henrick, K. (2005). Multiple Alignment of Protein Structures in Three Dimensions Computational Life Sciences. (Berthold, R., Glen, R., Diederichs, K., Kohlbacher, O. & Fischer, I., eds), pp. 67-78, Springer Berlin / Heidelberg.
10. Krissinel, E. B. & Henrick, K. (2004). Common subgraph isomorphism detection by backtracking search. *Softw: Pract. Exper.* **34**, 591-607.
11. Krissinel, E. B., Winn, M. D., Ballard, C. C., Ashton, A. W., Patel, P., Potterton, E. A., McNicholas, S. J., Cowtan, K. D. & Emsley, P. (2004). The new CCP4 Coordinate Library as a toolkit for the design of coordinate-related applications in protein crystallography. *Acta Crystallographica Section D* **60**, 2250-2255.
12. Krissinel, E. (2007). On the relationship between sequence and structure similarities in proteomics. *Bioinformatics* **23**, 717-723.
13. Brower-Toland, B., Wacker, D. A., Fulbright, R. M., Lis, J. T., Kraus, W. L. & Wang, M. D. (2005). Specific Contributions of Histone Tails and their Acetylation to the Mechanical Stability of Nucleosomes. *Journal of Molecular Biology* **346**, 135-146.
14. Mir, M., Das, S. & Dasgupta, D. (2004). N-terminal tail domains of core histones in nucleosome block the access of anticancer drugs, mithramycin and daunomycin, to the nucleosomal DNA. *Biophysical Chemistry* **109**, 121-135.
15. LaPenna, G., Furlan, S. & Perico, A. (2006). Modeling H3 histone N-terminal tail and linker DNA interactions. *Biopolymers* **83**, 135-147.

CHAPTER 6

General Discussion and Conclusion

6.1 DISCUSSION

The histone tails have for decades been regarded as unstructured polypeptide chains[1, 2]. More recently they have been shown to act as molecular beacons to protein effectors which modify chromatin[3]. Some CD experiments have, however, shown that there was a structural contribution from the histone tails to CD spectra [4, 5], and elementary MD studies have also shown some structural features in the tails [6-9]. In this study we have expanded on the current structural knowledge available for the histone H3 N-terminal tail with the use of MD and molecular docking. How is this new structural information interpreted in terms of the biology and current experimental information?

Early studies on the histone tails and the role of PTMs were done with hyper acetylated nucleosome cores *in vitro*[10-12]. Wallace and co – workers found that hyper – acetylation of rat liver chromatin with acetic anhydride increased the accessibility of the DNA to both DNase I and staphylococcal nuclease. DNA hydrolysis is generally inhibited for both these enzymes due to the close association of the DNA duplex with the octamer surface in the nucleosome. Thermal denaturation was also performed, and showed that the pre-melting curve of the nucleosome was shifted to a lower temperature in the acetylated samples. Yau and co – workers corroborated this observation in their thermal denaturation study[13]. The source of the pre-melting curve was shown to be the entering and exiting linker DNA, which was not strongly associated with the histone octamer. Because the H3 and the H4 tails constitute the majority of the lysine acetylation sites, this suggested that the H3 and H4 tails were associated with the entering and exiting DNA. Based on the position of the H3 tail [2, 14], and the observation that the H4 tail contacted the acidic patch of the nucleosome in co – crystals[2] and because H3 contains more potential acetylation sites, it seemed likely that the H3 tail marked the majority of the interactions with the entering and exiting DNA. Results from other experimental techniques reached the same conclusion. It was shown in mechanical pulling experiments that the force required to pull the inter turn of DNA (i.e. the entering and exiting linker DNA) from the nucleosome was significantly reduced when the tails were hyper acetylated[15]. Single – Pair Förster Resonance Energy (spFRET) experiments showed that the point to point distance of the ends of the two entering/exiting DNA was increased with acetylation, indicating that the DNA ends were more mobile[16]. Norton and co-workers also showed a decrease in the linking number difference per nucleosome in reconstituted mini - chromosomes upon acetylation[12]. Wang and Hayes directly implicated that the H3 tail was involved contacts with the entering/exiting linker DNA and that acetylation weakened these interaction through the use of biochemical assays with glutamine substitutions to mimic acetylation in model oligonucleosomes [17].

We observed in our docking experiments that the H3 tail was bound in close proximity to the exit point of the H3 tail from the nucleosome and where the linker DNA enters/exits the nucleosome. The tail was sandwiched between the DNA and the lateral side of the octamer and contact analysis showed that the majority of the contacts were made with the DNA. This binding position was also found with the other modifications, however with the hyper – acetylated tail, the binding energy was significantly reduced. This supported the view that the contacts between the H3 tail and the entering and exiting linker DNA (due to the location of the binding site) were disrupted with acetylation. When looking at the representative structure of the most populated cluster of the HYPER-ALY tail (Figure 4.28.D.i), the structure is shown to retract somewhat into a bulge, with the N-terminal tip closely associated with the bulge. We thus propose that the non – acetylated H3 tail is folded over the entering and exiting linker DNA and bound as shown in our docking experiments, essentially keeping the linker DNA immobile and potentially keeping the angle of entry/exit within an acute rage. When the tail is hyper – acetylated, the contacts with the DNA are abolished, and the tail retracts into the bulge, allowing the DNA to become mobile, thus changing the angle of the entry/exit of the DNA and increasing the length of the linker DNA. Wong and co –

workers showed in an all atom model of the chromatin fiber that the change of the angle of the entering/exiting linker DNA and the length of the linker DNA between nucleosomes could produce different chromatin structures. Thus, our results, taken together with available experimental results, gives a plausible model of how particularly acetylation can lead to a more open and accessible chromatin structure. Interestingly, the active tail exhibited a similar structure (Figure 4.28.B.i), but in this case the N-terminus was released from the bulge structure. This was due to a hydrogen bond pair between K4 and A31 that was present in the HYPER-ALY tail (Figure 4.20.B), and which occurred between 5000 and 10000 times, which was absent in the ACTIVE TAIL (Figure 4.18.B). Although this hydrogen bond involved the backbone amide of K4, the tri – methylation of K4 in the ACTIVE tail could be preventing the ACTIVE tail from making this hydrogen bond. Also, only K9 and K14 were acetylated in the ACTIVE structure, which could lead one to speculate that only certain lysines in the H3 tail should be acetylated to promote a more open chromatin structure found in regions of active transcription. The inactive tail was also bound to the nucleosome with very low binding energy. However, as seen in Figure 4.29.B, the side chain of R17 hydrogen bonded with two of the three phosphor oxygens. Thus, a potential screening effect might have been lost due to the use of only 15 residue of the N-terminal H3 tail in the docking study. This absence of this screening effect would cause the H3 tail to be repelled by the electronegative DNA, which would yield the low binding energy.

The unique structure stabilized in the INACTIVE tail simulations could be preferable to the binding of the entering/exiting linker DNA. In a MD study of the histone H4 tail by Yang and Arya[18], the authors made the observation that an α – helix stably orientated all the basic residues, occurring at multiples of approximately 4 residues, in a single direction direction. In light of the electrostatic interaction between the tail and the DNA, this might serve as a mechanism which would allow the maximum number of basic residues to contact the DNA, thus making the interaction stronger. This would explain why there was an increase in α – helical content in tail modifications patterns associated with compact chromatin, i.e. transcriptionally inactive chromatin, and a general loss of α – helical content in modifications associated with transcriptionally active chromatin.

6.2 CONCLUSION

We have derived a model from MD and molecular docking results that provides insight into the process of chromatin compaction at the level of the nucleosome [19]. Our model accurately accounts for the effect of PTMs in the dynamic and regulated transition of chromatin from an inaccessible, transcriptionally silent state, to an accessible and transcriptionally active state. This model is summarized in Figure 6.1.

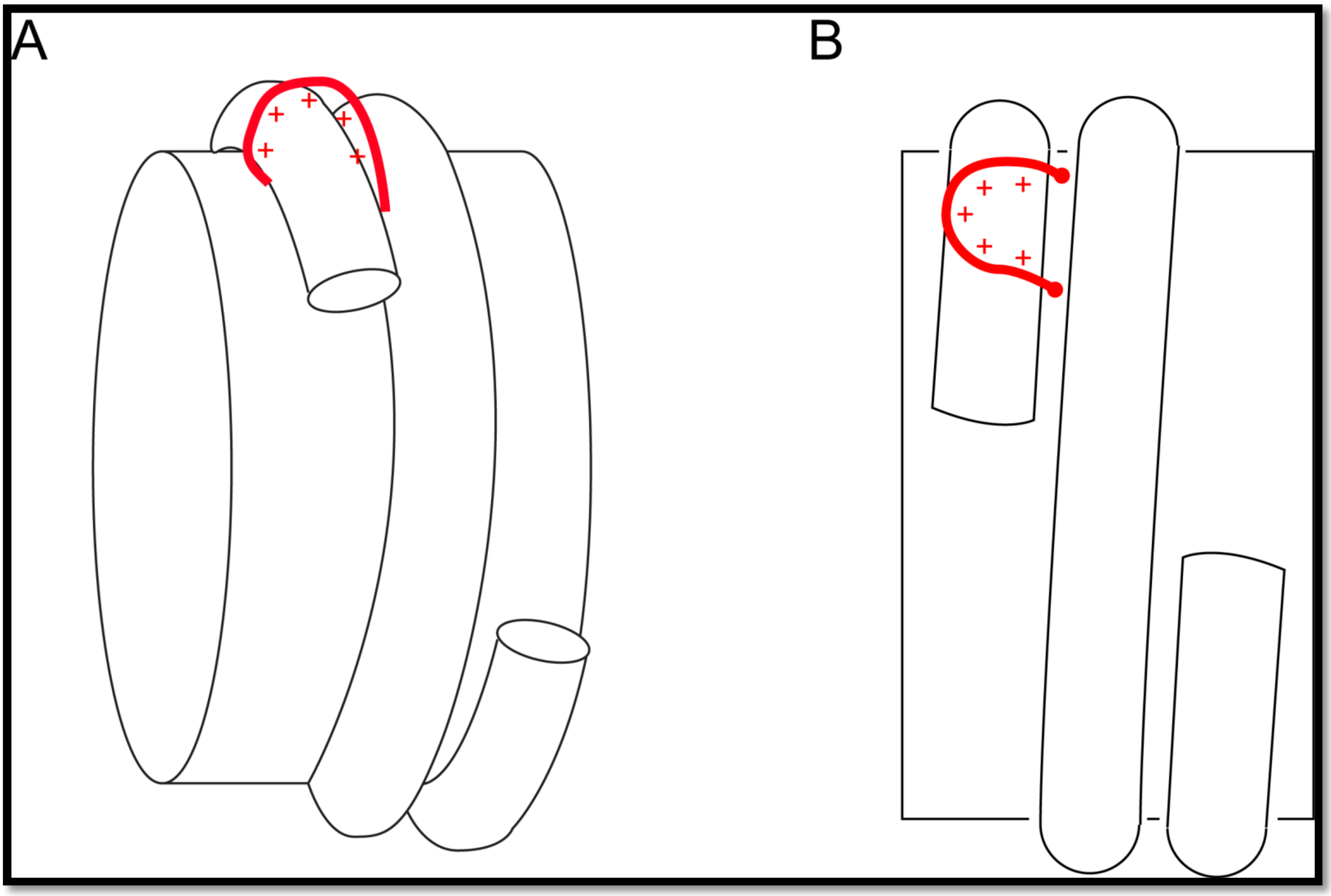


Figure 6.1 Derived model of binding of the histone H3 N-terminal tail in the nucleosome. A shows the positively charged tail bound over the negative DNA and contacting the side of the nucleosome. B shows a side - on view.

Future work would include going into the laboratory and attempting to physically verify the model *in vitro* or preferably *in vivo*, with chemical cross-linking coupled to mass spectrometry. This study presents a fresh perspective to our attempts to describe the molecular mechanisms whereby eukaryotic organisms produce and maintain the wonderful and exquisite diversity of life.

6.3 REFERENCES

1. Luger, K. & Richmond, T. J. (1998). The histone tails of the nucleosome. *Current Opinion in Genetics &amp; Development* **8**, 140-146.
2. Luger, K., Mader, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. (1997). Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* **389**, 251-260.
3. Strahl, B. D. & Allis, C. D. (2000). The language of covalent histone modifications. *Nature* **403**, 41-45.
4. Banères, J. L., Martin, A. & Parello, J. (1997). The N tails of histones H3 and H4 adopt a highly structured conformation in the nucleosome. *J. Mol. Biol.* **273**, 503-508.
5. Wang, X., Moore, S. C., Laszckzak, M. & Ausio, J. (2000). Acetylation increases the alpha-helical content of the histone tails of the nucleosome. *J. Biol. Chem.* **275**, 35013-35020.
6. Liu, H. & Duan, Y. (2008). Effects of post-translational modifications on the structure and dynamics of histone H3 N-terminal peptide. *Biophys. J.* **94**, 4579-4585.
7. Yang, D. & Arya, G. (2011). Structure and binding of the H4 histone tail and the effects of lysine 16 acetylation. *Phys. Chem. Chem. Phys.* **13**, 2911-2921.
8. Potoyan, D. A. & Papoian, G. A. (2011). Energy Landscape Analyses of Disordered Histone Tails Reveal Special Organization of Their Conformational Dynamics. *J. Am. Chem. Soc.* **133**, 7405-7415.
9. LaPenna, G., Furlan, S. & Perico, A. (2006). Modeling H3 histone N-terminal tail and linker DNA interactions. *Biopolymers* **83**, 135-147.
10. Allfrey, V. G., Faulkner, R. R. & Mirsky, A. E. (1964). Acetylation and methylation of histones and their possible role in the regulation of RNA synthesis. *Proc. Natl. Acad. Sci. U. S. A.* **51**, 786-794.
11. Wallace, R. B., Sargent, T. D., Murphy, R. F. & Bonner, J. (1977). Physical properties of chemically acetylated rat liver chromatin. *Proceedings of the National Academy of Sciences* **74**, 3244-3248.

12. Norton, V. G., Marvin, K. W., Yau, P. & Bradbury, E. M. (1990). Nucleosome linking number change controlled by acetylation of histones H3 and H4. *Journal of Biological Chemistry* **265**, 19848-19852.

13. YAU, P., THORNE, A. W., IMAI, B. S., MATTHEWS, H. R. & BRADBURY, E. M. (1982). Thermal Denaturation Studies of Acetylated Nucleosomes and Oligonucleosmes. *European Journal of Biochemistry* **129**, 281-288.

14. Davey, C. A., Sargent, D. F., Luger, K., Maeder, A. W. & Richmond, T. J. (2002). Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 Å resolution. *J. Mol. Biol.* **319**, 1097-1113.

15. Brower-Toland, B., Wacker, D. A., Fulbright, R. M., Lis, J. T., Kraus, W. L. & Wang, M. D. (2005). Specific Contributions of Histone Tails and their Acetylation to the Mechanical Stability of Nucleosomes. *Journal of Molecular Biology* **346**, 135-146.

16. Gansen, A., To¦üth, K., Schwarz, N. & Langowski, J. ê. (2008). Structural Variability of Nucleosomes Detected by Single-Pair Fo¦êrster Resonance Energy Transfer: Histone Acetylation, Sequence Variation, and Salt EffectsΓÇá. *J. Phys. Chem. B* **113**, 2604-2613.

17. Wang, X. & Hayes, J. J. (2008). Acetylation Mimics within Individual Core Histone Tail Domains Indicate Distinct Roles in Regulating the Stability of Higher-Order Chromatin Structure. *Molecular and Cellular Biology* **28**, 227-236.

18. Yang, D. & Arya, G. (2011). Structure and binding of the H4 histone tail and the effects of lysine 16 acetylation. *Phys. Chem. Chem. Phys.* **13**, 2911-2921.

19. Tse, C. & Hansen, J. C. (1997). Hybrid Trypsinized Nucleosomal Arrays: Identification of Multiple Functional Roles of the H2A/H2B and H3/H4 N-Termini in Chromatin Fiber Compaction. *Biochemistry* **36**, 11381-11388.

Summary

The histone tails have for decades been regarded as unstructured, polypeptide chains[1, 2] which simply served as molecular beacons to protein effectors which modify chromatin[3]. However, some Circular Dichroism (CD) experiments have shown that the histone tails make a structural contribution to CD spectra [4, 5]. Molecular Dynamics (MD) studies have also found some structural features in the tails [6-9]. We therefore decided to undertake a computational study of the histone H3 N-terminal tail and some of its post translation modifications (PTMs). To this end we develop computational tools to store and analyse a high volume of data. We subsequently developed the database simDB, which stored all analysis data obtained from custom Python (www.python.org) analysis software developed to analyse our molecular modeling results generated with the software YASARA[10].

Because protein – protein docking is still a developing field, we wanted to confirm the veracity of our approach [11, 12]. We thus tested AutoDock [13], implemented in YASARA, with the small hairpin peptide of KSHV - LANA, which has been experimentally shown to bind to an acidic patch on the nucleosome surface[14]. We found the most accurate docking approach to be rigid ligand docking, which faithfully docked the KSHV LANA peptide to the nucleosome surface in a position virtually indistinguishable to that seen in the co-crystal. We also designed a custom docking protocol for docking to large receptors, such as the nucleosome. This protocol essentially split the search space into smaller overlapping search spaces or cells, and then performed independent docking in each of these cells, thereby increasing coverage and accuracy.

We proceeded with the explicit MD simulations of 11 differently modified 38 – residue H3 N-terminal tails for 500 ns each. We found that, in agreement with CD experimental work [4, 5], other MD studies [6, 8, 9], and secondary structure prediction algorithms, the unmodified tail showed the

stabilization of two α – helices. The first was located near the N-terminal tip of the structure and the second was located in the middle of the structure. We subsequently showed that there was a distinct lack of α – helical content observed in tails which contained either acetylated lysine residues, and were therefore enriched in nucleosomes that are typically associated with actively transcribed regions of chromatin. In comparison, tails with modification patterns associated with silent chromatin showed an overall increase in α – helical content..

We subsequently docked three 15 – residue H3 tail structures from each of the MD trajectories to the nucleosome using our docking protocol. We found that the unmodified and K9 methylated (without phosphorylation) tails docked laterally to the nucleosome between the exiting points of the H3 and H2B tails, in between the DNA and octamer, and between the two DNA gyres. The structures from the other trajectories yielded low to negative binding energy values; however they were bound in the same position as the top docking poses. Thus, the phosphorylated serine and acetylated lysines in these tails made binding at this specific site energetically less favorable.

We proposed a molecular mechanism whereby chromatin compaction is carried out at a nucleosome level, and is regulated by transitions in the N-terminal H3 tail structures, which, in turn, are modulated by specific epigenetic PTM patterns.

Keywords

chromatin, molecular dynamics, molecular docking, histone H3, N-terminal, acetylation, methylation, phosphorylation, database, python

References

1. Luger, K., Mader, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. (1997). Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* **389**, 251-260.

2. Luger, K. & Richmond, T. J. (1998). The histone tails of the nucleosome. *Current Opinion in Genetics &amp; Development* **8**, 140-146.

3. Strahl, B. D. & Allis, C. D. (2000). The language of covalent histone modifications. *Nature* **403**, 41-45.

4. Banères, J. L., Martin, A. & Parello, J. (1997). The N tails of histones H3 and H4 adopt a highly structured conformation in the nucleosome. *J. Mol. Biol.* **273**, 503-508.

5. Wang, X., Moore, S. C., Laszckzak, M. & Ausio, J. (2000). Acetylation increases the alpha-helical content of the histone tails of the nucleosome. *J. Biol. Chem.* **275**, 35013-35020.

6. Liu, H. & Duan, Y. (2008). Effects of post-translational modifications on the structure and dynamics of histone H3 N-terminal peptide. *Biophys. J.* **94**, 4579-4585.

7. Yang, D. & Arya, G. (2011). Structure and binding of the H4 histone tail and the effects of lysine 16 acetylation. *Phys. Chem. Chem. Phys.* **13**, 2911-2921.

8. Potoyan, D. A. & Papoian, G. A. (2011). Energy Landscape Analyses of Disordered Histone Tails Reveal Special Organization of Their Conformational Dynamics. *J. Am. Chem. Soc.* **133**, 7405-7415.

9. LaPenna, G., Furlan, S. & Perico, A. (2006). Modeling H3 histone N-terminal tail and linker DNA interactions. *Biopolymers* **83**, 135-147.

10. Krieger, E., Koraimann, G. & Vriend, G. (2002). Increasing the precision of comparative models with YASARA NOVA – a self-parameterizing force field. *Proteins* **47**, 393-402.

11. Andrusier, N., Mashiach, E., Nussinov, R. & Wolfson, H. J. (2008). Principles of flexible protein – protein docking. *Proteins* **73**, 271-289.

12. Wang, C., Bradley, P. & Baker, D. (2007). Protein – Protein Docking with Backbone Flexibility. *Journal of Molecular Biology* **373**, 503-519.

13. Morris, G. M., Huey, R., Lindstrom, W., Sanner, M. F., Belew, R. K., Goodsell, D. S. & Olson, A. J. (2009). AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *Journal of Computational Chemistry* **30**, 2785-2791.

14. Barbera, A. J., Chodaparambil, J. V., Kelley-Clarke, B., Joukov, V., Walter, J. C., Luger, K. & Kaye, K. M. (2006). The nucleosomal surface as a docking station for Kaposi's sarcoma herpesvirus LANA. *Science* **311**, 856-861.

Opsomming

Die histoon sterte is al vir die afgelope paar dekades beskou as ongestruktueerde polipeptied kettings wat slegs dien as molekulêre bakens vir proteïen effektore wat verandering in chromatien bewerkstellig. Dit was egter deur CD experimente getoon dat die histoon sterte wel ‘n struktuele bydrae maak tot CD spektra van die nukleosoom. Molekulêre Dinamika (MD) studies het ook getoon dat die sterte ‘n struktuele komponent het. Dus het ons besluit om ‘n *in silco* studie van die histoon H3 N-terminale stert met van sy chemiese modifikasies te onderneem. Met hierdie doel voor oë het ons rekenaar sagteware ontwikkel om ‘n hoë hoeveelheid data te stoor en te analiseer. Gevolglik het ons Python (www.python.org) geskrewe sagteware onwikkel om die rou data wat deur die MD program YASARA gegenereer is te ontgin en na die databasis simDB aan te stuur, wat gevolglik die stoor van hierdie data behartig het. SimDB was ook inhuis ontwikkel.

Omdat proteïen – proteïen dokking nog ‘n ontwikkelende veld is, wou ons die geskiktheid van ons dokkings benadering vas te stel. Dus het ons die YASARA geimplementeerde AutoDock sagteware getoets met die klein haarnaald peptied van KSHV LANA, wat eksperimenteel bewys was om op die oppervlak van die nukleosoom te bind. Ons het gevind dat dokking met ‘n rigiede ligand die akkuraatste benadering was omdat hierdie benadering die KSHV – LANA peptied op die nukleosoom oppervlak op so ‘n manier kon dok dat dit ononderskeibaar was van die oorspronklike kristal struktuur. Ons het ook ‘n nuwe dokking protokol geskep om na groot reseptore soos die nukleosoom te kan dok. Hierdie protokol behels basies die skeiding van die soek ruimte in kleiner, oorvleuelende soek ruimtes sodat onafhanklike dokking in elkeen van hierdie soek ruimtes uitgevoer kan word. Hierdie benadering verhoog gevolglik die dekking en akkuraatheid van die dokking.

Hierna het ons MD simulasies uitgevoer op 11 histoon H3 sterte (38 residue in lengte) met verskillende modifikasies vir ‘n tydperk van 500 ns elk. Ons het gevind dat die ongemodifiseerde stert twee α – helikse gestabiliseer het, ineenstemming met CD eksperimentele werk, ander MD studies en sekondêre struktuur voorspelling algoritmes. Die eerste α – heliks was naby aan die N-terminale punt van die struktuur gevind en die tweede α – heliks was meer na die middel van die struktuur gevind. Ons het gevolglik gewys dat sterte met geassetieleerde lisien residue, wat met geneties aktiewe gebiede verbind word, ‘n algemene tekort aan α – heliks inhoud getoon het. In vergelyking het die modifikasies, wat met geneties onaktiewe gebied verbind word, ‘n toename in α – heliks inhoud getoon.

Hierna het ons drie 15 – residu H3 stert strukture van elk van die MD bane gedok na die nukleosoom met ons nuwe dokkings protokol.Ons het bevind dat die ongemodifiseerde stert en die K9 gemetieleerde (sonder fosforilering) sterte lateraal and die nukleosoom gebind het tussen die ingangs – en uitgangs punte van die H3 and H2B sterte: tussen die DNA en die histoon oktamer en tussen die twee DNA draaie. Die strukture van die ander bane het lae tot negatiewe bindings energie waardes opgelewer, alhoewel die strukture op dieselfde posisie gebind het as die top dokkings strukture. Dus het die gefosforileerde serien en die geassetieleerde lisien residue in hierdie sterte die binding in hierdie spesifieke gebied minder gunstig gemaak.

Ons het ‘n molekulêre meganisme voorgestel waar chromatien kompaksie uitgevoer word op ‘n nukleosoom vlak en gereguleer word deur oorgang tussen die verskillende N-terminale H3 stert strukure, wat op hul beurt deur spesifieke epigenetiese modifkasie patrone gemoduleer word.

Abstract

The histone tails have for decades been regarded as unstructured polypeptide chains which simply served as molecular beacons to protein effectors which modify chromatin. However some experimental evidence shows that the tails may contain structure. Thus we conducted a Molecular Dynamics study of the Histone H3 tail and it's most important post translationally modified isoforms. The 500 ns experiments showed the evolution of different secondary structure conformations for the different modified isoforms. More interestingly the active isoform showed a statistically significant longer reach compared to the inactive isoform. We next conducted a molecular docking study of the 15 – residue tip of the H3 tail to the nucleosome surface. The starting structures were sampled from the Molecular Dynamics trajectories. The tips showed binding to nucleosome where the H3 tail exits the nucleosome, between the DNA and the octamer. This binding position did not cha nge between the different isoforms. We thus propose a molecular mechanism whereby chromatin compaction is carried out at a nucleosome level, and is regulated by transitions in the N-terminal H3 tail structures, which, in turn, are modulated by specific epigenetic PTM patterns.

Keywords

chromatin, molecular dynamics, molecular docking, histone H3, N-terminal, acetylation, methylation, phosphorylation, database, python