Enhancing the solubility of intrinsically disordered HIV-1 Tat protein at physiological pH and structural investigation by NMR spectroscopy

by

Kiran Krishnamurthy

A Thesis Submitted to the Faculty of Graduate Studies of The University of Manitoba in partial Fulfillment of the Requirements for the Degree of

Master of Science

Department of Chemistry

University of Manitoba

Winnipeg

Abstract

Human Immunodeficiency Virus-1 (HIV-1) Transactivator of transcription (Tat) protein is a 101-residue intrinsically disordered protein, responsible for enhancing the transcription process and ultimately viral replication. To understand its mechanistic role in enhancing transcription and viral replication, a study of its structure and dynamics in the presence of its binding partners is required. However, the protein is soluble only under acidic conditions (pH 4) and it precipitates at pH 7, which precludes the determination of its structure, dynamics and interactions under physiological conditions. Hence, the primary objective of this research was to solubilize Tat-protein near pH 7 so that further studies may be carried out to discern the mechanistic role of the Tat protein in viral replication.

Multiple approaches were employed to improve the solubility of the Tat protein at pH 7. A sequence-specific nickel-assisted cleavage (SNAC) approach that involves cleaving the polyhistidine-tagged Tat was employed to produce full-length Tat protein minus the purification tag, but this was unsuccessful owing to an unpredictable internal cleavage site. Poly-anionic RNA from Torula yeast was used as a solubilizing agent and it was found to increase the solubility of Tat but Nuclear Magnetic Resonance (NMR) spectra were only marginally improved. Tat was observed to be insoluble at pH 4−7 in the presence of TAR (TransActivation Response) RNA, one of the binding partners of Tat which is vital for the functioning of the protein. Moreover, the solubility of Tat was tested in a solution containing HIV-1 TAR RNA and Torula yeast RNA. Here too, the protein was soluble but no major improvement in the NMR spectra of Tat was observed. Tat protein tagged with a supercharged sequence at the N-terminal was genetically engineered and expressed to test the effect of a high net-charge on Tat's solubility. The increased net charge did not improve Tat's solubility. Genetic engineering was used to replace seven cysteine residues of

the Tat protein with aspartic acid to study the role of cysteine residues in the aggregation of the protein at neutral pH. This approach produced highly soluble Tat at pH 7 with well-resolved NMR spectra leading to the realization of the important role that Cys oxidation plays in the solubility of the protein. The carboxy terminal domain of RNA polymerase II (RNAP II) undergoes liquid-liquid phase separation in the presence of a crowding agent. With the expectation that Tat might be soluble in a liquid-liquid phase separated medium at physiological pH, a RNAP II domain fusion protein was expressed, but poor yields of proteolyzed protein precluded Tat solubility studies.

Several analytical techniques were employed to characterize the structure of Tat protein. Fluorescence studies were carried out to understand the structural changes taking place as the pH is elevated. Somewhat surprisingly, the fluorescence spectra indicated that the single Trp residue resides in a solvent-restricted region of intrinsically-disordered Tat. Infrared spectroscopy was used to study the secondary structure of the protein and to quantify the fractions of different secondary structures comprising the Tat protein in the range of pH 4−7. Multi-dimensional solution- and solid-state NMR spectroscopy were employed as the primary analytical tool in the structural analysis of Tat protein. In addition to conventional indirect-detection NMR methods, ^{15}N-direct-detection NMR experiments were attempted to help monitor solubility, structure and dynamics.

Keywords: HIV, IDP, NMR, RNAP II, SNAC, Tat, TAR.

Acknowledgements

First and foremost, I want to thank my advisor Prof. Joe O'Neil for giving me a wonderful opportunity to study at the University of Manitoba and work on this high-impact project. His immense knowledge in the area and his encouragement to explore new ideas have been decisive in my growth as a graduate student and have moulded my understanding of the subject. Moreover, the excellent lab-facilities that were available in Prof. O'Neil's lab helped me dearly to sharpen and master my biochemistry skills. I would like to thank my committee: Dr. Mazdak Khajehpour and Dr. Gerd Prehna, for inspiring discussions during progress meetings and constructive wonderful feedback on my thesis proposal and the thesis. These discussions moulded my understanding of the subject further. Thanks to Dr. David Davidson, the lab manager at the Prairie NMR facility, University of Manitoba, for training me on the Varian NMR instrument and all his help in setting up solution- and solid-state NMR experiments. I will always cherish the discussions I had with him while optimizing the experiments. Special thanks to Dr. Vu To for his help in setting up the NMR pipe software and holding my hand during my initial days of working on the Tat project. I want to thank Chun Hin Wong (Michael) for helping me understand the fundamentals of biochemistry and guiding me with protein expression and purification. Thanks to my lab mates Herby Cadet and Dr. Mary Harnando for joyful discussions in the lab. I want to acknowledge Prof. Sean McKenna's group (Dr. Amit Koul, Nikhat Lubna, Taegi Choi, Dr. Evan Booy and Daniel Gussakovsky) for providing TAR plasmid and helping me in the *in-vitro* synthesis of TAR RNA and giving me access to the FPLC instrument. The Faculty of Graduate Studies, University of Manitoba is acknowledged for accepting my application to the graduate studies. Finally, I am

grateful to my family for all their support, encouragement and sacrifices that have led me here, today.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AA	amino acids
AIDS	acquired immunodeficiency syndrome
ATR	attenuated total reflection
BBB	blood brain barrier
BPTI	bovine pancreatic trypsin inhibitor
CBP	cAMP-response element binding protein
CD	circular dichroism
CP	capsid protein
cDNA	complementary DNA
CCR5	C-C chemokine receptor type 5
CXCR4	C-X-C chemokine receptor type 4
CDK	cyclin dependent kinase
CDK9	cyclin dependent kinase 9
CPP	cell penetrating peptide
CSI	chemical shift index
CPMAS	cross-polarization magic angle spinning
CPMG	Carr-Purcell Meiboom-Gill
CAI	codon adaptation index
COSY	correlation spectroscopy
dsDNA	double stranded DNA
DSIF	5,6-dichloro-1-β-D-ribofuronosylbenzimidazole sensitivity inducing factor
DDX4	DEAD-Box helicase 4

DTT	dithiothreitol
DSS	2,2-dimethyl-2-silapentane-5-sulfonate
E.coli	*Escherichia coli*
EDTA	ethylenediaminetetraacetate acid
EMF	electromotive force
EXSY	exchange spectroscopy
Env	envelope protein
FTIR	fourier transform infrared spectroscopy
FPLC	fast protein liquid chromatography
FT-ICR	fourier transform ion-cyclotron resonance
HIV	human immunodeficiency virus
HD	hydrogen-deuterium exchange
HSQC	hetero nuclear single quantum
HEPES	4-(2-hydroxyethyl)-1-piperazineethane sulfonic acid
IDP	intrinsically disordered proteins
IDR	intrinsically disordered regions
IN	integrase
INEPT	insensitive nuclei enhanced by polarization transfer
IPTG	isopropyl β-D-1 thiogalactopyranoside
LLPS	liquid-liquid phase separation
LTR	long terminal repeat
MC	membraneless compartments
MHC	major histocompatibility complexes

MA	matrix protein
MES	2-(N-morpholino) ethane sulfonic acid
MALDI	matrix assisted laser desorption/ionisation
NMR	nuclear magnetic resonance
Nef	negative factor
NC	nucleo protein
NRTI	nucleoside reverse transcriptase inhibitors
NtRTIs	nucleotide reverse transcriptase inhibitors
NNRTI	non-nucleoside reverse transcriptase inhibitors
NELF	negative elongation factor
ncRNA	non-coding RNA
NCD	net charge density
NOE	nuclear Overhauser effect
NTP	nucleotide triphosphate
ORD	optical rotatory dispersion
PR	protease
PI	protease inhibitor
pTEFb	positive transcription elongation factor b
PDSD	proton driven spin dynamics
Q-TOF	quadrupole-time of flight
RNP	ribonucleo protein
RT	reverse transcriptase
RF	radio frequency

SU	surface glycoprotein
Sn RNA	small nuclear RNA
Snc RNA	small nucleolar RNA
SEP	solubility enhancement peptide
SSP	secondary structure propensity
ssNMR	solid-state nuclear magnetic resonance
SEC	size exclusion chromatography
SDS-PAGE	sodium dodecyl sulfate polyacrylamide gel electrophoresis
Tat	Transactivator of transcription
TM	transmembrane protein
TAR RNA	transactivation response element RNA
TOCSY	total correlation spectroscopy
TCEP.HCl	tris(2-caboxyethyl) phosphine hydrochloride
Vif	viral infectivity factor
Vpu	virus protein U.

1. Introduction

Proteins are a major class of biomolecules that are vital for life on earth and are responsible for a myriad of biological processes. They perform a vast array of functions such as catalyzing metabolic reactions [1], replicating DNA [2], helping cells to respond to stimuli [3], translocation of molecules [4] and many others. Catalytic functions of proteins are highly specific; for example, the enzyme Invertase present in the yeast-extract hydrolyses α-glucosides but not β-glucosides whereas, the enzyme emulsin exhibits a contrasting behavior by hydrolyzing only β-glucosides. Emil Fischer made this important observation in 1894 and concluded that the enzyme-substrate interactions are highly distinct, brought about by the key chemical interactions and substrate-specific changes in the three-dimensional structure of enzymes, popularly referred to as the *lock and key model* [5,6]. In the mid-1930s, Mirsky and Pauling scrutinized the structures of native and denatured proteins and concluded that, upon denaturation, the native-structure of a protein loses its key 3D structural features, rendering it futile [7]. In the last six decades, there have been many direct- and indirect-studies on the three-dimensional structure of proteins [8–18]. Notably, in 1958, Kendrew *et al.* published the first three-dimensional structure of myoglobin [19] (**Figure 1**), and in 1958, Perutz *et al.* elucidated the structure of haemoglobin [20]. These two scientific papers laid out the foundation for detailed studies on the classification of proteins and their structures [21].

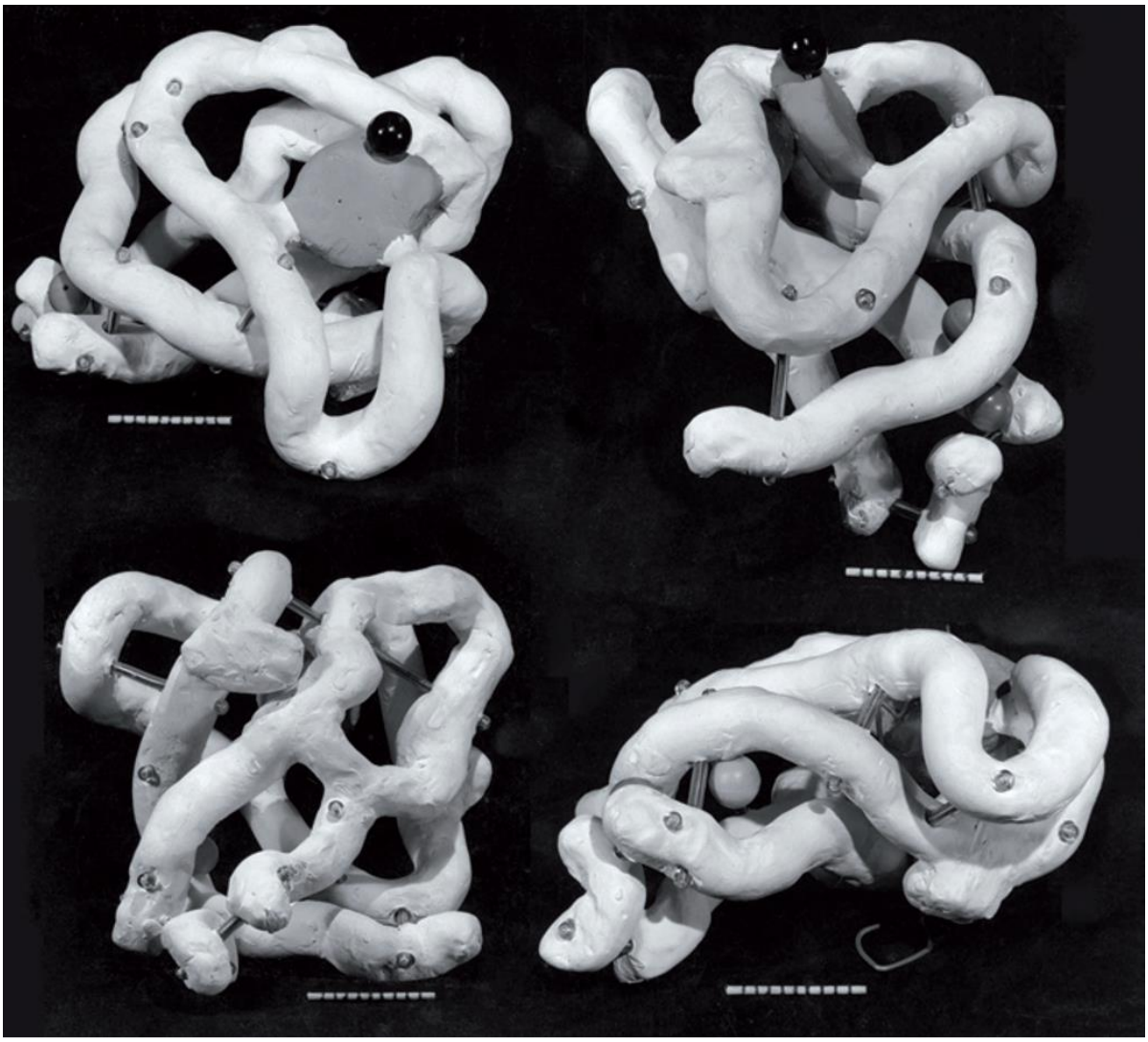


Figure 1. A 3D picture of Myoglobin reconstructed from the images in the archives of the medical research council Laboratory, Cambridge, UK. The Polypeptide chains and the heme group are represented by white-ribbons and grey discs, respectively. *Reprinted with permission from reference [19]. Copyright © 1958 Springer Nature.*

1.1 Protein structure

The carboxyl and the amino groups from two different amino acids (AAs) undergo a condensation reaction forming covalent bonds called *peptide bonds* (O=C-N-H). A sequence of such reactions

between different amino acids generates a long-chain of amino acids called a *polypeptide (PP) chains* [22–25] (**Figure 2**). The non-covalent interactions between amino acids (side-chain interactions), brought about by the residue-specific functional groups, help these chains attain a 3D structure. Based on the length of the polypeptide chain and the extent of its folding, protein structures are classified into four categories: primary, secondary, tertiary and quaternary [22,23].

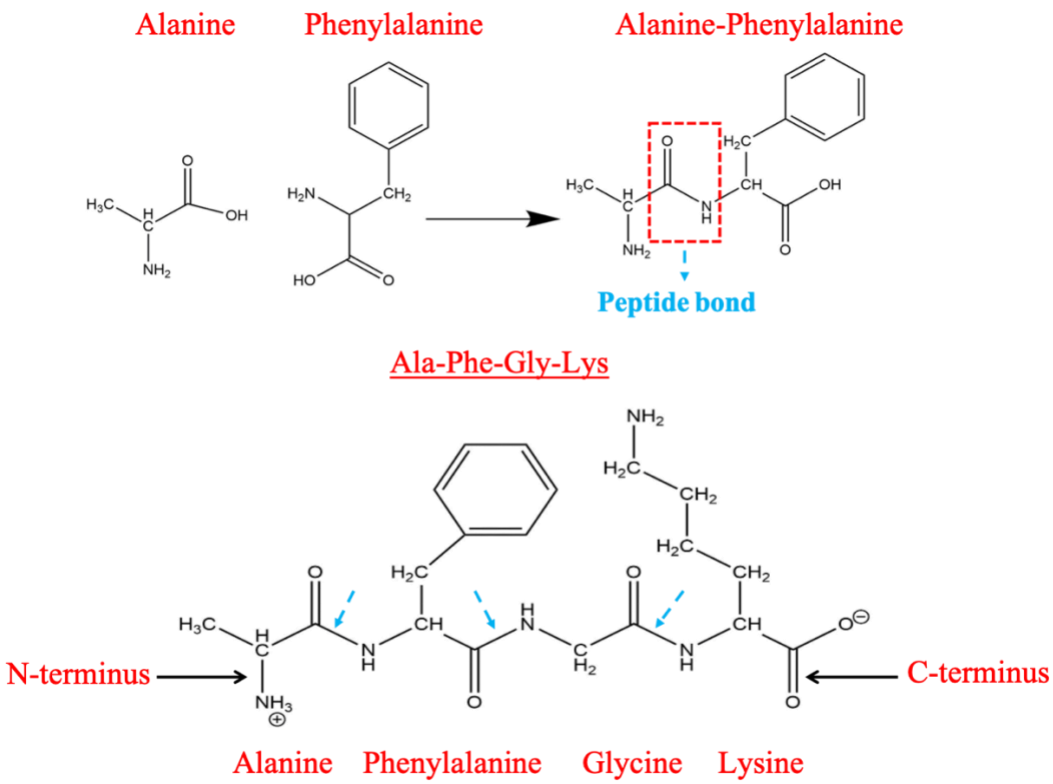


Figure 2. Chemical structures of amino acids alanine and phenylalanine and the alanine-phenylalanine dipeptide (**top**). A polypeptide chain of alanine, phenylalanine, glycine and lysine (Ala-Phe-Gly-Lys) (N→C) with an N-terminus and a C-terminus is also shown wherein the peptide bonds are marked with blue arrows (**bottom**).

1.1.1 Primary structure:

All proteins constitute PP chains, and their chemical identities are signified by their amino acid sequences. The amino acid sequences encoded by gene-specific mRNAs are considered primary structures. The chains terminate with an amino group and a carboxyl group on two terminals, referred to as N-terminus and C-terminus, respectively (**Figure 2**). The rule of thumb to read an amino acid sequence is from N-terminus to C-terminus [25]. Insulin is a classic example of a protein, which is made of two polypeptide chains, A (51 AAs) and B (20 AAs) **(Figure 3)**, connected through intra-chain disulfide (S−S) linkages [18]. It belongs to a small class of signaling molecules called peptide-hormones responsible for important physiological reactions. Other examples of peptide-hormones are glucagon [26], oxytocin [27] and somatostatin [28].

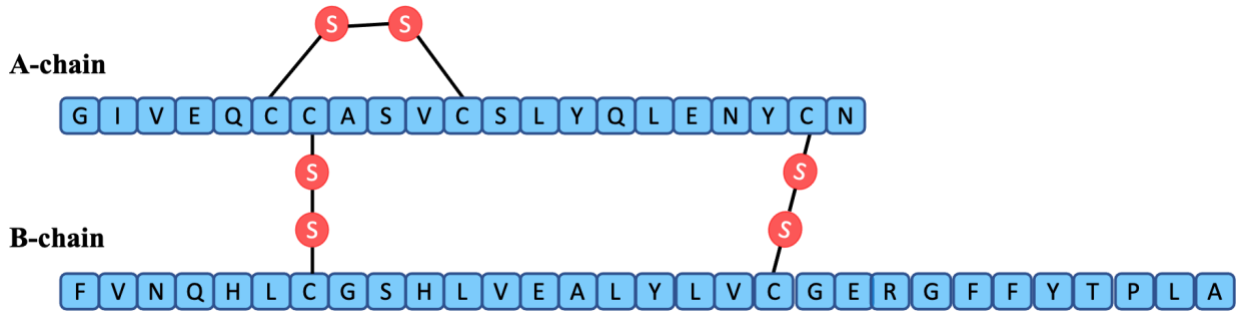


Figure 3. An illustration of the primary structure of insulin consisting of two polypeptide chains interconnected by disulfide bridges [23].

1.1.2 Secondary structure:

The hydrogen bonds between the amide protons and the carbonyl oxygens and non-covalent inter-residue interactions make the PP chains fold and twist, leading to secondary structure conformers with characteristic backbone dihedral angles, phi (ϕ) and psi (ψ) (**Figure 4**). When these ϕ and ψ torsional angles are plotted (x- and y-axis, respectively), the dihedral angles of residues within a PP chain cluster into specific regions of the $\phi - \psi$ plot, popularly called the *Ramachandran plot* [29,30]. Each of these regions represents a particular type of secondary structure and the plot is utilized to deconstruct complex peptide and protein structures. The secondary structures are of three types: *α-helix, β-sheet and β-turn.*

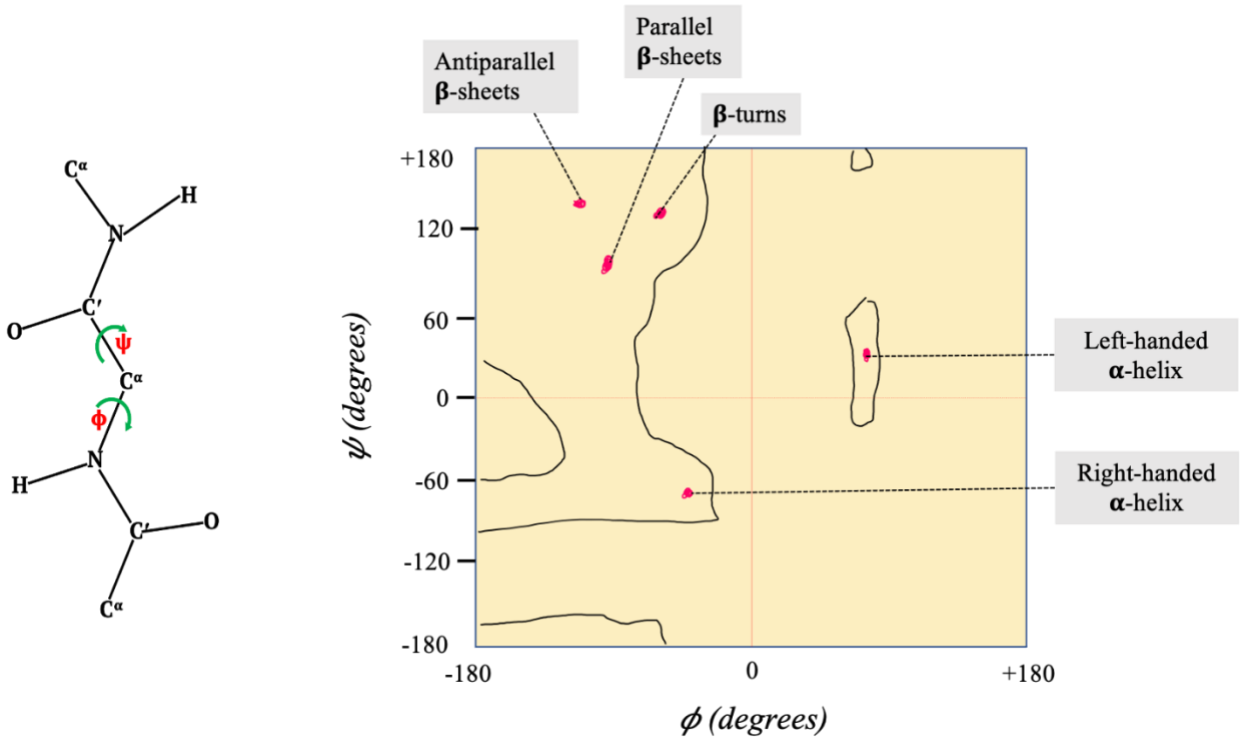


Figure 4. A schematic of a dipeptide with the ϕ and ψ angles drawn (**left**) and the Ramachandran plot (**right**) wherein the ϕ and ψ ranges of different secondary structures are listed [31]**.**

α-helices:

In a polypeptide chain, the carbonyl group on one residue (*i*) interacts with an amino group on the fourth residue (*i+4*) through hydrogen bonds, forcing the chain to twist into a helical form. The backbone of the chain aligns along the helical axis, and the side-chain groups face away. The helical conformation of peptides **(Figure 5)** was first proposed by Pauling *et al.* [8] in 1951, based on an in-depth structural analysis of amino acids and peptides. In their model, molecular features such as interatomic distances, bond angles and other configurational parameters were matched with the experimental values derived for different amino acids. A single helical turn is stretched about 0.54 nm along the axis and constitutes 3.6 amino acid residues [22,24]. The most commonly found secondary structure in proteins is the right-handed *α-helix* (α_R) [24]. Additionally, there are two more helical conformations called *3_{10}-helix* and *π-helix,* which are compared with the *α-helix* in **Table 1.** As the amino acid proline lacks an amide-hydrogen, it cannot form hydrogen bonds with other residues and hence, the PP chains with proline cannot form helices, giving it the name helix breaker. Examples of α-helical proteins include, α-keratin [10], myosin [16], haemoglobin [15], myoglobin [19].

Table 1. Structural features of different helices [8,32,33].

	Helical-turn contribution per residue	**Residues per turn**	**Translation along the helical axis per residue**	**Residues involved in the hydrogen bonding**	**Helical pitch**	**Number of Atoms**
α-helix	100°	3.6	1.54 Å	i + 4 → *i*	5.4 Å	13
3_{10}-helix	120°	3.3	1.93-2.0 Å	i + 3 → *i*	5.8-6 Å	10
π-helix	85.2°	4.4	1.28 Å	i + 5 → *i*	1.5 Å	16

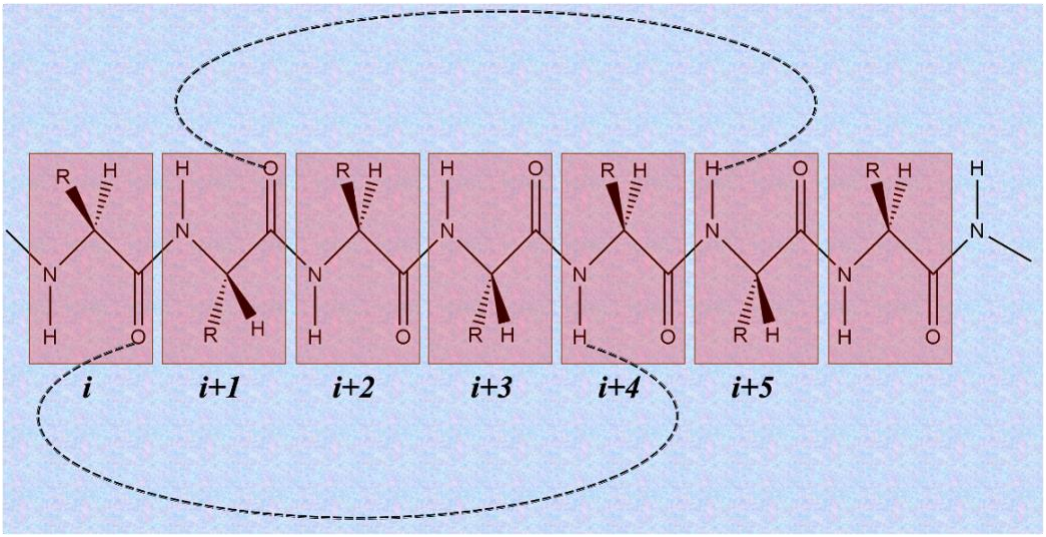


Figure 5. A schematic of the α-helix. Black dashed-lines mark the internal hydrogen bonds in the α-helix. The carboxyl oxygen of the i^{th} residue and the amide-hydrogen of the $(i+4)^{th}$ residue are marked in dashed line, respectively.

β-sheets:

The β-sheets are the second major secondary-structures in proteins, formed by inter-chain hydrogen bonding between the amide proton of one chain and the carbonyl group of a peptide bond in the second chain. These chains are 8 to 10 residues long and are referred to as ***β-strands*** [34]. The sidechain (R) functional groups are aligned above and below the protein backbones. If the chains run in the same direction, for example, N-terminus to C-terminus, the resulting sheets are called parallel β-sheets, and if they are oppositely aligned (N→C *vs*. C→N), the sheets are considered anti-parallel. A structural schematic of both parallel and anti-parallel arrangement of β-sheets is shown in **Figure 6**.

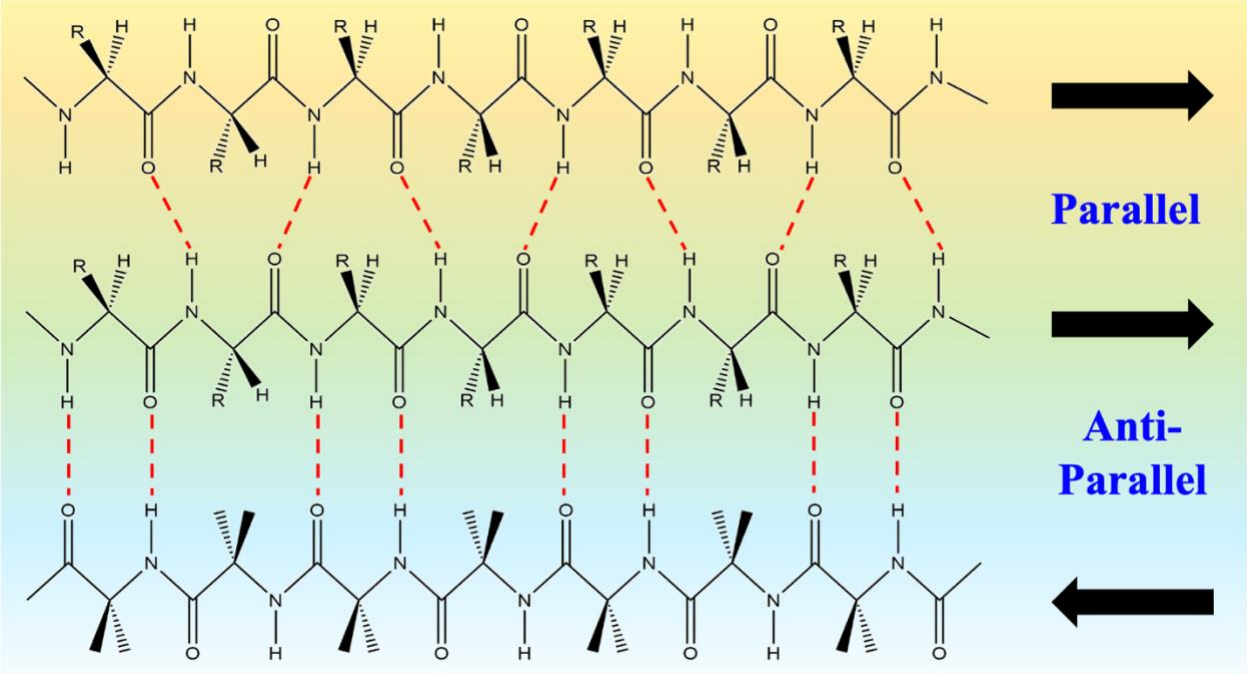


Figure 6. A picture depicting the parallel and anti-parallel arrangement of β-sheets. The sheets are interconnected through hydrogen bonds, which are marked with red dashed lines.

β-turn (β-bend):

The three-dimensional structure of proteins is mainly possible due to *β-turn*s in the secondary structure, which folds a protein chain onto itself, allowing proteins to adopt biologically active conformations. The *β-turn*s are comprised of four amino acids, mainly glycine and proline, due to their structural flexibility and cyclic structure, respectively [24,35]. The turns are stabilized by hydrogen bonds between the carbonyl oxygen of the first residue and the fourth amide proton in the β-turn **(Figure 7).**

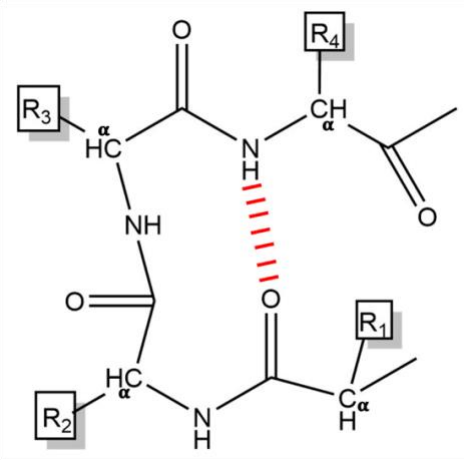


Figure 7. Structure of a typical *β-Turn.* The red dashed line represents the hydrogen bond between the carbonyl group of the first residue and the amide proton on the fourth residue (Adapted from [22]).

Polyproline II helix:

Poly-L-proline type-II (PP-II) is a fourth class of secondary structure, which was first found in structured fibrillary proteins present in collagen. Many structured and disordered proteins have been identified with this secondary structure in their sequences [36] but, the fraction of PP-II in not high and generally corresponds to that of 3_{10}-helix [36,37]. The PP-II structure functions as a binding site, aiding the formation of protein-protein and protein-DNA(RNA) complexes [36]. It typically adopts a trigonal prism-like structure with backbone dihedral angles (φ, ψ) of *ca.* -75° and 150°, respectively. It has three residues per turn and the translational helical rise per turn is 3.1 Å vis-à-vis α-helix, which is 1.5 Å. Due to the presence of fewer amino acids, regular inter-residue hydrogen bonds are rarely present [36,37]. Although proline is found ubiquitously in the PP-II structure, it is not essential for the PP-II structural conformation.

1.1.3 Tertiary structure:

The alpha-helices and beta-sheets interact through covalent and non-covalent (hydrophobic, hydrophilic, ionic, and van der Waal's) interactions, facilitating proteins to acquire a three-dimensional structure (**Figure 8**). The fraction of polar and non-polar amino acids in the peptide chains dictate the folding ability of the proteins. While the polar side-chains involved in hydrophilic interactions at the surface improve the solubility of proteins, the hydrophobic sidechains are buried within the chains, which stabilize the tertiary structure [38]. Moreover, the disulfide bridges between two cysteine residues and the ionic interactions called *salt-bridges* between charged residues add additional stability to the tertiary structure.

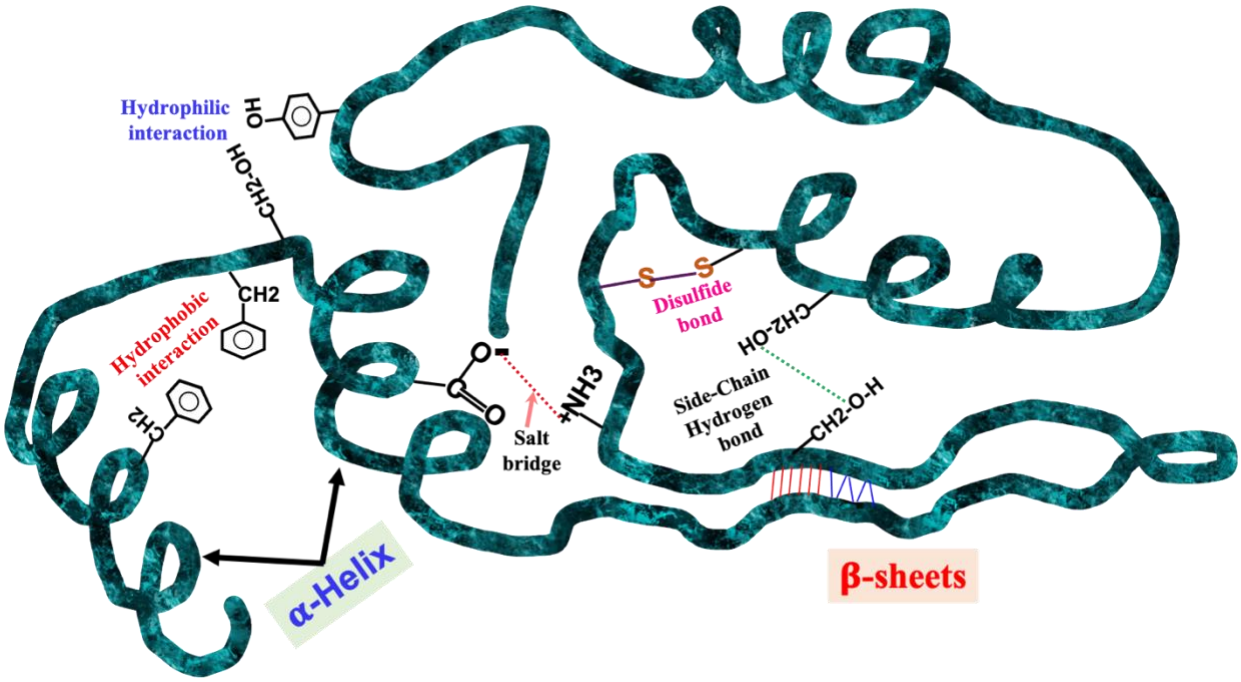


Figure 8. Overview of the tertiary structure formed by secondary structure interactions both covalent and non-covalent.

1.1.4 Quaternary structure:

Proteins that require an assemblage of several polypeptide subunits [22] to carry out their functions are considered to have a quaternary structure. Depending on the number of subunits in the assembly, the quaternary structures are classified as dimers (2 subunits), trimers (3 subunits) and tetramers (4 subunits). If identical subunits are involved, the prefix "homo" is used (homodimer), and if the subunits are chemically distinct, the dimer is referred to as a heterodimer. The protein haemoglobin is a well-known example of a quaternary structure as it constitutes four polypeptide subunits (two α- and two β-subunits). Each α- and β-subunit interact to form a heterodimer, and two such heterodimers integrate to form the biological form of hemoglobin [15].

1.2 Intrinsically disordered proteins

Researchers have identified cases where non-structured segments in a protein also play an equally important role in its functioning, which topples the consensus that protein-function depends exclusively on the protein's 3D structure. For example, the egg-protein Phosvitin was discovered to lack three-dimensional structures by Jirgensons *et al.* using single-wavelength optical rotatory dispersion (ORD) in 1958 [39]. The disordered structure of Phosvitin was further confirmed by Grizzuti and Perlmann in 1970 by Circular Dichroism (CD) spectroscopy [40] and Hans Vogel by Nuclear Magnetic Resonance (NMR) spectroscopy in 1983 [41]. Based on the alpha-helical content of proteins derived from the ORD results, Jirgensons *et al.* classified the proteins into three groups: a) high alpha-helical, b) low alpha-helical, and c) non-helical. Notably, they grouped the non-structured proteins as non-helical [42]. Up to 17% of eukaryotic proteins are completely disordered, and it is estimated that about 40-50% of eukaryotic proteins show the presence of long intrinsically-disordered regions (IDRs) (≥30 consecutive residues) [43]. These disordered proteins and regions are involved in a plethora of cellular functions, including transcription, translation, controlling cell-cycle and cell-signaling [44,45].

In the late 1990s, research on proteins lacking a stable three-dimensional structure gained thrust during which, many disordered proteins were characterized and were tagged using different names emphasizing their structure such as flexible [46], mobile [47], intrinsically-unstructured [48], disordered [49], denatured [50], unfolded [51], vulnerable [52], dancing protein [53], protein cloud [54], *et cetera*. At the beginning of the 21st century, the term "Intrinsically disordered proteins" (IDP) was adopted as the standard nomenclature to address such disordered proteins. Once the ***Genomic*** era started, scientists speculated on a few protein sequences obtained from genomic sequencing suggesting that they would not fold into normal protein structures and

therefore, would be non-functional [55]. However, recent studies on protein chemistry reveal that such proteins remain fully functional, although the whole protein or parts are completely unstructured/disordered in an aqueous medium. [56,57]. The intrinsically disordered proteins (IDPs) are defined as *"Proteins which are different from other proteins lacking fixed or ordered secondary or tertiary structures with high intramolecular flexibility"* [44,45,48,58–60].

The striking feature of IDPs is their biological activity despite their failure to adopt a well-defined 3D structure. They appear to have a dynamic structural ensemble at the secondary or tertiary level [61]. The disordered structure of a protein is a function of its amino acid sequence as the residue-specific intra-, and inter-residue interactions are responsible for the stability and folding capacity of protein structures [62]. Many of the disordered proteins share a common amino acid composition in their polypeptide sequence [62]. A survey of 275 naturally folded and 91 intrinsically disordered proteins [63] concludes that a combination of the low proportion of hydrophobic amino acids and high net-charge due to high fractions of polar residues influences proteins to adopt disordered structures [64]. High net-charge leads to strong electrostatic charge repulsions [64], and low numbers of hydrophobic residues result in weaker intra-chain van der Waal's interactions. Specifically, the IDPs have fewer order-promoting amino acids (tryptophan, tyrosine and phenylalanine), bulky hydrophobic groups (valine, isoleucine and leucine), cysteines, and high numbers of disorder-promoting amino acids (arginine, glycine, glutamine, lysine, serine and alanine) and helix-breaking prolines [49,62,64,65]. The following order of amino acids is based on their abilities to promote order in protein structures wherein, the order-promoting amino acids are positioned to the left, and the disorder-promoting amino acids are positioned to the right: W, F, Y, I, M, L, V, N, C, T, A, G, R, D, H, Q, K, S, E, P [62,66]. Another distinctive feature of

IDPs is the low-complexity in their protein sequences and higher abundance of repetitive amino acids compared to folded proteins [64].

The IDPs actively interact with proteins through protein-protein interactions and function as connecting nodes or hubs within a protein complex [64]. Many disease-causing IDPs, for example, p21 [67], p27 [68], p53 [69], Breast cancer 1 (BRCA1) [70], Estrogen receptor [71], Xeroderma pigmentosum, complementary group A (XPA) [72], and α-synuclein [73] form protein-hubs with numerous binding partners [64,74]. The intrinsically disordered regions (IDRs) within the IDPs act as flexible linkers connecting two or more ordered domains [64,74,75]. Strikingly, many IDPs undergo folding upon binding to a partner protein [76,77] whereas others conserve their structural disorder even in the bound state [78].

1.2.1 IDP predictors

The puzzling features of IDPs, attributable to their peculiar amino acid compositions [62,64], are a topic of interest to researchers worldwide. The similar nature of IDPs in terms of their amino acid compositions promoted the development of computational programs to predict disordered regions in primary protein sequences [62]. The evolution of IDP predictors can be categorized into three generations [79,80] :

a) First generation: 1979−2001
b) Second generation: 2002−2006
c) Third generation: 2007 and later

Williams [80] proposed the first IDP predictor in 1979, which utilized the random coil conformations to identify regions lacking the globular structure. In depth inspections declared it as a poor predictor as it lacked experimental data [80]. Romero *et al.* [81] developed a predictor

by choosing 67 disordered (length: 1340 residues) and many ordered domains (length: 16,543 residues) from the protein data bank and studying their physio-chemical properties such as aromaticity, flexibility, hydropathy, hydrophobicity and amino acid composition [66]. Another IDP predictor based on the charge and hydropathy of the amino acids (CH plot) was developed by Uversky *et al.* in 2000 [63]. Later in 2005, the CH plot was extended to the FoldIndex method [82]. Between 2002 and 2006, many second-generation methods such as Globplot [83] and IUpred [84] were developed, which relied only on the amino acid composition of proteins. Besides, methods utilizing both amino acid composition as well as machine learning were also developed (PONDR [85], DisEMBL [86] and DISOPRED [87]). Methods based on position-specific score matrix files such as PONDR-VL3P [88], DISOPRED2 [89], PROFbval [90], DISpro [91] and NORSp [92] were developed at the end of the second generation. Post-2006, more sophisticated methods, which are a fusion of multiple algorithms and several individual predictor methods, have been developed involving machine learning and meta predictors. Example include, MFDp [93], MetaDisorder [94], MFDp2 [95] and DisMeta [96]. In general, the IDP predictors have shown great promise in identifying the disordered regions, making them an integral part of selecting targets in genomics [43]. Currently, there are more than 50 algorithms available for predicting the IDPs and IDRs [80].

1.2.2 Folding Mechanism in IDPs

The tightly-regulated interactions between large biomolecules which are highly specific in nature, are fundamental to all processes in living organisms [97]. The challenges of discerning folding and binding mechanisms in proteins have been simplified by various theoretical and experimental approaches, yet, many elementary concepts are still unclear [98]. Unlike structured proteins that fold into a biologically relevant three-dimensional shape, the IDPs abundant in eukaryotic cells remain largely unfolded and appear to be a dynamic ensemble of interconverting structures. Upon binding with suitable partner proteins, some IPDs adopt a well-defined structure, widely seen in regulatory and signaling processes [99].

The coupled folding and binding of IDPs to other IDPs and globular proteins is described by the *conformational selection* and *induced-fit* mechanisms **(Figure 9)**, which differ in the order of occurrence of the binding and folding processes [99]. In the conformational selection mechanism, the IDP undergoes a transient folding before binding to the partner protein. On the contrary, the binding occurs first in the induced-fit mechanism, followed by the conformation change. The former mechanism is also known as pre-existing, folding-before-binding and population-shift mechanisms and the latter as the binding-before-folding mechanism. In the first phase of the conformational selection mechanism, the partner proteins weakly bind to the native IDPs, which induces a conformational change. Once the IDP adopts a suitable conformation, tight-binding is established [98].

In the induced-fit mechanism, weak interactions are established first between the IDP and its partner-protein, which is succeeded by the conformational change of the IDP. These two mechanisms are demonstrated using an example (**Figure 9**) wherein, the N-terminal region of transactivation domains of transcription factor c-Myb [TAD c-Myb] depicted in red binds to the

KIX domain of the transcriptional coactivator cAMP-response element binding protein (CBP) (green) [100] using both methods.

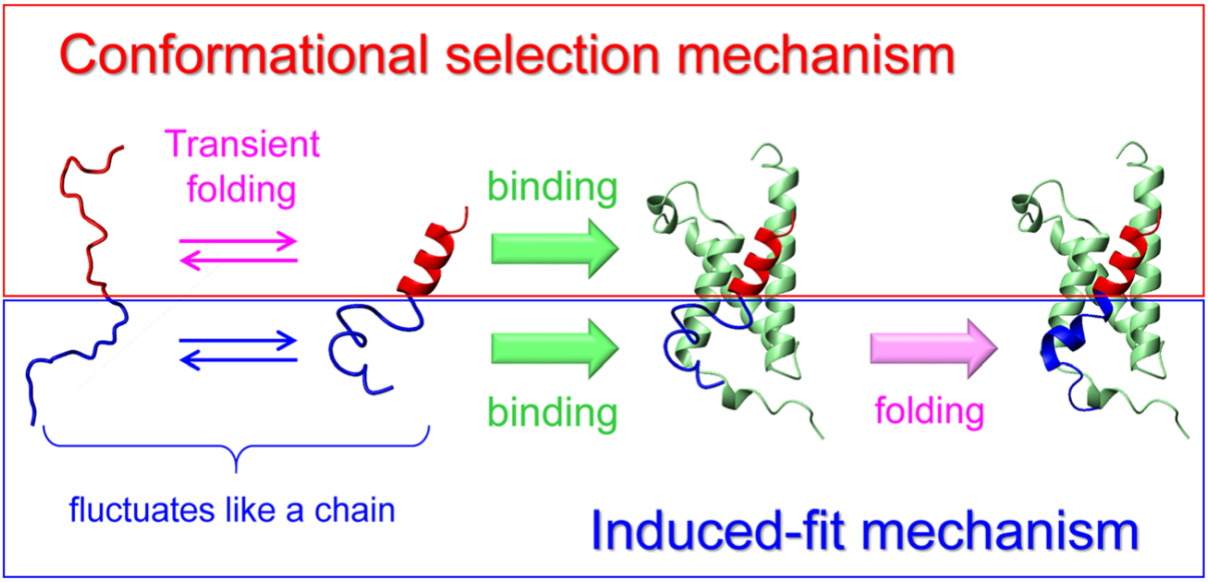


Figure 9. A schematic explaining the conformational selection (**top-section**) mechanism and the induced-fit mechanism (**bottom-section**) Adapted with kind permission from reference [98].

1.2.3 Fuzzy complexes

In physiological conditions, some IDPs interact with other IDPs and proteins [101] by molecular recognition adopting a well-defined secondary or tertiary structures [102], as the binding induces the transition from a disordered to an ordered state. It is also recognized that some IDPs have high specificity for target molecules and undergo rapid association and dissociation upon binding to them [103], *i.e.*, they exist in a dynamic equilibrium. A closer look at such complexes discloses that the IDPs achieve partial-ordering upon binding and sometimes can retain their disorderness [104,105]. Due to the observed dynamics in the binding of some IDPs in a protein-complex [106],

they. Are referred to as *Fuzzy complexes* [103] and are classified into four categories (**Figure 10**) [102].

a. **Polymorphic complex:** The bound molecule adopts two or more resolvable conformations. For example, the WH2 domain of Wiskott–Aldrich syndrome protein binds to the actin protein *via* a 3-residue or an 18-residue segment in alternative locations.
b. **Clamp complex:** Two folded-regions (structured) of a protein dock to its partner *via* a protein segment called a linker, which has a disordered structure even in its bound-state. The two folded-regions are referred to as clamps. In **Figure 10**, the two structured regions of the nonsense-mediated decay factor UPF2 binds to the UPF1 protein *via* a linker chain, whose structure remains ambiguous.
c. **Flanking complex:** The IDP binds to a partner protein through short motifs embedded in its disordered regions. Due to this minimal interaction, much of the disordered region neighbors the partner protein's ordered-regions in the complex without strongly interacting chemically. The binding of transcription factor Ultrabithorax to DNA is an example of the formation of flanking complexes.
d. **Random complex:** The dynamic binding of the IDPs to partner proteins generates random complexes, for example, the phosphorylation sites in the cyclin-dependent kinase inhibitor Sic 1 interchange upon contacting Cdc4.

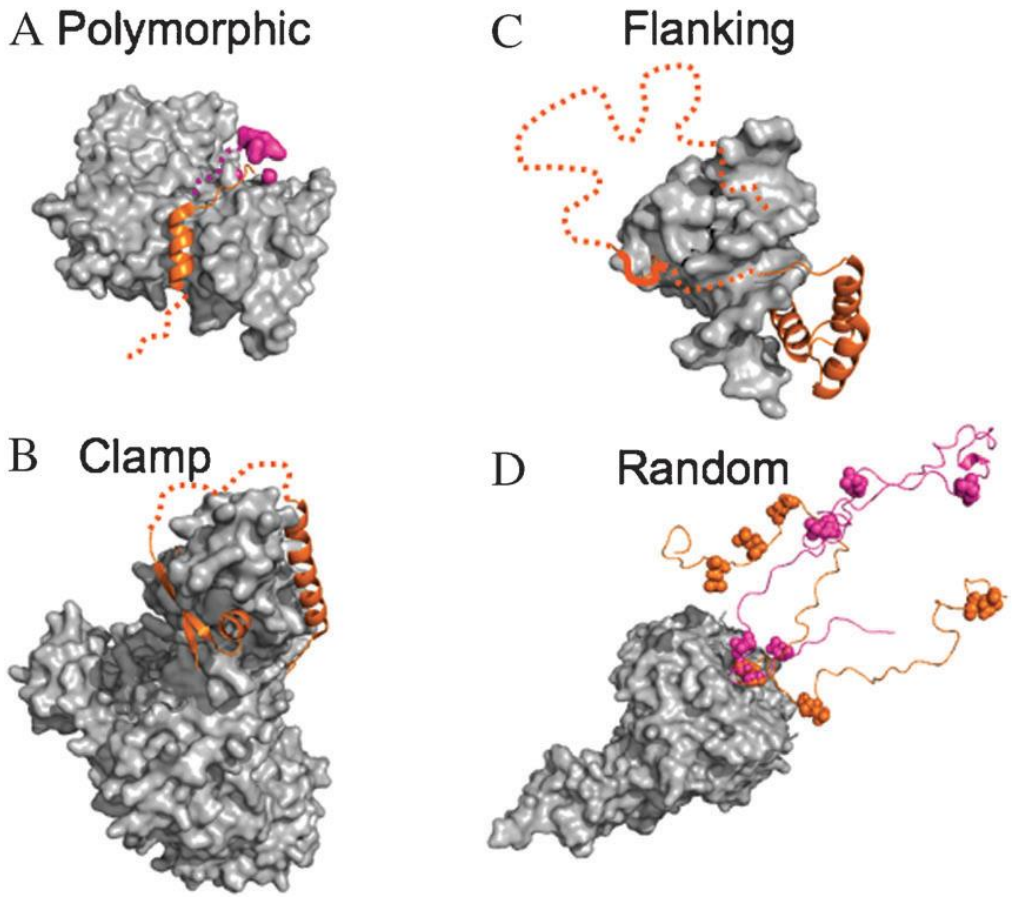


Figure 10. Conformations of different fuzzy complexes. The proteins involved in each case are described in the text. Adapted with permission from reference [102]. *Copyright © 2012 Royal society of chemistry.*

1.2.4 Functions of Intrinsically disordered proteins

The functioning of IDPs as hubs for the assembly of large multi-protein complexes has proven critical for many biological processes, and examples include the assembly and disassembly of microfilaments, organization of chromatin and the transportation of small molecules. Some chaperones that assist in folding RNA and proteins have also been identified to contain disordered regions in their sequence [107]. These disordered regions allow chaperones to interact with a broad range of proteins, both folded and unfolded through rapid macromolecular interactions [108] for example, α-synuclein [109] and α-casein [110]. Some IDPs form membrane-less compartments (MCs) within eukaryotic cells through a process called liquid-liquid phase separation (LLPS) [111]. The Ribonucleoprotein (RNP), granules, cluster of signaling complexes and nucleoli consist of MCs [112]. Signaling complexes control the signal transduction in living cells, often made of IDPs and their partner proteins [45].

1.3 Human Immunodeficiency Virus (HIV)

Acquired immunodeficiency syndrome (AIDS) is one of the most devastating epidemics in history [113]. In September 1982, the Centers for Disease Control and prevention coined this term for the very first time and described it as: *"a disease at least moderately predictive of a defect in cell-mediated immunity, occurring in a person with no known cause for diminished resistance to that disease"* [114]. In 1983, Luc Montagnier *et al.* from the Pasteur Institute, Paris and Gallo *et al.*, from the National Cancer Institute, Maryland, discovered a retrovirus responsible for acute compromisation of the immune system in humans. Initially, they were labeled as the lymphadenopathy-associated virus (LAV) and the human t-lymphotropic virus-III (HTLV-III), respectively, but were dropped for a new term, Human Immunodeficiency Virus (HIV) [115,116]. The chronic, life-threatening condition caused by HIV is called acquired immunodeficiency syndrome (AIDS). According to the Joint United Nations Program on HIV and AIDS (UNAIDS), *ca.* 38 million people were infected with HIV in 2019 [117]. Among them, *ca.* 36.2 million were adults, and *ca.* 1.7 million were children who were less than 15 years old. The estimated number of new HIV infections was around *ca.* 1.7 million across the globe in 2019 [117].

1.3.1 HIV genome and viral structure

HIV belongs to the genus called Lentivirus, grouped under the family Retroviridae and the subfamily Orthoretrovirinae [118]. Based on the genetic information and differences in the viral antigens, HIV is categorized into HIV-1 and HIV-2 [119]. The HIV-1 evolved from the Simian Immunodeficiency virus, mainly from central African chimpanzees (SIVcpz), and the HIV-2 evolved from sooty mangabeys (SIVsm) located in West Africa [119]. While the HIV-1 is responsible for the global AIDS pandemic, the HIV-2 is confined to certain regions of central- and

West-Asia [120]. In retroviruses, RNA is the genetic information carrier, unlike other viruses where DNA is the genetic material [117,120,]. The HIV-1 virus has a 9719 nucleotide-long RNA with multiple open reading frames **(Figure 11)** [120]. The open reading frame is 5′ to 3′, and both ends are flanked by long terminal repeat (LTR) sequences. The body of the HIV-1 RNA is composed of structural (gag, pol, env), regulatory (tat, rev) and accessory genes (nef, vpr, vif, vpu) [119]. A complete list of proteins encoded by these genes and their functions are listed in **Table 2.** The proteins are labeled as per their molecular size and chemistry; for example, a glycoprotein and a protein with sizes of 5 and 10 kDa are labelled gp5 and p10.

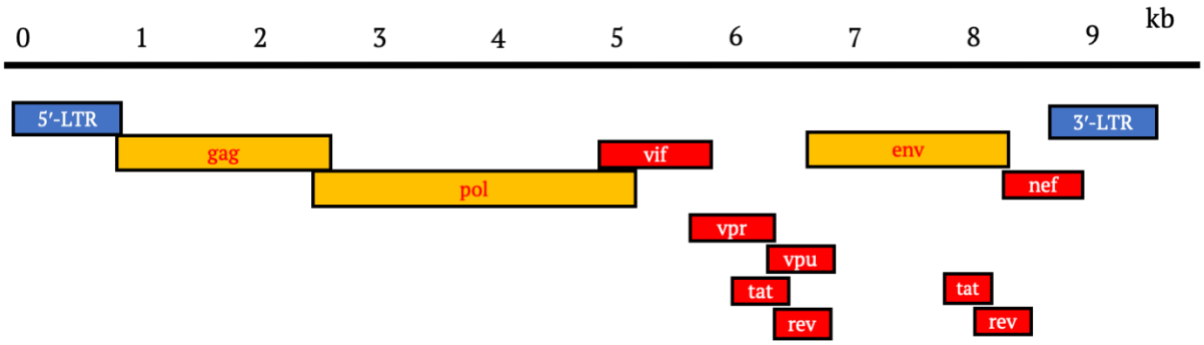


Figure 11. The HIV-1 genome composition with open reading frames marked in blue-rectangles (Adapted from [119]). LTR: long terminal repeat; *gag*: group-specific antigen; *pol*: polymerase; *vif*: viral infectivity factor; *vpr*: virus protein r; *tat*: Transactivator of Transcription; *vpu*: virus protein unique; *rev*: RNA splicing regulator; *env*: envelope; *nef:* negative regulating factor

The HIV virion **(Figure 12)** is approximately 100 nm in diameter and has an outer lipid-bilayer embedded with glycoproteins and transmembrane proteins. The protein assembly is in a knob form wherein the protein gp120 and gp41 form the cap (surface) and the stalk (transmembrane) of the knob, respectively. The bilayer also consists of major histocompatibility complexes (MHC, in the figure). While the matrix protein (p17) lines the inner membrane, the capsid protein (p24) forms the capsid around the viral genome. The capsid hosts two copies of single-stranded RNA, reverse transcriptase (p51, p55, p66), integrase (p32) and nucleoproteins (p7).

Table 2. A list of genes present in the HIV genome, the encoded proteins and their functions [119].

Gene	Protein	Function
gag	Pr55Gag[a], capsid protein (CP)- p24, matrix protein (MA)- p17, nucleoprotein (NC)- p7	Formation of the capsid and inner membrane layers. Form protein-RNA complexes that help in the viral particle-release.
pol	Pr160GagPol[a], Protease (PR)-p10, Reverse transcriptase (RT)-p51, RNase H- p15(66), Integrase (IN)-p32	Proteolytic cleavage. Transcription of viral RNA to proviral DNA and its integration into the host genome. Degradation of viral RNA in the viral RNA/DNA replication complex.
env	PrGp160[a], Surface glycoprotein (SU)-gp120, Transmembrane protein (TM)-gp41	Help in the virus's attachment to the target cell and the fusion of viral and cell membranes.
tat	Transactivator protein-p14	Activates transcription of viral genes.
rev	RNA splicing regulator-p19	Regulates the export of partially spliced and non-spliced viral mRNA.
nef	Negative regulating factor-p27	Influences HIV replication, enhances virulence and downregulates CD4, a glycoprotein present on the surfaces of immune cells.
vif	Viral infectivity protein-p23	Critical for *in vivo* viral production
vpr	Virus protein r-p15	Affects cell-cycle and facilitates virus infectivity
vpu	Virus protein unique-p16	Controls viral particle release and CD4 degradation

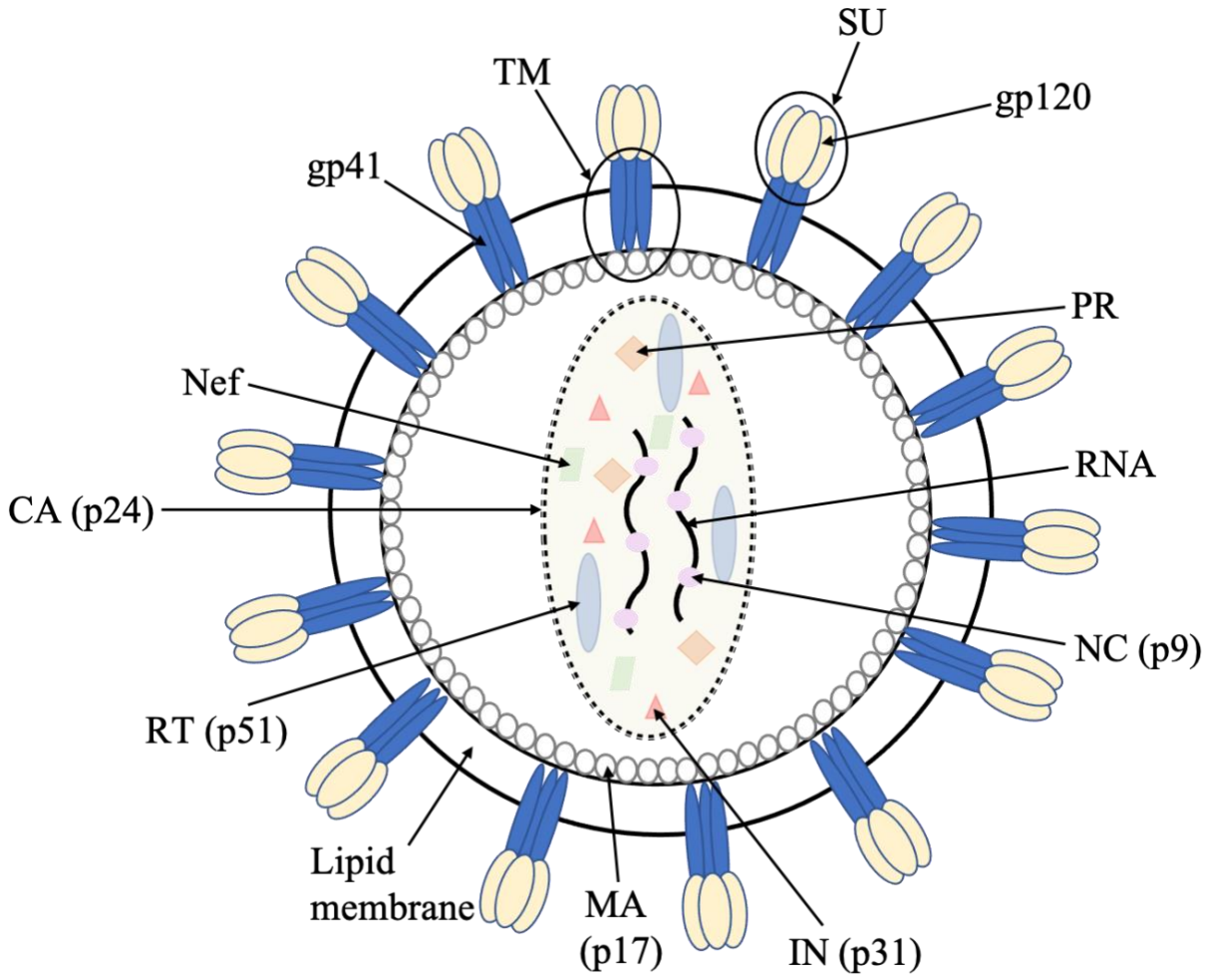


Figure 12. Diagram of the HIV-1 virion structure. SU: surface protein; RT: reverse transcriptase; TM: transmembrane protein; IN: integrase; CA: capsid protein; PR: protease; LI: linker protein; MA: matrix protein; NC: nucleic acid-binding protein; RNA: ribonucleic acid.

1.3.2 HIV infection and life cycle

The human immunodeficiency virus (HIV-1) exhibits a complicated life-cycle with six distinct phases: (i) binding and fusion, (ii) uncoating, (iii) reverse transcription, (iv) integration, (v) replication and assembly and (vi) budding **(Figure 13)** [120].

Binding and fusion: The surface glycoprotein (gp120) present on the viral envelope interacts with the CD4 receptors present on the cell surface of the immune cells: T-helper cells, dendritic cells, astrocytes and macrophages [119]. Upon binding, both the proteins undergo conformational changes generating an additional site for gp120, enabling it to bind to the co-receptors CCR5 and CXCR4 as well. The gp120 and gp41 undergo further conformational changes so that the N-terminus of gp41 integrates into the plasma membrane forming a channel, completing the fusion of the cell membrane and viral envelope.

Uncoating: Post-fusion, the viral capsid is translocated to phagosomes within the cytoplasm by endosomes. The capsid is digested, and the contents are released back into the cytoplasm. The uncoating process generates free viral RNA inside the cytoplasm wherein, the reverse transcriptase (RT) enzyme is activated.

Reverse transcription: The RT converts the single-stranded viral RNA (ssRNA) into complementary DNA (cDNA) with a parallel degradation of the ssRNA by RNase H. By the DNA-dependent DNA polymerase activity of the RT, the cDNA is transformed into a double-stranded DNA (proviral DNA).

Integration: The proviral DNA is transported to the nucleus through nucleopores in a linear or a circular form in conjunction with the integrase. The integrase inserts the proviral DNA into the host genome at random, completing the infection process.

Replication: The proviral DNA can be replicated alone or as a part of the host cell genome during cell division. Once the infected cells are activated, the viral genome's LTR promoter serves as an attachment site for cellular RNA polymerases and other transcription factors, initiating tat-regulated viral genomic RNA and mRNA synthesis. While the tat protein controls the synthesis of full-length viral mRNA, the rev protein promotes the transcription of genomic RNA and further translates structural and enzymatic genes needed for viral assembly.

Budding: The newly synthesized viral-RNA and viral-proteins move towards the cell membrane where new immature virions are assembled. The viral membranes are formed from the lipid bilayer and membrane-associated proteins of the host cell. [120,122,123]

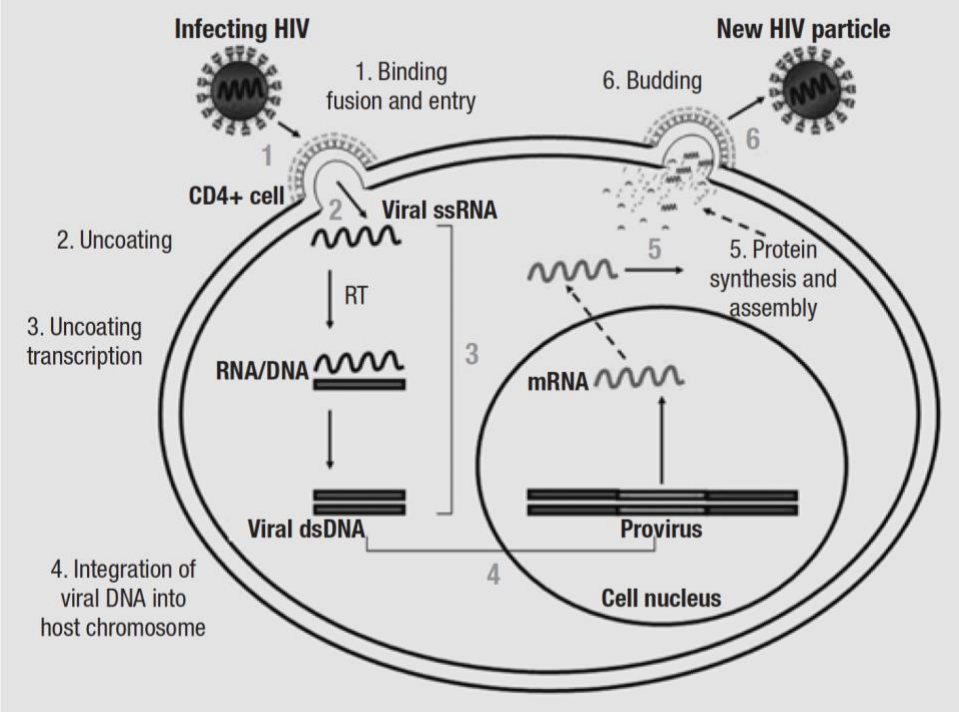


Figure 13. The HIV life-cycle. *Reproduced with kind permission from: Fanales-Belasio E, Raimondo M, Suligoi B, Buttò S. HIV virology and pathogenetic mechanisms of infection: a brief overview. Ann Ist Super Sanità. 2010;46(1): 5-14. DOI: 10.4415/ANN_10_01_02.*

1.3.3 Anti-HIV drugs

In the early days of HIV viral discovery, the infection of HIV was deadly as there were no medical treatments whatsoever. The isolation of the virus and a subsequent molecular-level life cycle characterization laid the foundation for anti-HIV drug research. In 1987, Azidothymidine (AZT), the nucleotide reverse transcriptase inhibitor (NRTI), was used to decrease mortality and infection rate in AIDS patients [122]. A combinatorial approach called antiretroviral therapy (ART) has become the standard HIV-infection treatment, which uses a combination of drugs [121,124]. In recent years many approaches have been adopted wherein different stages of the HIV life cycle are targeted. The different types of drugs commercially available are nucleoside reverse transcriptase inhibitors (NRTIs), nucleotide reverse transcriptase inhibitors (NtRTIs), non-nucleoside reverse transcriptase inhibitors (NNRTIs), protease inhibitors (PI), integrase inhibitors (INI), entry inhibitors or CCR5 antagonists (CCR5s) [125–127]. A comprehensive list of anti-HIV drugs is given in **Table 3**.

Table 3. A list of different classes of anti-HIV drugs and examples [119].

Anti-HIV drug classes	Drugs
Nucleoside/nucleotide reverse transcriptase inhibitors	Azidothymidine – Zidovudine, Didanosine, Zalcitabine, Stavudine, Lamivudine, Abacavir, Tenofovir, Emtricitabine
Non-nucleoside reverse transcriptase inhibitors	Nevirapine, Efavirenz, Delavirdine, Etravirine, Rilpivirine
Protease inhibitors	Saquinavir, Indinavir, Ritonavir, Nelfinavir, Lopanivir+Ritonavir, Atazanivir, Fosamprenavir, Tipranavir
Integrase inhibitors	Raltegravir, Elvitegravir, Dolutegravir
Entry inhibitors	Enfuvirtide, Maraviroc

1.3.4 Transactivator of Transcription (Tat) Protein

Transactivator of transcription protein, abbreviated as Tat-protein, is a 101 amino acid (11.5 kDa) long intrinsically disordered protein [128]. Two exons encode it, wherein the first exon encodes 72 amino acids, and the second exon encodes the remaining 29 amino acids. Based on the Tat-protein's amino acid distribution, the Tat sequence (**Figure 14**) is classified into six regions:

(a) The N-terminus proline-rich acidic region includes residues 1−21. The amino acids present in positions 1 and 2 play an important role in giving Tat a clean passage into the cells and act as pH sensors [129]. The activation of long terminal repeats (LTR) is served by the acidic region through interactions with the cyclin-T1, the cysteine-rich and the hydrophobic-core regions [130]. At the 11^{th} position in the acidic-region, the tryptophan is conserved and important in Tat's secretion and transportation into the cytosol [129,130].

(b) The cysteine-rich region runs from amino acids 22−37. It has seven well-conserved cysteines placed at positions 22, 25, 27, 30, 31, 34 and 37. These cysteine residues will form intramolecular disulfide bonds, and there is a high chance of mutation of cysteine to serine. Mutation at residue-31 may result in loss of transcriptional activity of Tat [131].

(c) The amino acids from the 38^{th} to 48^{th} positions constitute the hydrophobic-core region. The phenylalanine at the 38^{th} position and the residues from 41−48 (^{41}KGLGISYG^{48}) are conserved. In conjunction with the cysteine-rich region, it takes part in binding, specifically with co-factors like the histone acetyl-transferase (HAT), the Sp1 transcription factor and the CREB-binding protein (CBP)/p300 [132].

(d) The arginine-rich basic region has a well-conserved sequence ^{49}RKKRRQRR^{57}. This region plays an important role in binding to the trans-activation response element (TAR) RNA and relocating Tat into the nucleus [133].

(e) The glutamine-rich region constitutes the last part of the first-exon and includes residues from 58−72. When co-existent with the arginine-rich region, it is called the basic region and is involved in microtubule polymerization and apoptosis of T-cells [129,130].

(f) The second exon codes for the carboxyl-terminal domain region and includes amino acids from position 73 to 101 [134]. An RGD motif (amino acids 78 to 80) present in the second exon acts similar to the basic region, aiding in cell adhesion. The residues from positions 86 to 92 (^{86}ESKKKVE^{92}) participate in NF-kB assisted transcription of HIV genes [135][128].

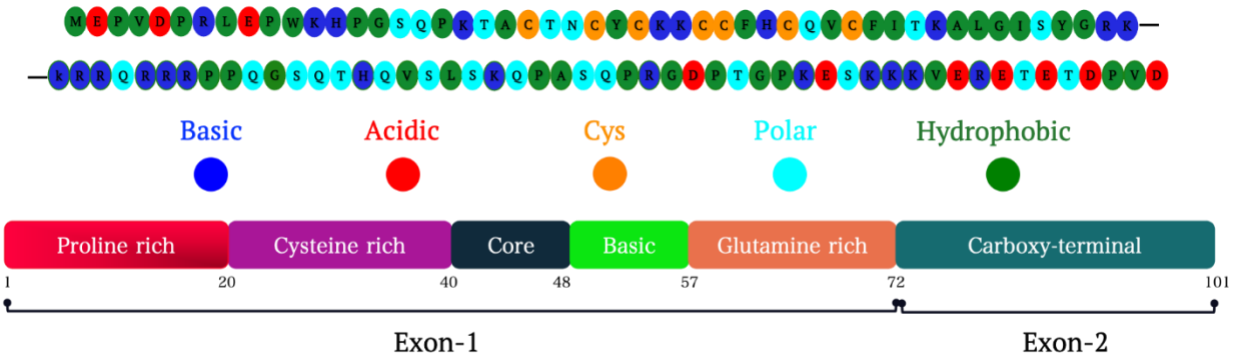


Figure 14. The complete sequence of the Tat protein with domains listed underneath the sequence.

The solubility of proteins is determined by their charged and polar residues. A protein is least soluble at its isoelectric point (pI), as the net-charge of a protein will be zero [136]. Being a basic protein, Tat (pI of 9.7) exhibits poor solubility in physiological conditions because the pI of the Tat (9.5) is too close to the physiological pH. As the pH shifts towards the acidic regime, the protein acquires a high net positive charge (**Table 4**), improving its solubility in an aqueous medium [137].

Table 4. The net charge of Tat protein at different pHs [138].

pH	Net charge
1	+32.992
2	+32.919
3	+32.284
4	+28.994
5	+23.154
6	+17.564
7	+12.215
8	+8.583
9	+3.245
10	-4.4.38
11	-9.451
12	-12.312
13	-17.588
14	-19.693

Here is an overview of how Tat-protein helps in the replication of HIV. Human immunodeficiency virus-1 (HIV-1) infects the T-cells of an individual and compromises the immune system making the person susceptible to all kinds of infections. During the initial phase of infection, the single-stranded viral RNA is reverse-transcribed into viral double-stranded DNA (dsDNA) and integrated into the host-DNA. Post-integration, the viral dsDNA is transcribed into new viral RNA, used both as genomic RNA and to make proteins for new virions [139].

After initiation, the transcription of viral DNA was halted due to the binding of RNA polymerase II (RNAP II) to the negative transcription elongation factor (NELF) and 5,6-dicholoro-1-β-D-ribofuronosylbenzimidazole sensitivity inducing factor (DSIF). The early translated viral protein Tat transports back into the nucleus from cytoplasm, where it binds to the trans-activation

response (TAR) element, a 59-nucleotide RNA sequence present at the 5' end of the viral transcript [134]. Upon binding to TAR-RNA, Tat recruits a hetero-dimer positive transcription elongation factor (p-TEFb) composed of cellular CDK9 (cyclin-dependent kinase) and Cyclin T1 (**Figure 15**). Tat activates the transcription by forming a complex with p-TEFb/Tat-TAR, when CDK9 in close vicinity of RNAP II, hyperphosphorylates the carboxy terminal domain (CTD) of RNAP II (**Figure 15**). Eventually, NELF and DSIF also get phosphorylated by CDK9 and fall off the DNA and this results in the replication of full-length viral RNA [139].

In the absence of Tat, only short viral transcripts are transcribed due to hypophosphorylation of RNA polymerase II [134,137]. Besides, evidence suggests that Tat-protein affects cellular functions due to its pleiotropic effects on cellular genes and host cell metabolism [140]. Tat is a multifunctional disordered protein that interacts with multiple proteins due to its flexible structure, adopting different local conformations. This helps it to associate with other binding partners like CDK9, TAR RNA, and Cyclin-T1 [128,137].

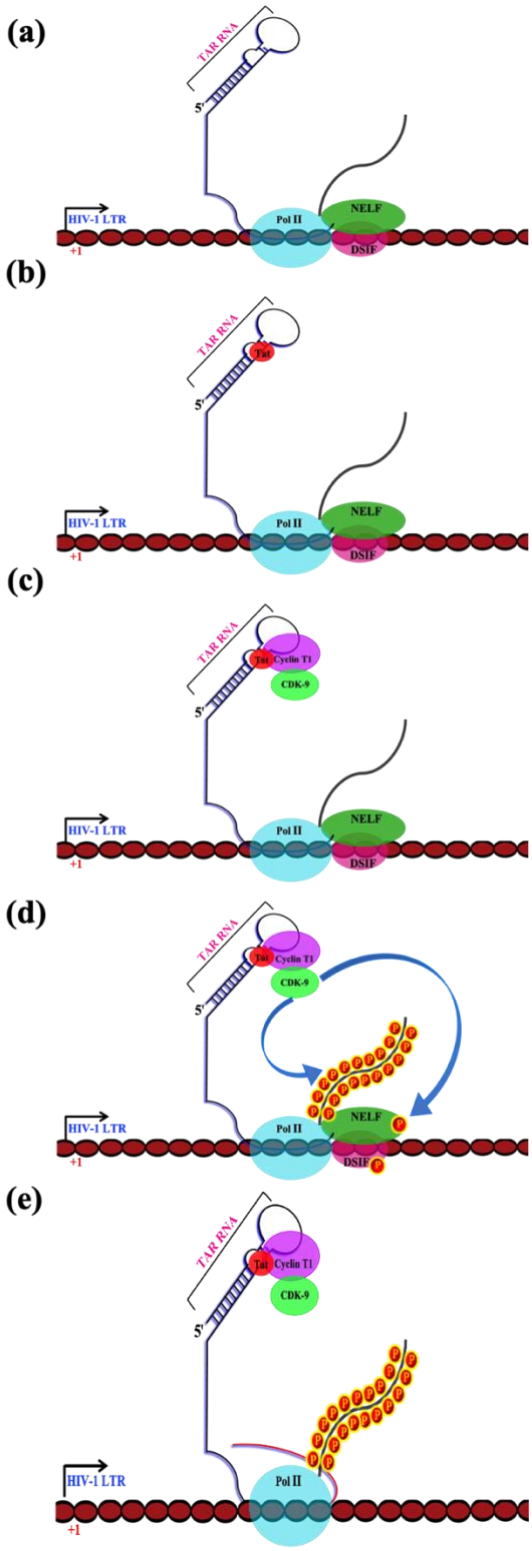


Figure 15. Events of Tat activating transcription of human immunodeficiency virus (HIV-1). (a) The initiation of HIV-1 viral DNA transcription (b) Tat binds to TAR-RNA (c) Tat recruits pTEFb

(Cyclin T1 and CDK9). (d) CDK9 hyperphosphorylate the CTD of RNAP II. (e) Hyperphosphorylation results in enhancement of viral transcription.

1.4 Liquid-liquid phase separation in biology

Eukaryotic cells consist of several membrane-less and membrane-bound organelles that permit various cellular activities and biochemical reactions to be done in a controlled spatiotemporal fashion [141]. Membrane-less species include Ribonucleoprotein (RNP), puncta, granules, clusters of signalling complexes, and nucleoli, all of which consist of proteins that exhibit high flexibility in their properties and functions, attributed to their lack of a secondary and tertiary structure [142]. Many studies have pointed out key molecular mechanisms that govern the formation of membrane-less compartments such as, fission and fusion, vesicle-mediated trafficking and inter-organelle interactions [141]. Another important mechanism which aids in the formation and development of protein- and nucleic acid-based membrane-less compartments is *liquid-liquid phase separation* (LLPS), and the so formed organelles are known as biomolecular condensates [143]. For example, T-cell receptor protein which aids in signal transduction, phase-separates upon phosphorylation [144]. Another example is the mTORC1 protein that modulates the phase-separation of PGL granules and plays a decisive role in controlling their autophagy degradation [145]. The aqueous solution and the biomolecular condensates form the phase-separated system. The phase-separation process is influenced by many factors such as, pH, temperature, osmotic pressure and ionic strength, which control the molecule-level interactions [143,146]. Furthermore, the condensates are stabilized by electrostatic interactions, Van der Waals forces, hydrogen bonds, hydrophobic interactions, cation-π interactions, and π-π stacking between aromatic residues (tyrosine, phenylalanine and tryptophan) [141,143].

In some cases, phase-separation occurs when proteins with positive and negative charges on their surfaces interact with multiple molecules, which could be either structured or intrinsically disordered. The interactions between IDRs are frequently weak and transient whereas the structured regions of proteins interact strongly with higher specificity. The weak interactions observed in IDR regions are attributed to their lack of a three-dimensional structure and the low-complexity amino acid sequence with a high content of amino acids such as Asn, Phe, and Tyr, and repetitive stretches of poly-glycine, poly-serine and poly-glutamine [147–149]. The aromatic residues play a crucial role in promoting some phase-separations. For instance, the disordered regions in the DEAD-box helicase 4 (DDX4) protein [150] constitute several Phe-Gly repeats wherein the Phe's participate in inter- and intra-molecular interaction with Arg's and establish cation$-\pi$ and $\pi-\pi$ interactions, thus promoting phase separation [150]. Gln, Asn and Ser residues can drive the phase separation through dipolar interactions [151].

1.4.1 Intrinsically disordered carboxy terminal domain of RNA polymerase II:

RNA polymerase is a multiprotein complex that catalyses the DNA-dependent transcription process of protein-coding genes in both prokaryotic and eukaryotic cells. It was first discovered by Weiss and Gladstone in 1959 in rat-liver cells [152]. In prokaryotes, all types of mRNA are transcribed only by a single RNA polymerase enzyme, whereas in eukaryotes, transcription is mediated by three different DNA-dependent RNA polymerase (RNAP) enzymes: I, II and III. RNAP-I synthesizes precursors of ribosomal RNA (rRNA); RNAP-II transcribes protein-encoding genes (mRNA), noncoding RNA (ncRNA) and small nuclear/nucleolar RNA (sn/sncRNA) [153]; RNAP-III synthesizes 5S and tRNA genes [154]. RNAP-II comprises 12 subunits and is the most important one among the three polymerases [153]. Its transcription function can be divided into three phases: initiation, elongation, and termination, and it employs

additional proteins to ensure proper functioning of transcription and gene expression. Recent *in vitro* studies show that the large intrinsically disordered carboxy terminal domain (CTD) tail of RNAP-II undergoes phase separation [146].

1.4.2 CTD of RNA Polymerase II:

The origin of RNAP-II is traced back to the early stages of cellular evolution and its core structure is highly conserved in lower eukaryotes (unicellular: *e.g.*, protists and fungi), plants, insects and mammals. RNAP-II is composed of 12 subunits, the two largest subunits are labelled RPB1 and RPB2, and have molecular weights of 220 and 140 kDa, respectively. These large subunits have a tendency to degrade by proteolysis, but they also undergo phosphorylation and post-translational regulation. Compared to RNAP-I and -III, RPB1 has a peculiar conserved carboxy terminal domain (CTD), which serves as an anchoring site for associative proteins helping RNAP-II in the transcription process [155]. It was first discovered by Allison *et al.*, in yeast [156] and Corden *et al.*, [157] in mouse in the year 1985. The CTD performs numerous functions. A few among them are: transcription initiation, DNA binding, enzyme localization and modulation of enzyme activity [155,158,159]. The CTD consists of a heptapeptide repetitive sequence ($Y_1S_2P_3T_4S_5P_6S_7$), which is common from fungi to humans. The length of the CTD increases with genomic complexity; for example, the yeast-CTD has 26 repeats while the human-CTD has 52 repeats [155,160]. Within the heptapeptide, five residues can undergo phosphorylation, which equips the CTD to recruit different regulatory proteins and aid in signal transduction. Within the heptapeptide, the Tyr position is the most conserved. Recent studies have shown that several kinases mediate the phosphorylation of the CTD [161,162]. That the CTD is essential in transcription is illustrated by the fact that mutation or deletion of residues results in the inhibition of mRNA processing and proves lethal to the vitality of organisms [155].

1.4.3 Functionality of CTD in transcription:

The transcription process is heavily dependent on the phosphorylation of CTD [163]. At first, the promoter recruits the non-phosphorylated CTD, which forms a tight multi-component complex with the initiation transcription factors and the mediator complex [137]. The phosphorylation of the CTD induces a conformational change in RNAP-II, facilitating the interaction of the latter with elongation factors and other proteins involved in mRNA synthesis [164]. Phosphorylation is mainly carried out by three different cyclin-dependent kinases: CDK7, CDK8 and CDK9 [163]. Each individual kinase performs specific functions at different stages of transcription. A component of general transcription factor TFIIH contains the CDK7/cyclin-7 enzyme which phosphorylates the CTD during preinitiation complex formation. Generally, phosphorylation occurs on the serine-5 in the sequence ($Y_1S_2P_3T_4S_5P_6S_7$) and it is often observed that phosphorylation of serine-2 is facilitated by CDK9 [165]. CDK9 plays a crucial role in the elongation stage of transcription and in the activation of the transcription of HIV viral genome by interacting with Tat [137].

1.4.4 Liquid-liquid phase separation of the CTD:

Boehning *et al*., [146] and Zhou *et al*., [166] have shown that the carboxy terminal domain of RNAP II alone can undergo length-dependent *in vitro* liquid-liquid phase separation in the presence of a crowding agent. The same studies reveal that the phase-separation process is regulated by weak, multivalent interactions common to membraneless organelles [146]. Phase-separation of non-phosphorylated and phosphorylated CTD has been observed wherein the former undergoes phase-separation mainly driven by weak hydrophobic interactions [146] whereas, the latter phase-separates through electrostatic interactions [166]. IDRs in the phase-separated CTD

help in the binding of RNAP-II to partner proteins but, its phosphorylation status also has a significant influence on this binding process [167].

1.4.5 Carboxy terminal domains of RNAP II and Cyclin T1

Tat is a multifunctional protein [134] responsible for enhancing viral transcription upon binding to its partner proteins in the pTEFb complex (CDK-9 and Cyclin T1) and TAR RNA resulting in hyperphosphorylation of the CTD of RNAP II (see **Section 1.3.4**.) Thus, Tat is involved in the elongation phase of transcription. Interestingly, transcription initiation is also thought to be regulated by LLPS promoted by two low-complexity proteins FUS and TAF15 that can capture the CTD in a phase-separated liquid. Furthermore, under low salt conditions the long unstructured C-terminal domain of cyclin T1 also forms a phase-separated liquid that can solubilize the CTD of RNA polymerase *in vitro* [166]. The Histidine-Rich Domain (HRD) of cyclin T1 is responsible for this phase separation and its removal results in a loss of speckle formation in cells [166]. The same study showed that the HRD is essential for Tat/cyclin T1 activation of HIV-1 transcription elongation. These results strongly suggest that *in vivo*, Tat is localized in phase-separated Super Elongation Complexes [166].

1.5 Enhancing the solubility of HIV-1 Tat using a genetically engineered positively supercharged N-terminal Arginine (R_{10}) tag

Attractive and repulsive interactions between charged molecules, which are generally explained using Coulomb's law, play a crucial role in the functioning and interactions of biological molecules [168]. While the genetic material (DNA and RNA) is negatively charged, proteins could either be negatively or positively charged, depending on the side-chain functional groups present [168,169]. The net charge of a protein is determined by the fraction of cationic (Lys/Arg/His) and anionic (Asp/Glu) amino acid residues present in the sequence. Moreover, the labile-hydrogen containing amino acids (cysteine, SH; tyrosine, OH; histidine, NH) charge states are also susceptible to changes in the solution pH, and the net charge of protein will be pH dependent as well [168]. Supercharged proteins (SuP) are a sub-class of proteins, which encompass proteins bearing more than one net-charge per kDa of molecule [170]. Furthermore, these are classified as folded or unfolded, depending on the extent of folding [170]. Large numbers of natural SuP are disordered or unstructured, undergo direct phase separations and aid in maintaining ionic potentials in cells [171,172]. Due to the favorable attributes of supercharged proteins, significant efforts have been made to develop new genetically engineered biomolecules with enhanced properties and new functionalities [168]. Supercharged proteins have applications such as, assembly of bioliquids [173] intracellular formation of organelle-like compartments [174], artificial biological nano-containers [175], and many others.

1.5.1 Non-covalent interactions:

All proteins have basic (Arg/Lys/His) and acidic (Asp/Glu) amino acids, whose side-chain labile-hydrogens are in dynamic equilibrium with the hydronium ions (H_3O^+) in water. Van der Waals-dipole-dipole, dipole-charge, electrostatic, hydrophobic and hydrogen bonding interactions dictate the stability of protein structures and are generally referred to as weak interactions [176]. Electrostatic interactions, referred to as salt bridges, occur between the side-chain carboxyl oxygens of acidic and the nitrogen atoms of basic amino acids when they are in 4 Å proximity. While the electrostatic interactions are specific, the van der Waals interactions are non-specific, as they occur between dipoles formed due to the polarization of covalent bonds in a molecule. Electrostatic interactions regulate many processes pertaining to proteins such as,

a) Folding and stability of proteins [176,177]
b) Aggregation and solubility of proteins [178]
c) Ion transport *via* protein channels [179]
d) Electron shuttle between biological reaction centers [180]
e) Enzyme catalysis [181]
f) Protein-protein [182] and protein-nucleic acids [183] complex formation
g) Protein denaturation at non-physiological pHs [184]

1.5.2 Net-Charge density (NCD):

The net charge of a protein controls its solubility, aggregation, crystallization and electrophoretic mobility. It also regulates protein-protein associations [185], long-range electrostatic interactions [186] and enzymatic catalysis [187]. Charges could either be present on a protein surface, in an active site, or buried deep inside the protein structure. Many protein purification methods such as

ion-exchange chromatography and two-phase extraction are net-charge dependent [176]. The net-charge density (NCD, equation 1) of some of the highest net-charge unstructured cationic and anionic SuP and their functions are listed in **Tables 5** and **6** [168,169]. The net charge of a protein is zero at its isoelectric point (pI) and deviates to positive and negative values at pHs lower and higher than the pI, respectively. The net charge density of a protein is defined as follows [176]:

$$\text{NCD} = \frac{\sum(+ve) - \sum(-ve)}{N} \qquad (1)$$

where $\sum(+ve)$ and $\sum(-ve)$ are the total positive and negative charges, and N is the total number of amino acids present in the sequence.

The cationic and anionic supercharged-protein tables are topped by the sperm histone protamine (NCD = +0.58) and prothymosin-alpha (NCD = -0.38), respectively. The sperm histone is a small 62-residue protein but has 36 +ve charges, making it one of the highest net-charge density proteins. Prothymosin-alpha is a transcription factor involved in cell- cycle progression and cell proliferation. Upon binding to the bacterial membrane, the cationic antimicrobial peptides undergo a secondary structure transformation and concomitantly become amphiphilic.

Table 5. List of major cationic disordered Sup proteins obtained from UniProt databank [168,169].

Sl.no	Protein	Function	UniProt code	NCD*	Mw (kDa)	(+) AA#	(-) AA#	Organism
1	Sperm histone (protamine)	DNA Condensation	P15340	0.58	4.44	36	0	*Gallus gallus* (chicken)
2	Histone H5	DNA Condensation	P02259	0.32	2.95	66	5	*Gallus gallus* (chicken)
3	Histone H1.0	DNA Condensation	P10922	0.27	2.54	62	9	*Mus musculus* (Mouse)
4	Histone H1.2	DNA Condensation	P15865	0.27	2.68	66	*7*	*Rattus norvegicus (Rat)*
5	Genome polyprotein	Several	P06935	0.20	0.03	15	4	*HIV type-1*
6	Protein LLP	Transcriptional activator	B0FRH7	0.18	1.56	36	14	*Aplysia kurodia* (Kuroda's sea hare)
7	50S ribosomal protein L33	Protein synthesis	P0A7N9	0.18	1.56	15	5	*Escherichia coli*
8	Non-histone chromosomal protein HMG-17	Tuning DNA condensation	P02313	0.18	1.70	26	10	*Bos taurus* (Bovine)
9	Cyclin-dependent kinase inhibitor 2A [Isoform 3]	Negative regulator of proliferation	Q64364-1	0.18	1.56	38	8	*Mus musculus* (Mouse)
10	30S ribosomal protein S12	Protein synthesis	P0A7S3	0.17	1.53	28	7	*Escherichia coli*
11	Histone H1	DNA condensation	P53551	0.16	1.51	62	20	*Saccharomyces cerevisiae* (Baker's yeast)
12	Cathelicidin antimicrobial peptide (LL-37)	Antibacterial activity	P49913	0.16	0.31	11	5	*Homo sapiens* (Human)
13	30S ribosomal protein S18	Protein synthesis	P0A7T7	0.16	1.33	18	6	*Escherichia coli*
14	Beta-defensin 12	Antibacterial activity	P46170	0.16	1.46	6	0	*Bos taurus* (Bovine)

Table 6. List of major anionic disordered Sup proteins obtained from UniProt, databank [168,169].

Sl.no	Protein	Function	UniProt code	NCD*	No. of AA	(+) AA#	(-) AA#	Organism
1	Prothymosin alpha	Transcription factor (cell- cycle progression and proliferation)	P06302	-0.38	112	11	53	*Rattus norvegicus (Rat)*
2	26S proteasome complex subunit DSS1	Ubiquitin-dependent proteolysis	P60896	-0.31	70	5	27	*Homo sapiens* (Human)
3	Protein starmaker	Formation of otoliths in the inner ear	A2VD23	-0.24	613	70	216	*Danio rerio* (Zebrafish)
4	Cyclin nucleotide-gated cation channel beta-1 [Isoform GARP1]	Visual and olfactory signal transduction	Q28181-4	-0.20	590	46	165	*Bos taurus* (Bovine)
5	Calsequestrin-1	Internal calcium store in muscle	P07221	-0.20	395	32	111	*Oryctolagus cuniculus* (Rabbit)
6	Acyl carrier protein	Fatty acid biosynthesis	P0A6A8	-0.19	78	5	20	*Escherichia coli*
7	Prokaryotic ubiquitin-like protein pup	Marker for proteasomal degradation	P9WHN5	-0.19	64	7	19	*Mycobacterium tuberculosis*
8	Troponin C, slow skeletal and cardiac muscles	Striated muscle contraction	P63315	-0.18	161	17	46	*Bos taurus* (Bovine)
9	60S acidic ribosomal protein P1-alpha	Protein synthesis	P05318	-0.18	106	5	24	*Saccharomyces cerevisiae* (Baker's yeast)
10	Methylosome subunit pICln	Chloride conductance regulatory protein	P35521	-0.17	235	13	53	*Canis lupus familiaris* (Dog)
11	Bone sialoprotein 2	Integral part of mineralized matrix	P21815	-0.17	317	23	76	*Homo sapiens* (Human)
12	Calmodulin	Calcium signal transduction	P62152	-0.16	149	14	38	*Drosophila melanogaster* (Fruit fly)
13	Latent membrane protein 2A	Blocks tyrosine kinase signalling	A8CDV5	-0.16	118	4	23	Epstein-Barr virus (Human herpesvirus 4)
14	RWD domain—containing protein 1	Cell signalling	Q9CQK7	-0.16	243	24	62	*Mus musculus* (Mouse)

1.5.3 Genetically engineered Supercharged proteins:

Multiple factors such as pH, solvent composition and temperature influence protein aggregation. By increasing the total net-charge of a protein, conditions unfavorable for protein aggregation can be induced. The net-charge of an SuP can be modified using the following methods:

(a) Solvent-exposed site-specific mutagenesis:

This method involves chemical modification of solvent-exposed amino acids [168,188] wherein, the negatively charged residues are switched with positively charged residue or *vice-versa*. Variants of SuP can also be synthesized by replacing neutral amino acids with charged residues. In both cases, proteins will be expressed but their functionality is not guaranteed as the substitutions can distort electrostatic interactions, thus causing intramolecular repulsion which destabilizes the protein structure [168,169]. For example, variants of Green Fluorescent Protein (GFP) ranging in NCD were synthesized by Lawrence *et al.* [188] (**Figure 16)**.

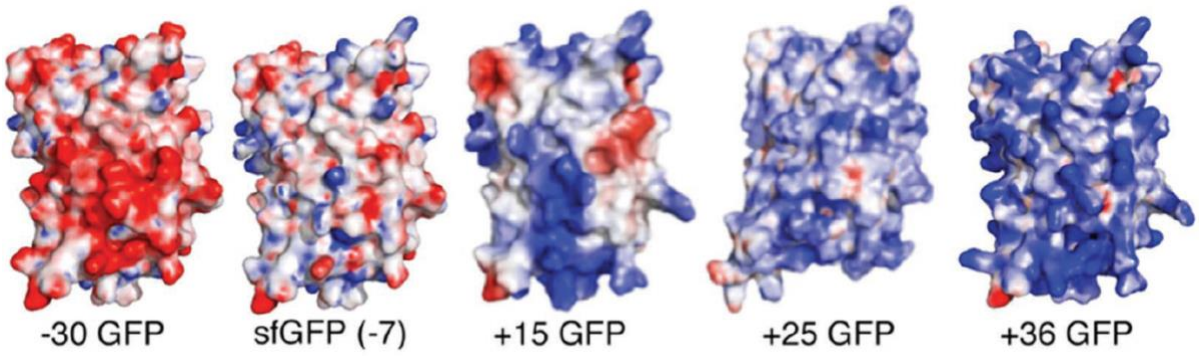


Figure 16. Genetically engineered variants of supercharged green fluorescent protein. Blue: Negatively charged GFP; Red: Positively charged GFP; sfGFp: super folder green fluorescent proteins. Adapted with permission form reference [189]. *Copyright 2009 © National Academy of Science, USA.*

The fluorescence and bioluminescence properties of the green fluorescent proteins (GFP) have been extensively studied in the recent years due to their applications in gene-tagging in recombinant protein production, biosensors, to probe protein-protein interactions and many others [190,191]. GFP is a 27 kDa protein, made of 238 amino acids, and is extracted from the crystal jellyfish *Aequorea victoria* [190]. The name represents its inherent fluorescence emission property, that matches the green wavelength of the visible spectrum when exposed to white light [191]. The crystal structure of GFP shows an unusual motif, containing 11 β-strands and a solvent-unexposed short distorted α-helix segment, which is a chromophore made of three specific amino acid residues, Ser, Tyr and Gly at positions 65, 66 and 67 in the sequence [190,191]. The fluorescent property is independent of any external factors such as a chromophoric co-factor, enzyme or substrate. In its folded state, the chromophore property is gained when the central helix undergoes autocatalytic dehydration, cyclization and oxidation in the presence of molecular oxygen [190]. The amino-terminus of GFP is short ($1-81$ residues) composed of three β-strands and a central helix and the carboxy terminus is a long polypeptide chain ($81-238$ residues) that consists of eight β-strands arranged in unique "Greek key" fashion [190,191].

(b) Post-translational chemical modification:

In this method, the net charge is modified by carrying out different chemical reactions involving side-chains of positive (Lys) and negative (Asp/Glu) amino acids. In lysine, the charge is neutralized [$(+1)\rightarrow 0$] or inverted [$(+1)\rightarrow (-1)$] by the acetylation or succinlyation of ε-amino groups, respectively (**Figure 17a**). For acidic amino acids, the charge is neutralized [$(-1)\rightarrow 0$] or inverted [-1 to $+1$] by the amidation/amination of the side-chain carboxylic group (**Figure 17b**)

[168,169]. Similar to the first method, unfavorable electrostatic interactions might cause enhanced protein aggregation.

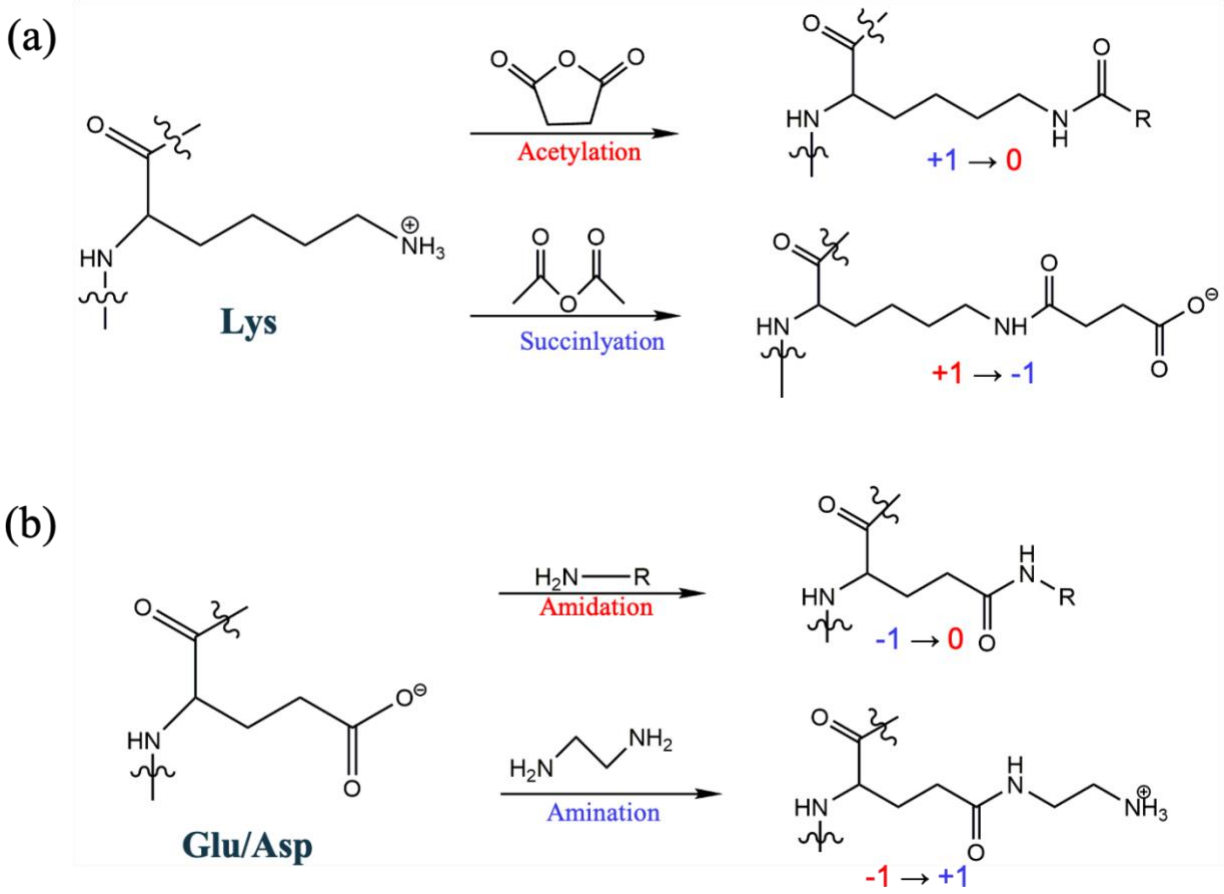


Figure 17. General scheme of post-translational chemical modification of (a) positively and (b) negatively charged amino acids. Based on reference [168].

(c) Supercharging by Unification of highly charged unstructured polypeptide tag:

In this third method, the net-charge of a protein is drastically increased by attaching a highly charged unstructured polypeptide tag. Since the native sequence of the protein is not disturbed, this method can be highly effective. The supercharged polypeptide chains with basic amino acids

are categorized as cell penetrating peptides (CPPs) and are used as therapeutics [192]. They also enhance the stability and solubility of proteins prone to aggregation [192].

1.5.4 Positively charged residues as cell penetrating peptides (CPPs):

Regulation of the absorption of albumin by tumor cells in culture by histones and long, positively charged polylysine was first observed by Ryser *et al.* [193] in 1965. A survey of supercharged cationic homopolymers revealed that medium-sized arginine polymers are potent in traversing the cell membrane compared to other polymers made of lysine, ornithine and histidine [194]. The integrity of biological membranes is essential to maintain homeostasis in tissues [195]. Among these membranes, the blood-brain barrier (BBB) is extremely important; it is permeable to lipid-soluble substances such as O_2, CO_2 and small hydrophobic molecules but blocks the transportation of disease causing bacteria, viruses and other large macromolecules [195,196]. As the BBB prevents the permeation of therapeutic drugs, treating neurological disorders is a big problem. However, if a drug is tagged with a cell penetrating peptide, it can be transported across the membrane easily and hence, many neurological or non-neurological disorders [195] might be treated effectively. The cell-penetrating peptides are short (<30 amino acids), water-soluble, cationic and amphiphilic (**Table 7**). Due to their transporting ability, they are also known as protein transduction domains [197].

The valuable property of CPPs is that they can traverse the cell membrane (both, *in vivo* and *in vitro*) even at low concentrations without any help form external receptors and without harming the host [198]. The non-toxic CPPs have the upper hand in therapeutic delivery, in contrast to cytoplasmic delivery methods that utilize liposomes and encapsulation [199]. In physiological conditions, CPPs with basic amino acids have large NCD and because of this, they

easily penetrate the plasma membrane [197]. This penetrating property is brought about either by the receptor-independent electrostatic interactions between the CPP and motifs on the cell membrane [197] or by the amphiphilic nature of the CPP [199].

Table 7. List of natural and synthetic cell-penetrating cationic peptides.

Cationic CPP	Sequence	Length	Origin	References
Tat peptide	RKKRRQRRR	9	Protein derived	[200]
DPV3	RKKRRRESRKKRRRES	16	Protein derived	[201]
DPV6	GRPRESGKKRKRKRLKP	17	Protein derived	[201]
Penetratin	RQIKIWFQNRRMKWKK	16	Protein derived	[202]
R8	RRRRRRRR	8	Synthetic	[203]

1.5.5 Tat peptide (Tatp):

The sequencing of the trans-activator of transcription protein revealed that it consisted a polycationic region called *Tatp* (residues, 49−57), which is responsible for the penetration of the protein into mammalian cells [204]. A synthetic Tat peptide (residues, 38−60) was synthesized in 1991, which exhibited an α-helical amphipathic conformation bearing residues 38−45 and an arginine-rich, random coil conformation involving residues 49−57 [205]. Deletion of arginine residues in the 9mer fragment reduced the cell penetration ability of the peptide [206], which underscores that the basic random coil conformation is more important than the α-helical amphipathic conformation for cellular peptide uptake [207]. Upon deletion of one Arg residue, the cell penetration capability of the peptide was reduced by 50% and the deletion of two residues decreased the activity by 75% [206].

Various cargo molecules of different sizes have been transported across the cellular and tissue barriers using Tatp as a tag [251]. A list of molecules that have been conjugated with the Tatp for transportation across cell membranes are listed in **Table 8**.

Table 8. List of cargo transported by Tatp and their applications.

Categories	Cargo	Application	Reference
Small molecules	Imaging agents and paramagnetic labels	Molecular imaging	[208,209]
Antibodies	Fab fragments, Toxin fragments	Tumor therapy, Neuroprotection against neurotoxins	[210,211]
Peptides and proteins	Exogenous protein; Recombinant antigen; Apoptin	Protien-based vaccine administration; Dendritic-cell-based-immunotherapy; Selective cancer cell apoptosis	[212–214]
Liposomes	Plain and PEGylated liposomes as carriers for drugs and DNA	Tumor studies	[215]
Nanoparticles	Dextran-coated superparamagnetic iron oxide particle; Superparamagnetic-derivatized nanoparticles	Cell magnetic labelling though magnetic resonance imaging (MRI)	[216]
Polymer-based approach	Polymer bound anticancer drug (doxorubicin)	Human ovarian carcinoma cell drug delivery	[217]

1.5.6 Solubility enhancement peptide (SEP) Tags:

The solubility of proteins and other biological molecules is a complex phenomenon, essential for many physical and biological property measurements. High concentrations of pure soluble protein are required for protein structural analysis. Different solvent conditions like pH, temperature and salt concentrations are experimentally determined to derive optimal conditions

for protein solubility [218]. Proteins can be solubilized using solubility enhancement tags, for instance, maltose binding protein [219] and glutathione S-transferase [220]. Sometimes, due to the large sizes of SEP tags the characteristic structure and functions of proteins could be perturbed by tagging and protein yields can be reduced owing to the metabolic burden of larger tags. Yutaka Kuroda *et al.* [218] used small charged-peptide tags (poly-Lys or poly-Arg) for addressing the solubility issue of bovine pancreatic trypsin inhibitor variant (BPTI-22). They tested many combinations of positively charged residues as a tag on both termini (C or N) of the protein sequence. For example, they tested 3R/K, 4 R/K and 6 R/K. To prevent direct interactions between the protein sequence and the tags, two glycine residues were added, known as spacers. The clusters of charged residues can induce repulsive electrostatic intermolecular interactions and can hinder the protein aggregation. Moreover, small peptide tags are more effective for small sized proteins. The addition of small peptide tags significantly enhanced the solubility of BPTI-22 variant (contains 22 alanines) [222]. They also showed that high-quality NMR spectra could be obtained at protein concentrations where BPTI-22 is normally aggregated [222]. **Table 9** shows some of the examples of tagged supercharged-proteins with improved solubility.

Table 9. List of solubility enhancement charged peptide tag.

Sl. No.	Protein	Tag	Solubility	Cysteine Residue	NMR	Enhance expression (soluble) and purification	Crystallization	Reference
1	Minibody	N-K3/C-K3	10 μM to 1 mM	−	Yes	−	−	[221]
2	BPT1	No Tag C-5K/C-5R	2-3 mg/mL 8-10 mg/mL	−	No	Yes	Yes	[222,223]
3	BPT1-22	No tag C-R6 C-R5 N-5R	1.7 mM 10.59 mM 8.23 mM 6.20 mM	−	Yes	−	−	[218]

1.6 Sequence-Specific Nickel Assisted cleavage (SNAC) of Tat

In recombinant protein expression, different tags are attached to amino or carboxy termini to enhance protein stability, solubility, folding, purity and expression yield [224]. Once the protein is purified, cleaving the purification tag from the target protein is normally an important option to avoid structural or functional effects of the tag [225]. At present, the only biocompatible approach is by using proteases such as the tobacco etch virus (TEV) protease and the thrombin protease [226] . This adds one more purification step and the enzymes are extremely expensive when used in large-scale protein production. Often, enzyme cleavage is not successful in membrane proteins when the enzyme is in close proximity to hydrophobic domains [226]. One other approach is chemical cleavage by using cyanogen bromide at the carboxyl side of Met residues but this method uses rather harsh conditions at low pH [225]. A third approach, where the attached tag undergoes cleavage with the assistance of a metal ion was recently proposed by Dang *et al*. [227] in 2019.

This is called the SNAC tag, which is a short five-residue peptide with the sequence, GSHHW. When the tagged protein is is exposed to nickel ions, the peptide bond between the glycine and serine residues are cleaved [228], releasing the full-length protein. Since the tag is short its synthesis does not impose a metabolic burden on the cell. Placement of the tag on the C-terminus of a protein enables the release of the target protein with its native sequence but for the addition of a single Gly at the C-terminus of proteins that do not normally have a Gly at their C-terminus. Since purification tags can interfere with the folding and function of a protein we decided to prepare Tat with a C-terminal His_6 SNAC purification tag.

1.6.1 SNAC Tag cleavage Mechanism

The cleavage mechanism is explained by the N-to-O acyl shift, which occurs with the assistance of the Ni^{2+} metal ions. This involves migration of the carbonyl group of Gly (P_1) to form a new bond with the side-chain hydroxyl group of serine (P_2) (**Figure 18**). It is important to have Gly in the first position due to its small steric properties, as residues other than Gly result in slower hydrolysis [227]. The protein sequence (SNAC-Tat), buffer and cleavage conditions are explained in **Section 3.3**.

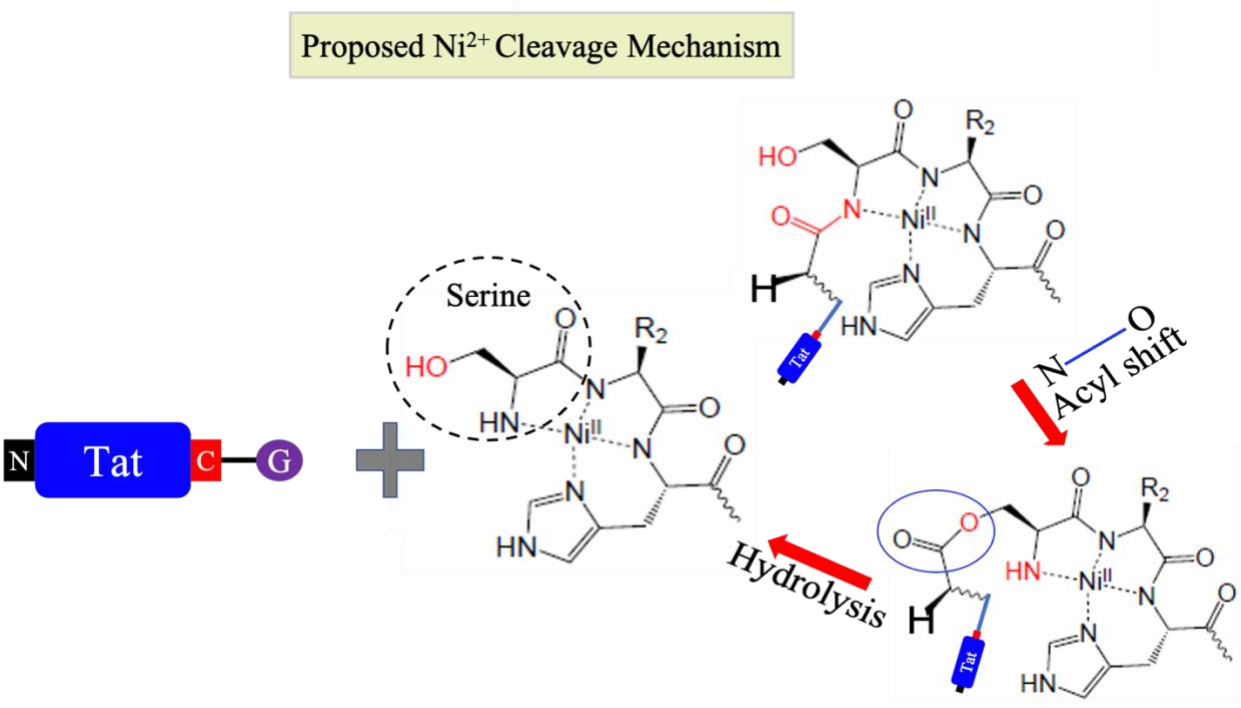


Figure 18. Schematic representation of Ni^{2+}-assisted cleavage of SNAC-tagged sequences. The blue circle highlights the site of hydrolysis responsible for the cleavage of the SNAC tag [227] .

1.6.2 SNAC Tag Cleavage Measurement by Mass spectrometry

Mass spectrometry (MS) is an important analytical tool used in all streams of chemistry, which measures the mass-to-charge ratios of intact molecules and molecular fragments yielding their molecular weights [229]. The m/z ratio (where m is the mass of the analyte and z is its total charge) is measured in the vacuum of the mass spectrometer and the obtained spectrum is a plot of ion abundance *vs* m/z ratio with units of Daltons (Da) per unit charge [229]. In brief, the mass spectrometer consists of three main components: an ionization source, mass analyzer and a detector. Standard ionization methods such as electrospray ionization (ESI) and matrix-assisted laser desorption/ionization (MALDI) are used to generate charged species for analysis and detection. Different mass-analyzing methods routinely used are time-of-flight (TOF), triple-

quadrupole mass filter, quadrupole ion-trap and Fourier-transform ion-cyclotron resonance (FT-ICR). In-depth information on the technique, its functioning and applications are given in these reviews [229,230].

2. Biophysical characterization of Intrinsically disordered proteins

Recent advancements in analytical techniques have unearthed the crucial roles of IDPs in many life-critical biological processes. The lack of functional dependency on the three-dimensional structure makes IDPs attractive but difficult to study due to their dynamic structural characteristics. Although there are many techniques that have been put to work to understand the structural aspects of disordered proteins, fluorescence, circular dichroism (CD) spectropolarimetry, Nuclear Magnetic Resonance (NMR), and Fourier-Transform Infrared spectroscopy (FTIR) techniques remain among the most powerful due to the unique insights they offer.

2.1 Fluorescence of proteins

Fluorescence is a luminescent process that occurs after some atoms and molecules undergo electronic excitation by absorbing light at specific wavelength [231]. After reaching the highest vibrational energy of the excited state, the molecules lose energy rapidly due to intermolecular collision and relax to the lowest energy level of the excited state, which is referred to as internal conversion (IC). Later, the molecules fully relax to the ground electronic state by emitting energy, which is referred to as *fluorescence* (**Figure 19**). Reduction in the fluorescence intensity of a sample is called *fluorescence quenching* [231]. Different molecular interactions result in quenching such as, excited-state reactions, molecular rearrangements, energy transfer and molecular collision. A broad range of molecules act as fluorescence-quenchers, and among them, molecular oxygen is the best-known collisional quencher [232].

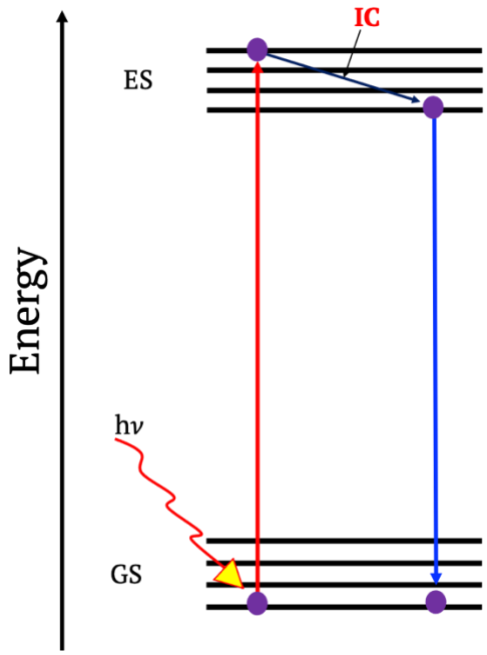


Figure 19. The graphical representation of the fluorescence process. GS-ground state; ES-Excited state; IC- internal conversion.

Many biomolecules, lipid membranes and polysaccharides are non-fluorescent, whereas proteins fluoresce due to phenylalanine, tyrosine and tryptophan amino acids. This intrinsic property of proteins is utilized in protein folding studies [233]. A typical fluorescence spectrum of a protein is excited at 280 nm or longer wavelength and the emission of Tyr and Trp is observed between 300−350 nm in water [231,233]. The quantum yield of these three residues is in the order Trp > Tyr > Phe [231,233]. Tryptophan is the more fluorescent among the three but its concentration in proteins is generally low. However, its well-resolved emission at 350 nm is highly sensitive to the residue's local surroundings therefore, even small changes in the spectrum may be assigned to conformational changes, ligand binding or denaturation [231]. Phenylalanine is not generally excited due to its low quantum yield, so the fluorescence emission from Phe is rarely

seen. Tyrosine emission is usually quenched in folded proteins, possibly due to interaction with the peptide chain or transfer of its energy to a tryptophan.

Tryptophan fluorescence is highly sensitive to the local environment of the amino acid, and at least two different fluorescence lifetimes (~0.5 and ~3.1 ns) have been observed and both are inherent to the Trp structure [234]. It is hypothesized that the different emission lifetimes result from emission from nearly identical electronic absorption transitions (^1L_a and ^1L_b states). Alternately, a widely accepted rotamer model argues that different lifetimes (ms and ns) observed for the tryptophan is a result of structural heterogeneity, *i.e.*, rotameric structures about the $C\alpha - C\beta$ bond. The current consensus is that the tryptophan emission occurs only from the ^1L_a state, except when the surrounding environment is nonpolar [231]. Thus, the emission from tryptophan is greatly affected by local surroundings and occurs from both ^1L_a and ^1L_b states, depending on the polarity of the solvent. If the tryptophan fluorescence emission occurs from the ^1L_a state the fluorescence spectrum shifts towards longer wavelengths (red shift) due to an increase in polarity of the environment; the emission from the ^1L_b state dominates in hydrophobic environments, meaning that tryptophan fluorescence shifts towards shorter wavelengths (blue shifts) due to increasing non-polar environments. For example: the Trp-45 in Azurin (a native folded protein) is located deep in a hydrophobic pocket and shows a maximum emission (λ_{EM}) at 308 nm [235]. In contrast, tryptophans in denatured proteins like glucagon and melittin, are solvent exposed with their emissions (λ_{EM}) at 352 and 346 nm, respectively [231].

2.2 Fourier Transform Infrared Spectroscopy (FTIR)

Infrared spectroscopy is a sensitive and powerful technique in determining molecular structures wherein, each functional group in a molecule vibrates at a specific frequency in the infrared region

(400 to 4000 cm^{-1}). The relative intensities of the peaks in an IR spectrum serve a dual purpose of identifying the functional groups and quantifying them. Infra-red spectroscopy allows us to understand the secondary structure and different conformational ensembles of a disordered protein [236]. This approach has wide applications in protein science, ranging from small soluble proteins to large membrane proteins. Often, this method requires a very short measurement time, low amounts of sample (10-100 µg), and is cost-effective [237]. This technique is already in use for studying the secondary structures of globular or folded proteins. Fourier Transform IR (FTIR) spectroscopy is also used in conjunction with other popular techniques such as CD spectroscopy [237]. Prominent IR absorption bands that are extensively used for characterising proteins are the **amide I** (1600-1700 cm^{-1}), **amide II** (1500-1600 cm^{-1}) and **amide III** (1200-1400 cm^{-1}) bands (**Figure 20**). Multiple secondary structures resulting from various hydrogen bonding patterns show one low-intense broad band in the amide I region. These broad bands hide important structural information which is mathematically extracted by Fourier self-deconvolution [238,239]. **Table 10** highlights the amide I vibration band regions and wavenumber ranges pertaining to different protein secondary structures. The region consists of vibrations from amide C=O groups which are highly sensitive to the strength of hydrogen bonding. It is not affected by the side-chains of amino acids and depends only on the backbone secondary structure, making it important in the analysis of the secondary structure of proteins [236]. The amide II region (1600-1500 cm-1) consists of vibrations from NH in-plane stretching, CN stretching and some vibrations from C=O bending. This region is used to study the flexibility of proteins [239]. The third amide region (1400-1200 cm^{-1}) represents in-plane NH bending, CN stretching and minor contributions from C=O bending [236].

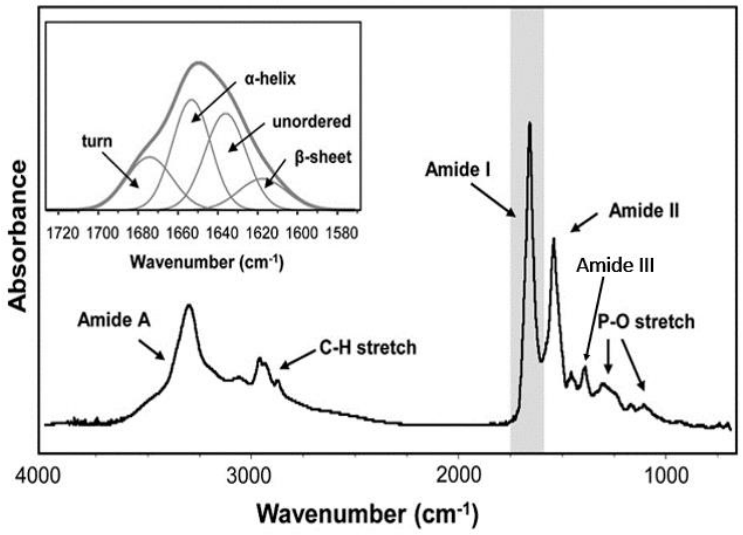


Figure 20. FTIR spectrum of a typical protein (Adopted from reference [240]).

Table 10. Amide band I for assignment of secondary structure [236,241].

Secondary structure	**Amide band I position (cm^{-1})** Solvent: H20	
	Average	Range
Turn	1672	1686-1662
α-Helix	1654	1660-1648
Random coil	1654	1657-1642
β-Sheet	1633	1640-1623
β-Sheet	1625	1630-1620

2.3 Nuclear Magnetic Resonance spectroscopy

The physical and chemical properties of all organic and inorganic materials is a direct consequence of their structures, *i.e.*, the way the atoms are arranged in the molecules and the modes of bonding between them. For instance, ethanol or and dimethyl ether are analogous in the number and the types of atoms present, yet they exhibit significant differences in their properties, attributed to their different structural configuration (**Figure 21**). While ethanol is a liquid at room temperature, dimethyl ether is a poisonous gas.

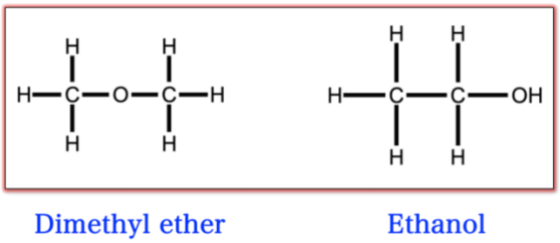


Figure 21. Chemical structures of dimethyl ether and ethanol.

Over the years, many spectroscopic techniques have been invented to elucidate the structures of a diverse class of molecules. Amongst these spectroscopic methods, Nuclear Magnetic Resonance (NMR) spectroscopy has emerged as a major and powerful technique, which was independently developed by Bloch and Purcell in 1945 [242,243] for which, they were awarded the Nobel prize in 1952. It is a versatile tool and has been a dominant analytical technique in the fields of Chemistry and Biology with far reaching applications. As it provides atomic-level structural details, it has paved the way for understanding the kinetics, structure, dynamics, thermodynamics and interactions involving both small and large biomolecules [244,245]. It is by far the most widely adopted technique in studying the structure of inorganic materials, as a majority of the elements in the periodic table are NMR accessible. Magnetic resonance imaging, which has the same working

principle as NMR has revolutionized the field of medicine. Its capability in observing spatio-temporal images of biological tissues and fluids has helped tremendously in identifying life-threatening conditions and in their diagnosis [246,247].

As proteins catalyze many life-critical biological reactions, studying their structures becomes all the more important. Since proteins are macromolecules, only a few techniques are equipped to discern their structural information and NMR is undoubtably one of the most important. Although, x-ray diffraction has positioned itself as the standard technique in studying protein structures, its insensitivity towards macromolecules with non-periodic atomic configurations and the requirement of samples to be in a crystal-form for data acquisition have hindered the structural analysis of many biologically relevant proteins. Moreover, the structural knowledge derived pertains to one local minimum on the potential energy surface of the probed molecule, which may or may not be the physiologically relevant structure of the molecule. NMR spectroscopy on the other hand, is capable of determining structure and dynamics information on small proteins dissolved in water. In-depth insights into the biologically active conformation of the molecule and its structural dynamics on the 10^{-3} to 10^{-10} s range are obtained [244][248], which underscores the utility of the technique.

The sensitivity and resolution of NMR are heavily dependent on the magnetic field at which the samples are analyzed. Due to lack of adequate technology, high-field spectrometers were practically impossible, and this technique was limited to studying proteins with molecular sizes less than 10 kDa. However, revolutionary technological advances such as superconducting magnets and great advancements in the electronics and semi-conductors have expanded the scope of the technique that is now routinely used to study molecules of sizes up to 100 kDa (**Figure 22**) [249]. The NMR contributions to the protein database, a library of protein structures has

significantly increased over the years (**Figure 23**), and has reached a total of 13389 structures as of May 17, 2021 [250].

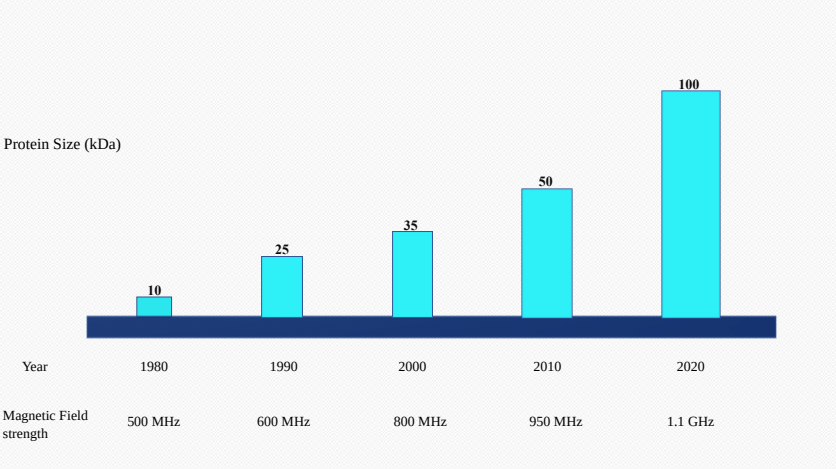


Figure 22. An account of the increase in the analytical magnetic field and the sizes of proteins studied by NMR over the decades [249].

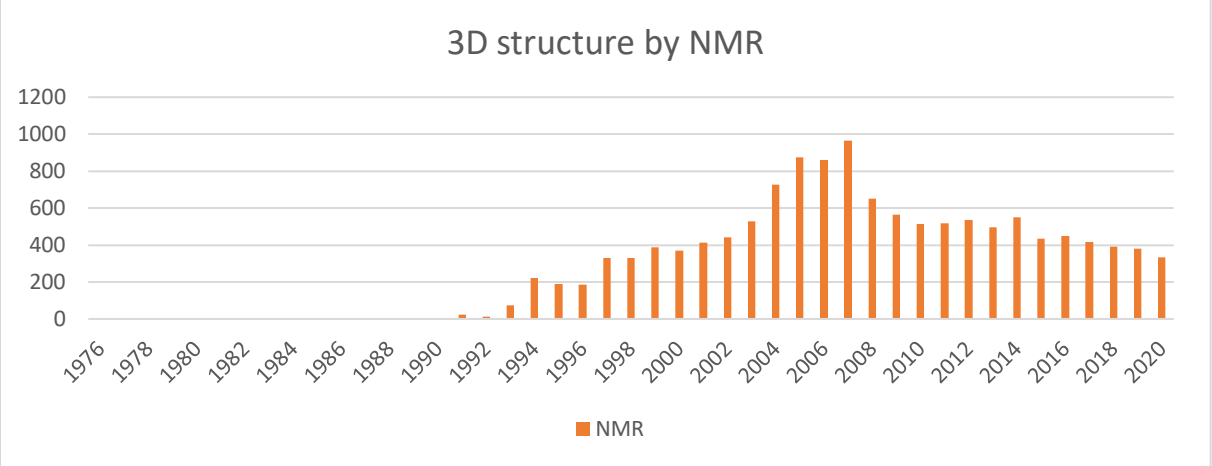


Figure 23. The number of three-dimensional protein structures deposited in the protein database over the years [250].

2.3.1 Fundamentals of NMR:

All atomic nuclei except those with an even number of protons and an even number of neutrons have an intrinsic quantum mechanical property called *spin*. Owing to their mass and spin they also possess a spin angular momentum and an associated nuclear magnetic dipole moment, μ. The nuclear spin is characterized by the spin quantum number I, which can have values of 0 and $\frac{n}{2}$ (n = 1 to 9). The magnitude of angular momentum vector (S) is given by equation 1.1 [251], where $\hbar$ is the reduced Plank constant.

$$S = \hbar(I\,(I+1))^{1/2} \qquad 1.1$$

The nuclei with I = 0 are NMR inactive and those with values of $\frac{1}{2}$ and higher are observable by NMR. Nuclei with non-integer nuclear spins have a non-spherical charge distribution and are called quadrupolar. The nuclear spin is a quantized physical constant, and the nuclei can adopt 2I+1 allowed spin states, described by the magnetic quantum number, m_I. The spin states will have integral values of +I to -I, with integer steps between the states. For example, spin states in ^{1}H which has a spin of ½ are -1/2 and +1/2. Similarly, the ^{11}B nucleus has a spin quantum number of 3/2 and consists of four spin states, -3/2, -1/2, +1/2 and +3/2. In the absence of a magnetic field, these spin states are degenerate but when a field is applied (B_0), the degeneracy is lifted, and the nuclei populate different spin states. This is called the *Zeeman interaction* or *Zeeman splitting*. For a spin ½ nucleus with two spin states +1/2 and -1/2 (α and β spin states, respectively), the nuclear magnetic moment opposes the applied field and is slightly higher in energy compared to +1/2, which is aligned with the field (**Figure 24c**). Although the energy difference is minute, it is the fundamental reason behind NMR spectroscopy. If all spin states were isoenergetic, there would be no macroscopic magnetization.

To understand the concept of bulk magnetization, individual nuclei may be considered as tiny bar magnets. Since the angular momentum is a vector, the tiny magnets may now be dubbed as individual magnetization vectors. In the absence of a magnetic field, all possible orientations of the magnetization vectors are present leading to no bulk magnetization, represented as M_0. When the field is applied, the direction of the field defines the z-axis along which the spins orient (**Figure 24a**). Although the nuclear magnetic moments of oppositely aligned spins cancel out, a difference in their population governed by the Boltzmann distribution (equation 1.2) [252] exists, resulting in a net magnetization aligned with the applied field.

$$N_\beta/N_\alpha = 1- (\gamma\hbar Bo/2\pi\ kT) \qquad 1.2$$

where N_β/N_α is the ratio of the population of two spin states, γ is the gyromagnetic ratio (μ/I), $\hbar$ is the reduced Planck's constant, Bo is the applied magnetic field, k is the Boltzmann's constant and T is the temperature. The population ratio is directly proportional to the applied field and inversely proportional to the temperature. Hence an increase in the magnetic-field strength and a decrease in the system temperature increases the population difference. Furthermore, the energy difference between the quantized spin states scales with the applied magnetic field and is expressed as :

$$\Delta E= (1/2\pi)\hbar\gamma B_o \qquad 1.3$$

In a magnetic field of strength B_0, the nuclear magnetic moment will precess about the axis of the field at a certain frequency called the *Larmor frequency* or *precession frequency* represented by ω_0 (equation 1.4) [253], which increases linearly with the strength of the magnetic field (**Figure 24b**).

$$\omega_0 = \gamma B_o \qquad 1.4$$

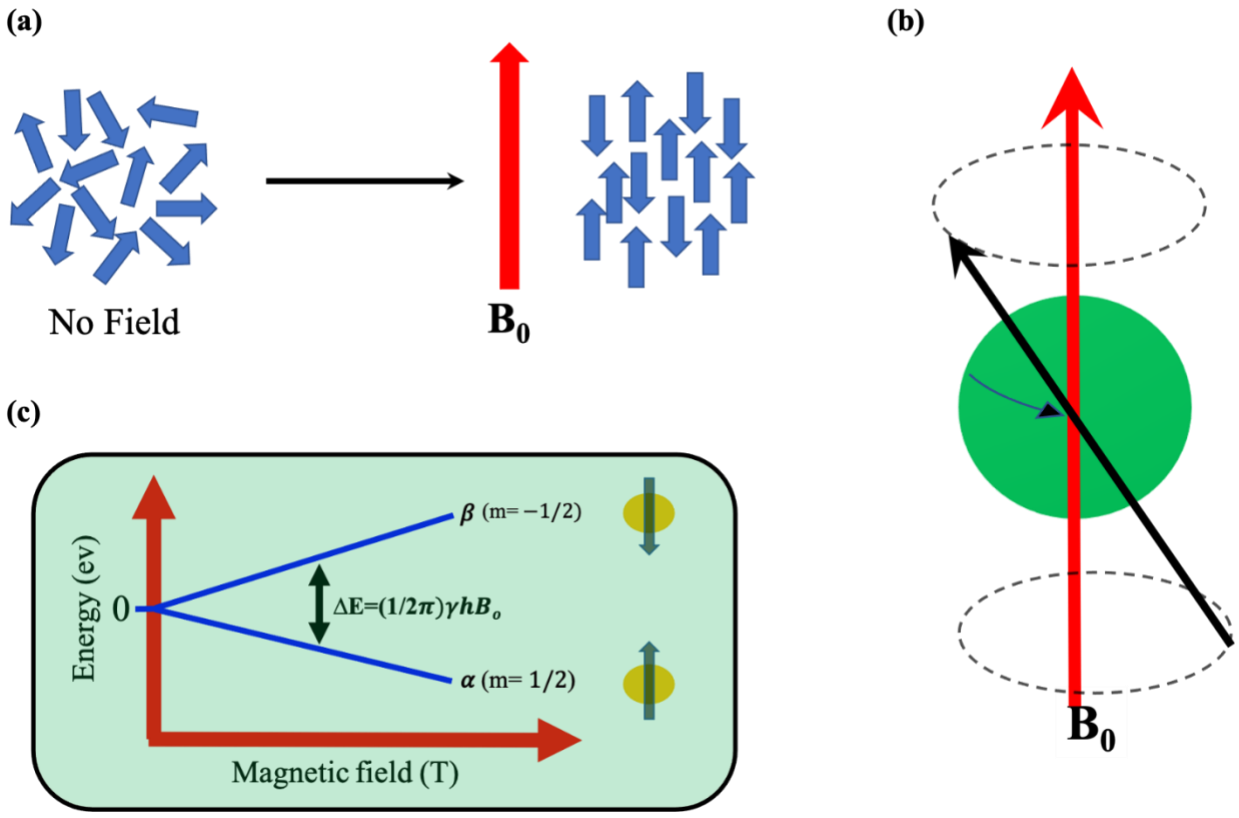


Figure 24. (a) The effect of the external magnetic field on spin coherence; (b) Nutation of a nuclear spin aligned with the applied magnetic field B_0; (c) Illustration of Zeeman splitting wherein, an increase in the energy difference between two spin states is demonstrated. Based on reference [253].

In order to understand the origin of the NMR signals, the bulk magnetization vector is placed in a cartesian coordinate system where the direction of the applied field defines the z-axis (**Figure 25**). The Larmor precession of the magnetization vector along the z-axis does not produce any observable signal. Its position can be manipulated by applying an electromagnetic radiofrequency (RF) pulse wherein the M_0 interacts with the magnetic component of the

electromagnetic radiation (B_1). The RF pulse is applied orthogonal to the main field (x- or y-axis) and when its energy matches with the energy difference between the two spin states, the M_0 will become sensitive to the new field B_1 and the condition is called *on resonance*. Provided the RF pulse duration is short, the M_0 now precesses about the B_1 and is tipped from z-axis to the xy-plane (**Figure 25**). The angle of rotation about the x or y axis, also called the *flip angle*, depends on the duration of the pulse applied and follows the relation:

$$\theta = \gamma B_1 t_p \qquad 1.5$$

where θ is the flip angle, γ is the magnetogyric ratio of the nuclei, B_1 is the field applied along the x-axis and t_p is the duration of the pulse in μs. A case of 90 and 180° rotation of the M_0 is shown in **Figure 25**. The precession of magnetisation in the xy-plane at its Larmor frequency induces an electromotive force (EMF) in the detection coils, which is recorded as a sinusoidal oscillating voltage. This induced current does not persist forever and will diminish as the magnetization vector precesses back to its equilibrium position. The acquired time-domain free induction decay (FID) is digitized, and Fourier-transformed to obtain a frequency spectrum (**Figure 26**).

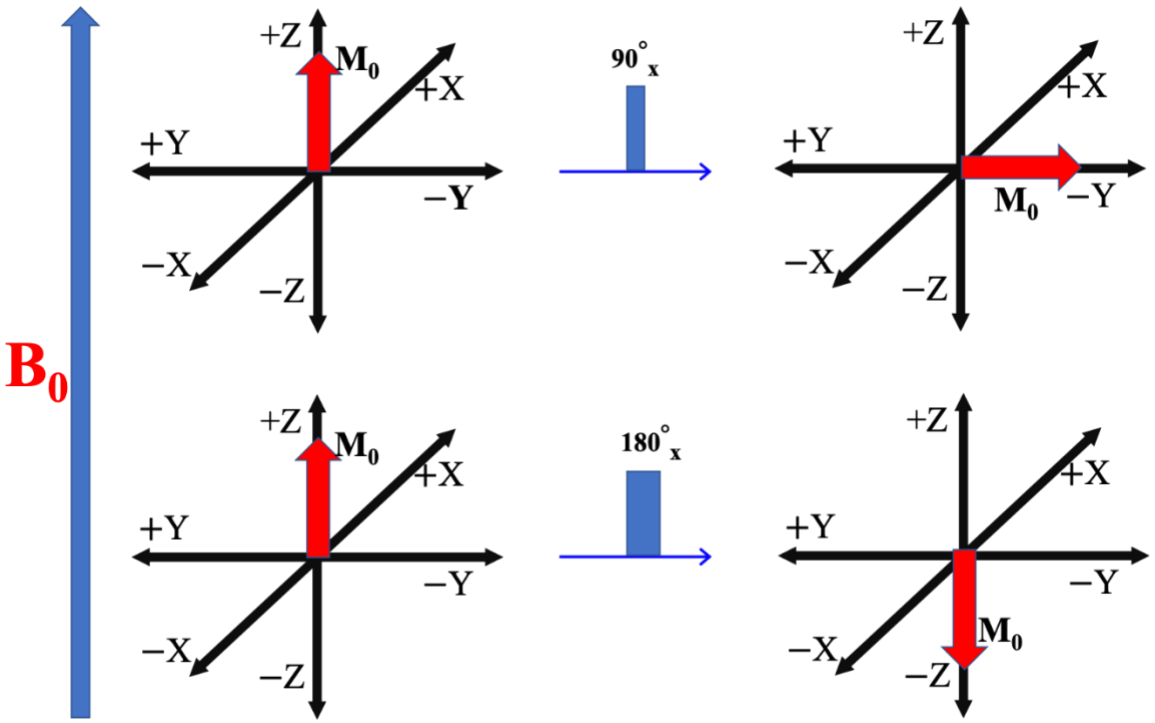


Figure 25. Upon applying rf pulses ($90°_x$ (top) and $180°_x$ (bottom)) the B_0 aligned magnetization (M_0) is re-aligned from the +Z-axis to the −Y ($90°_x$) or −Z axis ($180°_x$).

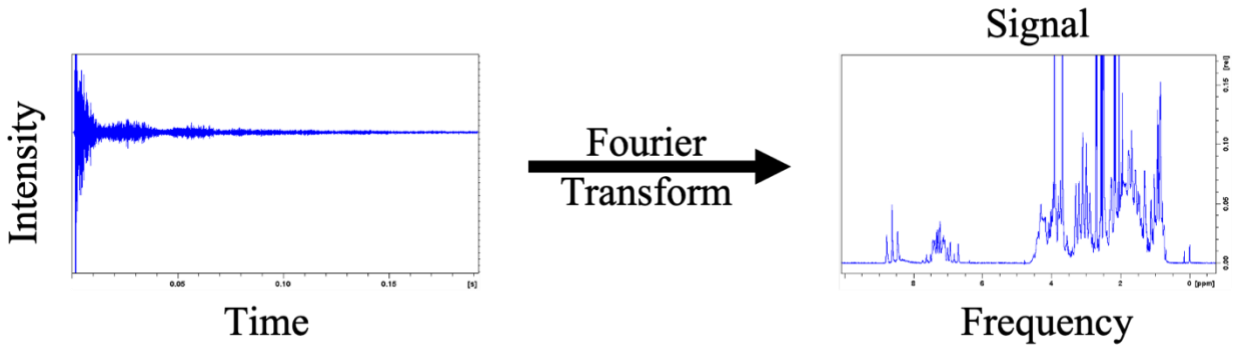


Figure 26. The extraction of frequency domain signals from raw time-domain data by performing a Fourier transformation.

2.3.2 NMR relaxation

As the Boltzmann spin distribution is conserved, the magnetization vector that was flipped to the xy-plane, or in other words excited, must return to its equilibrium position about the z-axis, and this process is called *relaxation*. There are two types of relaxation mechanisms involved: *longitudinal* relaxation and *transverse relaxation*, also referred to as T_1 and T_2 relaxations, respectively. In T_2 relaxation, the precessing magnetization vector flipped from z-axis to the xy-plane loses its amplitude over time due to the loss of coherence between the spins. On the other hand, in T_1-relaxation, the magnetization is reset to its equilibrium position along the z-axis (**Figure 27a**). The T_1 relaxation is also referred to as the *spin-lattice relaxation* as the energy is exchanged between the spins and the surroundings wherein the energy is lost to translational, vibrational and rotational degrees of freedom of the molecules. The T_1 decay of the magnetization follows the equation [252,253] :

$$M_t = M_{max}(1\text{-}e^{-t/T1}) \qquad 1.6$$

where M_t is the amplitude of the magnetization vector M_0 at time t, M_{max} is the maximum amplitude at full recovery and T_1 is the longitudinal relaxation time. At a time of one T_1, the recovery of the magnetization is 63%, which reaches to 95% when t = 3 T_1. As the amplitude of M_0 directly affects the NMR signal intensity, the recycle delay between the transients in a multi-scan experiment must be set to $3-5$ times the T_1 to allow enough time for the M_0 to fully relax. Since T_1 relaxation is a consequence of the interactions of the spins with the surrounding environment, structural and dynamics information about the molecules can be obtained [254].

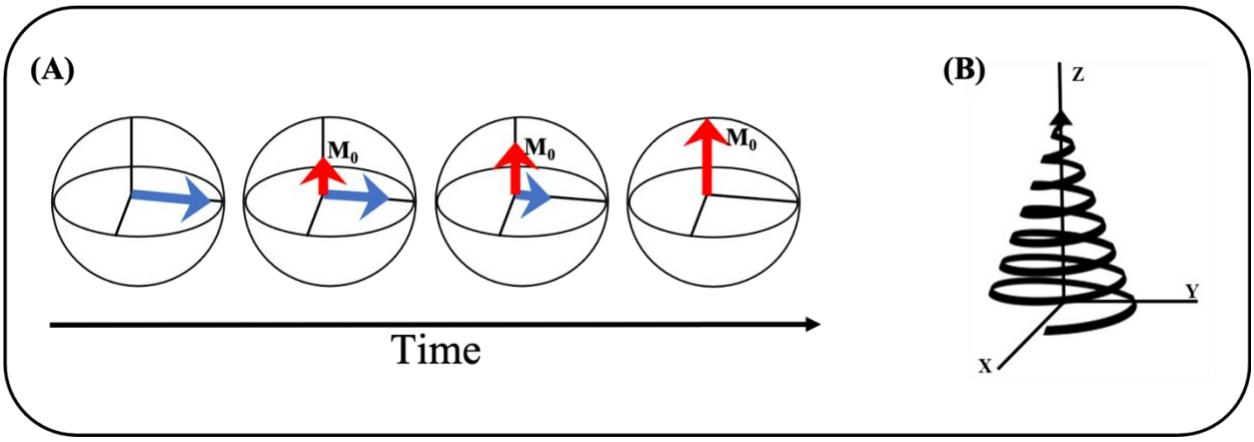


Figure 27. (a) Graphical illustration of the T_1 and T_2 relaxation processes after an RF pulse with a 90° flip angle. The blue arrow represents the xy-component of the magnetization vector M_0 (red arrow). The amplitude of the vector decreases in the xy-plane due to the loss of spin coherence (T_2 relaxation) and the resetting of the M_0 vector to its equilibrium position along the z-axis is due to energy exchange with the lattice (T_1 relaxation). (b) Resetting of the bulk magnetization from the xy-plane to the z-axis and the spiralling motion represents its precession.

During T_2 relaxation, magnetization in the xy-plane is lost due to the decoherence of the spins. Fluctuations of the spins' local magnetic field changes their Larmor frequency and as a result, they become out-of-phase with the rest of the spins, decreasing the amplitude of the net xy-magnetization (**Figure 27b**). T_2 relaxation can occur in conjunction with T_1 relaxation or independently of it. When a spin exchanges energy with the lattice, a concurrent change in the local magnetic field occurs, changing its precession frequency. In T_1-independent conditions, the dipolar interaction between the spins allows the exchange of magnetization, also changing the precession frequency of the spins. Since the amplitude of the magnetization does not change, T_1 relaxation is not induced.

2.3.3 Chemical shift

Structural features such as bond lengths, bond angles, and electronegativity of the bonded atoms significantly influence the electron densities about the nuclei. The electrons possess magnetic moments due to their precession in the applied field, which attenuates the actual field experienced by the nuclei. The new local field B_{eff} is expressed as [253]:

$$B_{eff} = (1 - \sigma)B_o \qquad 1.7$$

where σ represents the shielding of the nuclear spin by the surrounding electrons . Since the electrons surrounding a nucleus determine the latter's local magnetic field, differences in the electron densities causes a shift in the resonance frequencies of the nuclei, which are unique to each chemical environment. For example, in methyl (CH_3) and methylene (CH_2) functional groups, the carbon atoms are bonded to three and two hydrogen atoms, respectively. This difference in the number of bonded protons results in two different electron densities at the carbon centers, attenuating their precession frequencies to different extents. When the collected time-domain signal is Fourier-transformed, two peaks are observed in the ^{1}H and ^{13}C NMR spectra representing the methyl and methylene protons and carbons, respectively. This property of differentiating the functional groups is the fundamental reason for the success of NMR spectroscopy and its popularity as a structural characterization technique.

The difference in the frequency of a bare nucleus ($\nu_{nucleus}$) and the one in the chemical species under investigation (ν_{sample}) is considered the absolute magnetic shielding of that nucleus (σ_{sample}) and is represented as:

$$\sigma_{sample} \text{ (in ppm)} = 10^6 * (\nu_{nucleus} - \nu_{sample})/\nu_{nucleus} \qquad 1.8$$

The absolute shielding is transformed into an experimentally observable parameter called *chemical shift* (δ), which is the difference in the absolute shielding of the nucleus in the species under investigation (σ_{sample}) and the same nucleus in a reference compound (σ_{ref}) and is expressed as

$$\delta\ (ppm) = 10^6 * (\sigma_{ref} - \sigma_{sample})/(1-\sigma_{ref}) \qquad 1.9$$

This equation can be rewritten in frequencies as:

$$\delta\ (ppm) = 10^6 * (\nu_{sample} - \nu_0)/(\nu_0) \qquad 2.0$$

where, ν_{sample} is the resonance frequency of the sample and ν_0 is the spectrometer frequency [253]. The functional groups of a molecule have distinct chemical shifts and hence, NMR spectroscopy can be effectively applied to derive molecular structures [255,256]. The ^{1}H NMR spectrum of the Transactivator of transcription (Tat) protein with labelled peaks representing different functional groups is shown in **Figure 28**. The most widely used reference compounds in ^{1}H NMR are DSS (2,2-dimethyl-2silapentane-5-sulfonic acid), TMS (tetramethyl silane) and TSP (trimethylsilyl propanoic acid), with characteristic peaks at "0" ppm. In an NMR spectrum, the signals that appear at a frequency higher than the reference compound are considered deshielded and those appearing at lower frequencies are shielded. The higher the frequency of the peak, the more deshielded it is [257].

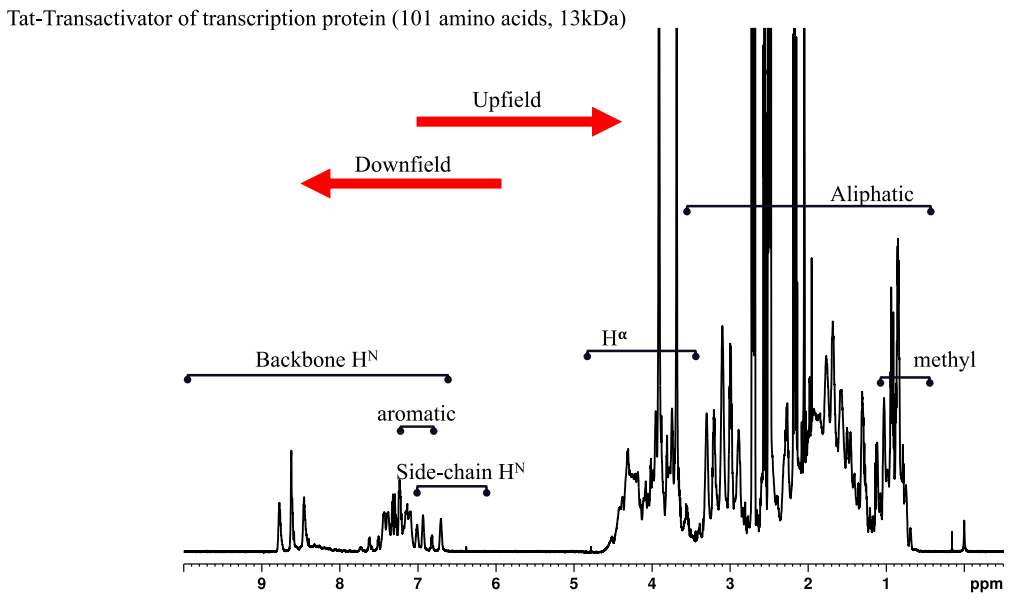


Figure 28. The ^{1}H NMR spectrum of the HIV-1 Transactivator of transcription (Tat) protein with different functional groups labelled.

NMR spectroscopy can obtain information about protein backbone and side-chain dihedral angles [258], intra- and inter-molecular atomic distances, proximity of aromatic rings, protein backbone dynamics and different conformational states of proteins [255,256,259]. NMR-based investigations of both ordered and disordered proteins rely heavily on the chemical shifts of the amino acid residues, which are determined by multi-dimensional ^{1}H, ^{13}C and ^{15}N NMR experiments [260][261]. The dihedral angles ϕ and ψ have a significant influence on the ^{13}C and ^{15}N chemical shifts [262] and hence, the type of secondary structure (alpha-helix, beta strands, random coil) can be established from the protein chemical shifts and can be used as constraints in structural calculations. In the case of a disordered protein, the secondary structure is formed only transiently and the chemical shifts are generally close to the "*random coil chemical shifts*". The difference between the measured chemical shifts and a set of random coil chemical shifts is

calculated, called the *secondary chemical shift*, which helps in identifying the location of the random coil regions and their population. In proteins with secondary structures, the chemical Shift Index (CSI) of the protein backbone is calculated by comparing experimental ^{1}H, ^{13}C and ^{15}N backbone shifts with a set of random coil chemical shifts (see **Table 11**). While the $^{1}H^{\alpha}$ chemical shifts of amino acid residues in helices are lower in frequency (upfield) compared to random coil shifts, a downfield shift is observed for $^{1}H^{\alpha}$ in beta strands. Similarly, the α-helices have positive $^{13}C^{\alpha}$ and negative $^{13}C^{\beta}$ shifts with respect to random coil chemical shifts, and a complimentary trend is observed in β-sheets [257]. In this graph-based technique, the assigned ^{1}H backbone shifts are converted to a three-state index, (−1, 0, +1). If the assigned shift is lower in frequency (upfield) than that of a random coil, it is labelled −1 and, if it is higher (downfield), it is labelled +1. No difference between the assigned and the random coil shift is marked with a 0. When these values are mapped along the protein backbone, clustering of −1 and +1 represents helices and beta strands, respectively in the protein structures. Programs such as chemical shift index (CSI) [263] and secondary structure propensity (SSP) [264] are used to predict secondary structures of protein using chemical shifts data.

Table 11. Random coil ^{1}H and ^{13}C chemical shifts of the 20 natural amino acids [265]

Amino acids	**NH**	**^{15}N**	**$^{13}C_{\alpha}$**	**$1H_{\alpha}$**	**$^{13}C_{\beta}$**	**CO**
Ala	8.24	123.8	52.5	4.32	19.1	177.8
Cys (Red)	8.32	118.8	58.2	4.55	28.0	174.6
Cys (Oxd)	8.43	118.6	55.4	4.71	41.1	174.6
Asp	8.34	120.4	54.2	4.64	41.1	176.3
Glu	8.42	120.2	56.6	4.35	29.9	176.6
Phe	8.30	120.3	57.7	4.62	39.6	175.8
Gly	8.33	108.8	45.1	3.96	–	174.9
His	8.42	118.2	55.0	4.73	29.0	174.1
Ile	8.00	119.9	61.1	4.17	38.8	176.4
Lys	8.29	120.4	56.2	4.32	33.1	176.6
Leu	8.16	121.8	55.1	4.34	42.4	177.6
Met	8.28	119.4	55.4	4.48	32.9	176.3
Asn	8.40	118.7	53.1	4.74	38.9	175.2
Pro	–	–	63.3	4.42	32.1	177.3
Gln	8.32	119.8	55.7	4.34	29.4	176.0
Arg	8.23	120.5	56.0	4.34	30.9	176.3
Ser	8.31	115.7	58.3	4.47	63.8	174.6
Thr	8.15	113.6	61.8	4.35	69.8	174.7
Trp	8.25	121.3	57.5	4.66	29.6	176.1
Tyr	8.12	120.3	57.9	4.55	38.8	175.9
Val	8.03	119.2	62.2	4.12	32.9	176.3

2.3.4 Scalar coupling

In a coupled spin system where the nuclei are bonded to each other through covalent bonds, the polarization of electrons present in the bond has a significant influence on nuclear resonance frequencies. This electron polarization-induced change in the local magnetic field about the nuclei is referred to as *scalar coupling* and leads to the splitting of resonance peaks in a typical NMR spectrum. It is also referred to as indirect *spin-spin coupling, J-coupling* or *through bond coupling* and follows the rule where the number of peaks in the splitting pattern is given by (*2I+1*) where, *I* is the spin quantum number of the bonded nuclei. The couplings are reported in Hertz (Hz) and are invariant to the external magnetic field. Scalar coupling therefore serves as an excellent tool in the structural analysis of both organic and inorganic molecules. More importantly, the J-couplings are sensitive to the conformations of molecules, and hence play a crucial role in protein secondary structure analysis. It is represented as $^{n}J_{AB}$ where, n is the number of bonds connecting the coupled nuclei A and B. The smaller the number of bonds between the nuclei, the greater is the magnitude of the scalar coupling. The $^{1}J_{C\alpha C\beta}$, $^{1}J_{C\alpha H\alpha}$, $^{1}J_{C\alpha C'}$, and $^{1}J_{C\alpha N'}$ coupling constants are dependent on the protein backbone torsions ϕ and ψ or the amino acid side-chain torsional angles χ_1 and χ_2 [266] **(Figure 29).** The coupling constants from the same spins (homonuclear) and chemically distinct spins (heteronuclear) are used to generate torsion angle distributions and predict secondary structures. For example, the predicted $^{3}J_{HN\alpha}$ coupling constants for α-helices, 3_{10} helices and β-strands are 4.8, 5.6 and 8.5 Hz respectively [267]. A list of homo- and heteronuclear couplings observed in proteins are given in **Table 12**. It should be noted that a slight variation in these coupling constants is generally observed based on the electronic environments of the observed spins.

Table 12. A list of homonuclear and heteronuclear J-coupling constants observed in proteins [267].

Homonuclear and Heteronuclear Spins	Couplings	^{1}H-^{1}H spins	Couplings
C-N	14 Hz	H-C=C-H	4-6 Hz (cis)/ 12-15 Hz (trans)
C-C	35 Hz	H-N-C-H	1-10 Hz (4-5 Hz α-Helix) (8-9 Hz β- strand)
H-N	92 Hz		
H-C	130 Hz		

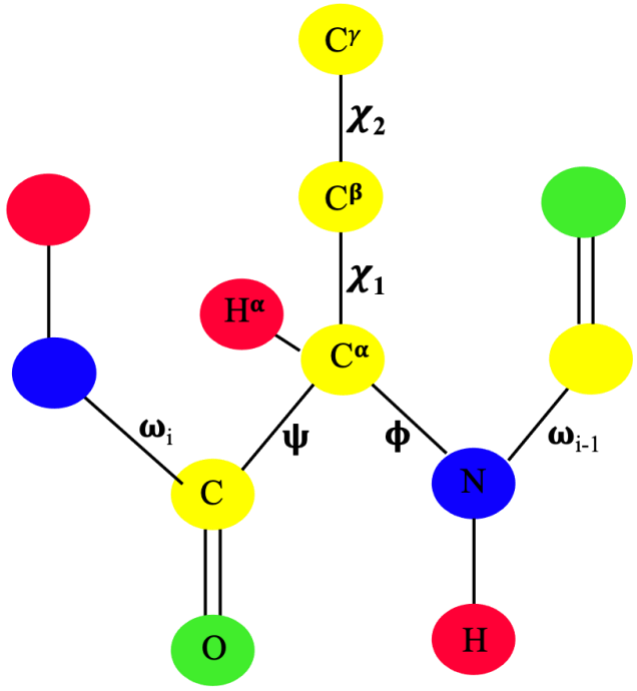


Figure 29. Structure of a typical amino acid with backbone torsion angles ϕ, ψ, χ_1, and χ_2 labelled. The topological representation of one-bond coupling constants like $^1J_{C\alpha C\beta}$, $^1J_{C\alpha H\alpha}$, $^1J_{C\alpha C'}$, and $^1J_{C\alpha N'}$ are also shown.

2.3.5 Dipolar coupling

Dipolar coupling is a through-space interaction between nuclear spins which attenuates the local magnetic field experienced by the nuclei. This interaction depends on the distance between a pair of nuclei, *i* and *j*, the gyromagnetic ratios of the nuclei (γ) and the angle between them when oriented in an applied field B_0 [268]. In solution, these spatial anisotropic interactions are negligible due to the rapid tumbling motion of the molecules [269]. Although many dipolar based NMR experiments have been developed, the Nuclear Overhauser Effect (NOE)-based NMR experiments such as NOESY and ROESY have gained a lot of traction due to their ability to measure interatomic distances in small proteins. In the NOE, the magnetization of one nucleus is transferred to a nearby nucleus through cross-relaxation. This could be between two chemically identical nuclei (homonuclear, ^{1}H-^{1}H, ^{15}N-^{15}N, ^{13}C-^{13}C) or two chemically distinct nuclei (heteronuclear, ^{1}H-^{15}N, ^{1}H-^{13}C). The dipole-dipole cross-relaxation rates of NOEs are directly proportional to r^{-6} where r is the distance between the atoms in the spin-pair. The extent of magnetization transfer is a function of the distance between the nuclei, which manifests as the intensity of an NOE cross-peaks in a typical NOE-based experiment [270–272]. Generally, the NOE is observable only when two NMR-active nuclei are in *ca.* 5 Å proximity [270,271]. Observed NOEs for IDPs are normally short-range, while the medium-range and long-range NOEs are rarely seen [272]. The NOEs are quite powerful in evaluating secondary structure [271].

2.3.6 NMR Protein Dynamics:

NMR is highly sensitive to the local structure and dynamics of proteins and other biological molecules at timescales ranging from pico- to milli-seconds [273–277]. NMR has thus been extensively used in deducing structural information on protein folding [278], for detecting poorly

populated excited ‘ghost/invisible’ states structures [248] and in enzyme catalysis studies [279]. Proteins are inherently flexible in solution at ambient temperature. Although the three-dimensional structure of a protein represents its ground state, biological functions often rely on the excited molecular states. For example, in the case of the allosteric molecule haemoglobin, its quaternary structure changes remarkably upon binding to oxygen [276]. Hence a fruitful interpretation of the complete structures of proteins and their transformation with time becomes crucial. The study of protein dynamics fills the gap between ground and excited states and explains how motion affects protein functions. Protein dynamics influences enzyme catalytic pathways, cellular signalling/regulation pathways and thermostability. Protein dynamics governs the rate and pathway of protein folding along with aggregation and misfolding which are responsible for fatal neurodegenerative diseases. In proteins, the timescale of atomic and molecular motions are different for example, atomic vibrations occur in the sub-picosecond regime, backbone and side-chain rotations occur on pico- to nanosecond timescales, conformational fluctuations happen at millisecond rates and some interactions occur on the order of seconds [279] **(Figure 30).** Further descriptions of molecular dynamics, influencing factors, and techniques devised to study protein dynamics are listed elsewhere [272,273,278–285].

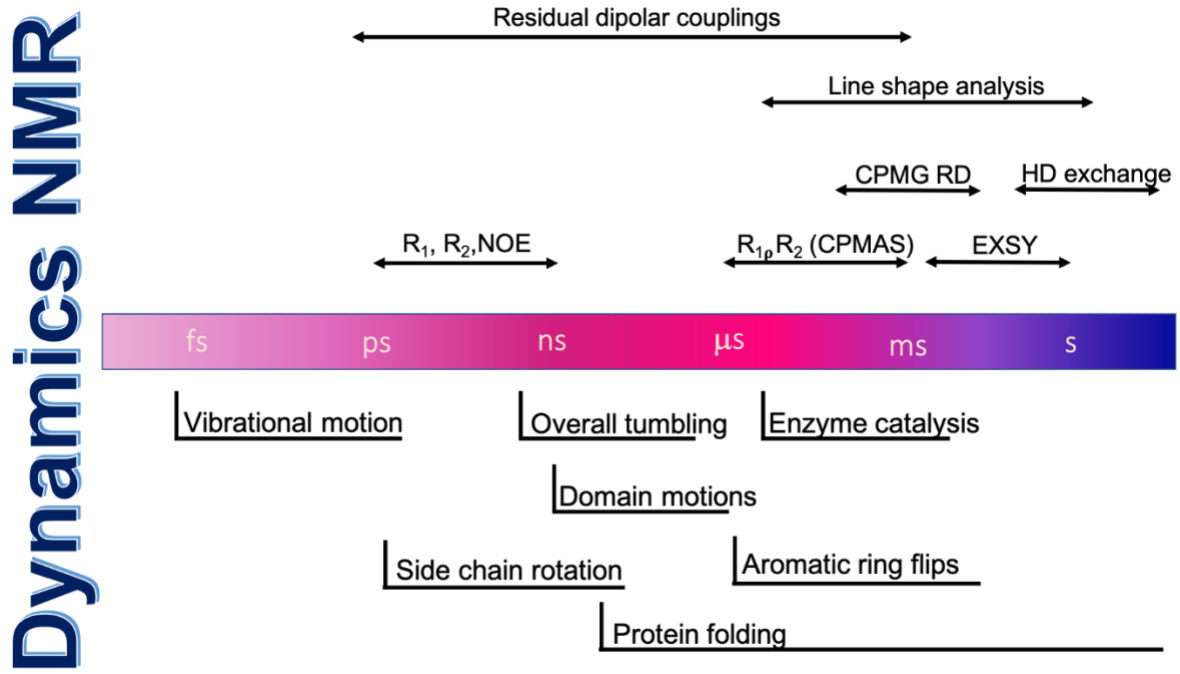


Figure 30. Timescale of protein dynamics (coloured bar). Aspects of conformational changes of proteins and some biological functions (bottom) and various methods (top) to study dynamics by NMR spectroscopy. CPMAS: Cross-Polarization Magic-Angle Spinning, CPMG RD: Carr-purcell Meiboom-Gill Relaxation Dispersion, EXSY: EXchange Spectroscopy (also known as ZZ-exchange), HD: Hydrogen-Deuterium exchange [286,287].

2.3.7 Two-dimensional NMR spectroscopy:

A two-dimensional experiment gives more insights on the structure of the target molecules compared to one-dimensional experiments. The two-dimensional correlation NMR spectra can provide correlations between both homo- and heteronuclei, either connected through one or many bonds, therefore the covalent connectivity within a molecule can be clearly established. In a homonuclear experiment such as COSY [288] and TOCSY [289], the obtained spectrum is symmetrical wherein the resonances are aligned along the spectrum diagonal, while the correlations between different nuclei appear as cross peaks (**Figure 32**). In a heteronuclear two-

dimensional experiment such as HSQC, the contours represent the correlation between the two nuclei which are scalar coupled. A schematic of a 2D experiment is shown in **Figure 31** and is comprised of four parts: preparation, evolution (t1), mixing and detection (t2). The two frequency axes arise from the Fourier transformation of the two time-variables, t1 and t2, which includes information about the free precision of the magnetization after excitation and the *spin-talking* phase where the magnetization is exchanged between the coupled spins generating correlations.

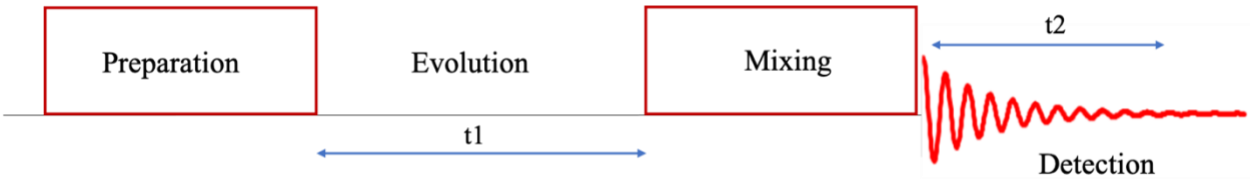


Figure 31. Scheme of a typical two-dimensional NMR experiment with all four parts outlined.

In the preparation step, the coherence is created in the xy plane, which evolves during the evolution time without any observation. In the mixing period, coherence is transferred between the J-coupled or dipolar-coupled spins by applying an RF pulse and the signals are collected in the detection period. During mixing, the magnetization is exchanged between the coupled spins which occurs in two ways, either by scalar coupling (through bond) or dipolar interactions (NOE). The mixing-stage may contain one or more RF pulses and may include further time-delays, depending on the type of experiment. The experiment starts with a zero evolution-time, which is incremented by Δt, which generates a matrix of time-domain points. The process of incrementing t1 and collecting data (t2) is repeated until there are enough data points for a two-dimensional Fourier transform.

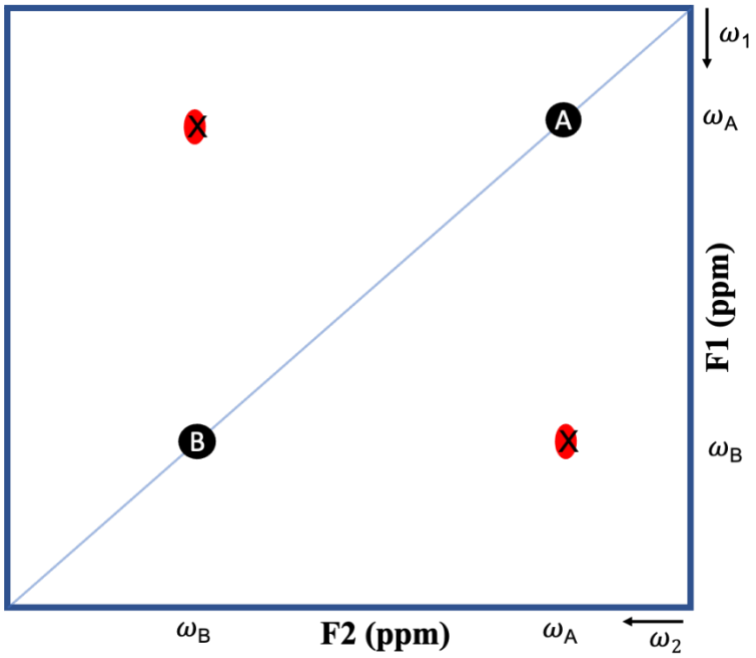


Figure 32. Graphical illustration of a two-dimensional NMR spectrum of spins A and B (Black diagonal) with the two frequency axes labelled. The cross peaks which are a resultant of A-B correlation are presented in red.

2.3.8 Heteronuclear single quantum coherence (HSQC):

In the NMR studies of proteins, the ^{1}H-^{15}N HSQC has emerged as a standard two-dimensional experiment, which is ubiquitously employed to study all classes of proteins. The chemical shift correlation between protons and directly-bonded nitrogen atoms in all of the amino acids is obtained, with the exemption of proline, which lacks the amide proton. The amide protons of each amino acid exhibit cross peaks at a characteristic shift in the obtained spectrum, which greatly enhances the ease of resonance assignment and protein structure determination. However, the experiment becomes less efficient when it comes to the IDPs, where the resolution drops due to severe signal overlap owing to poor dispersion of the resonances and in some cases owing to the addition of peaks due to multiple conformations of the IDP.

2.3.9 Direct-detection NMR

The conventional mode of studying structured proteins using two- and three-dimensional NMR experiments is through ^{1}H detection to enhance the sensitivity of the low-γ nuclei (**Table 13**). As mentioned above, this approach is not optimal in the case of intrinsically disordered proteins, as they exhibit severe overlap of resonances in their spectra due to poor chemical shift dispersion of the protons, conformational heterogeneity, aggregation and poor solubility [290]. As a remedy to this problem, direct detection methods (^{13}C and ^{15}N) can be adopted, as the chemical shift ranges of these nuclei are broad and they exhibit long T_2 relaxation rates, leading to sharp signals in comparison to the ^{1}H-detection methods. Furthermore, IDP protein sequences frequently contain a high number of proline residues, which lack amide protons. The ^{1}H detection-based experiments are insensitive to these proline residues and consequently structural information is missing for them. Recent improvements in probe design such as cryogenically-cooled direct-detection probes (cryoprobes) and isotopic labelling have improved the sensitivity of direct-detection experiments, allowing for the development of multi-dimensional experiments that can characterize IDPs with increased complexity [291].

Table 13. The important nuclei of proteins and their Gyromagnetic ratios and natural abundance for NMR studies [290].

Nuclei	Spin I	Gyromagnetic ratio γ_n (10^6 rad s^{-1} T^{-1})	Gyromagnetic ratio/2π (MHz T^{-1})	Nuclei natural abundance (%)
^{1}H	$^{1}/_{2}$	267.513	42.576	99.98
^{13}C	$^{1}/_{2}$	67.262	10.705	1.108
^{15}N	$^{1}/_{2}$	-27.116	-4.316	0.37

^{15}N direct detection

Generally, in disordered or unstructured proteins, the amide protons and other labile hydrogens undergo fast exchange with solvent. On some NMR time scales these can go undetected and do not appear in ^{1}H-based NMR experiments leading to incomplete data sets. On the other hand, low gamma nuclei-based experiments (^{15}N and ^{13}C) suffer no such intensity losses. A further problem with ^{1}H NMR is that as a consequence of fast transverse relaxation (short T_2) of ^{1}H spins in a macromolecular system, the peaks are inherently broad (line-width α $1/T_2$) and this is one of the biggest challenges for obtaining a well-resolved NMR spectrum [292]. In contrast, the slow relaxation properties (long T_1 and T_2) of low-gyromagnetic nuclei in comparison to proton leads to narrower line-widths than in ^{1}H NMR [137][293]. ^{15}N is an NMR-active isotope which has a low magnetogyric ratio and appreciable relaxation times. Moreover, all amino acids have amide groups in them which form the amide backbone of a protein chain. Furthermore, one need not worry about huge signals originating from water when running direct detection ^{15}N NMR experiments. Thus, no water suppression is required in ^{15}N-based experiments and hence we would expect to see an improved resolution in the spectra obtained [292]. Another advantage of direct ^{15}N detection is that signals from proline residues, which lack backbone amide protons and are therefore invisible to detection in conventional 2D ^{1}H-^{15}N HSQC experiments, are detectable in the ^{15}N direct-detection experiments. However, ^{15}N NMR experiments are inherently of low sensitivity due to the low gyromagnetic ratio of the observed spin. The polarization-transfer scheme adopted in these NMR experiments on low-gamma nuclei is most commonly scalar (*J*) coupling-mediated *e.g.*, INEPT proton polarisation transfer [293].

One prominent ^{15}N-based experiment is the 2D hCaN experiment [292], which can identify proline nitrogen's which are generally not observed in conventional ^{1}H-based experiments. The

spin coherence transfer pathway in a molecule which leads to an observable signal from a proline residue is depicted in **Figure 33**.

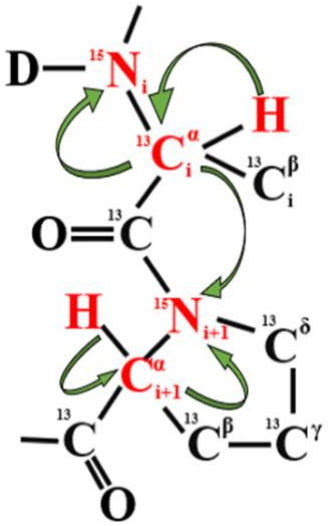


Figure 33. Coherence transfer pathway in the hCaN experiment. The nuclei which are involved in the experiment are in red colour and the arrows indicate the coherence pathway steps involved in the hCaN experiment [292].

Here, the magnetization is transferred from an alpha proton H^{i}_{α} to an attached alpha carbon $^{13}C_{\alpha}$, (step 1) and on to N^{i} and N^{i+1} atoms (step 2) [294]. By optimising the time involved in the magnetisation transfer of the second step, we are able to detect both intra and inter-residual chemical shift values. For example, if the time for the transfer of magnetization from $^{13}C_{a}$ to N^{i} and N^{i+1} atom is set to 22 ms, both intra (N^{i}) and inter (N^{i+1}) residue coherence transfers can be detected simultaneously. But with a setting of 70 ms, selective detection of sequential residual (N^{i+1}) coherence transfer can be selectively achieved [292].

2.4 Solid-State NMR Spectroscopy

The disordered structure of IDPs is not limited to the solution-state but has been observed in insoluble systems as well, for example, amyloid fibrils and membrane proteins [295,296]. Dipolar and J-based solid-state NMR (ssNMR) experiments have been used to probe both static and dynamic domains of IDPs [297,298]. When proteins exhibit low or no solubility, solid-state NMR experiments become good alternatives to study their structure even though, the information obtained represents one local minimum on their potential energy surface. This is particularly not advantageous in studying the IDPs as they are characterised by an ensemble of conformations [299,300] but, failure to derive any structural information at all by solution NMR makes solid-state NMR experiments acceptable. Another important feature of ssNMR is that anisotropic interactions (chemical shift anisotropy, dipolar coupling and quadrupolar interactions) can be effectively probed, which are averaged in solution NMR due to molecular tumbling. Since the numbers and types of ssNMR experiments are extremely large, only a few important experiments which were adopted in this work are presented here.

While the fundamental principles are the same [301,302], the hardware, experimental setup and sample preparation methods are significantly different between solid and solution NMR methods. The NMR resonances observed in a ssNMR spectrum are severely broadened due to the anisotropic interactions mentioned above. Since the Hamiltonians of the anisotropic interactions contain $(3\cos^2\theta\text{-}1)$ terms, where θ is the sample is inclined at an angle of 54.74° with respect to the external magnetic field, B_0, they become θ-dependent, and spinning the sample at $\theta = 54.74°$ averages them out yielding sharp signals. This angle is termed the *magic-angle*, and the samples could be spun or not, depending on the choice of the NMR experiment that needs to be performed.

In ssNMR of proteins, the lyophilized protein is packed into ceramic rotors of varying sizes (0.7−4 mm) and spun at the magic-angle inside the superconducting magnet (**Figure 34**).

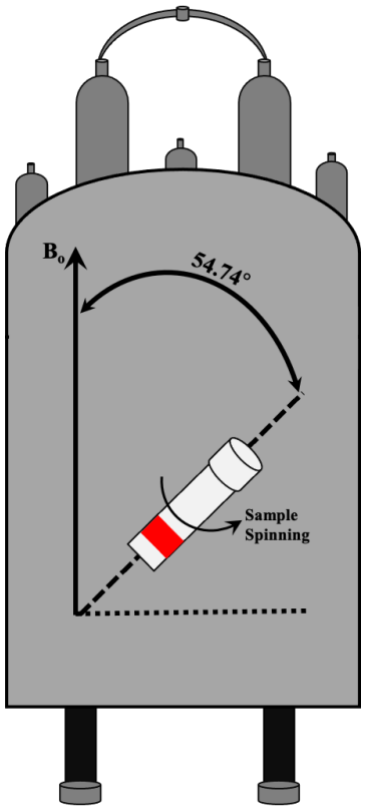


Figure 34. Graphical illustration of a rotor spinning at the magic-angle inside a superconducting magnet. The sample is inclined at an angle of 54.74° with respect to the external magnetic field, B_0.

2.4.1 Cross Polarization Magic-Angle Spinning NMR spectroscopy (CP-MAS)

Cross-polarization is a popular ssNMR technique that is employed on a routine basis to study molecules both, organic and inorganic, including macromolecules such as the proteins. It is mainly employed to observe spins with low gyromagnetic ratios and low natural abundance that are relevant to the structural analysis of proteins (for example, ^{13}C and ^{15}N). Long relaxation times of

these nuclei and low natural abundance warrant several thousand scans which will make experiments time consuming, which is particularly problematic for studying unstable and fast degrading proteins. This is a dipolar-based experiment, which relies on the transfer of magnetization from a spin with high-γ and high natural abundance, typically ^{1}H, to low-γ and less abundant nuclei such as ^{13}C and ^{15}N. For the magnetization transfer to occur, the energy gap between the spin states (α and β) of the abundant spin (I) and the dilute spin (S) has to be the same, which is achieved by applying RF pulses of varying amplitudes on the S channel, while the I spins are spin-locked in the xy-plane. This experiment depends on a complex spin dynamics which has been explained in detail elsewhere [301,302] (**Figure 35**). This is called the Hartmann-Hahn condition [303] and is given as [293],

$$\gamma_H B_{1H} = \gamma_X B_{1X}$$

Where, γ is the gyromagnetic ratio of the nuclei (γ_H: ^{1}H; and γ_X: ^{13}C,^{15}N), B_1 is the applied magnetic field, facilitating the matching of energy gaps between the spin states. Since spinning the sample at the magic-angle averages the dipolar interaction between I and S, lower spinning frequencies must be employed. Moreover, since the dipolar interaction is distance-dependent, the intensity of the signals in the CPMAS spectrum provides the spatial information of the molecule and the dynamics of the molecule as the molecular tumbling averages the dipolar interaction too.

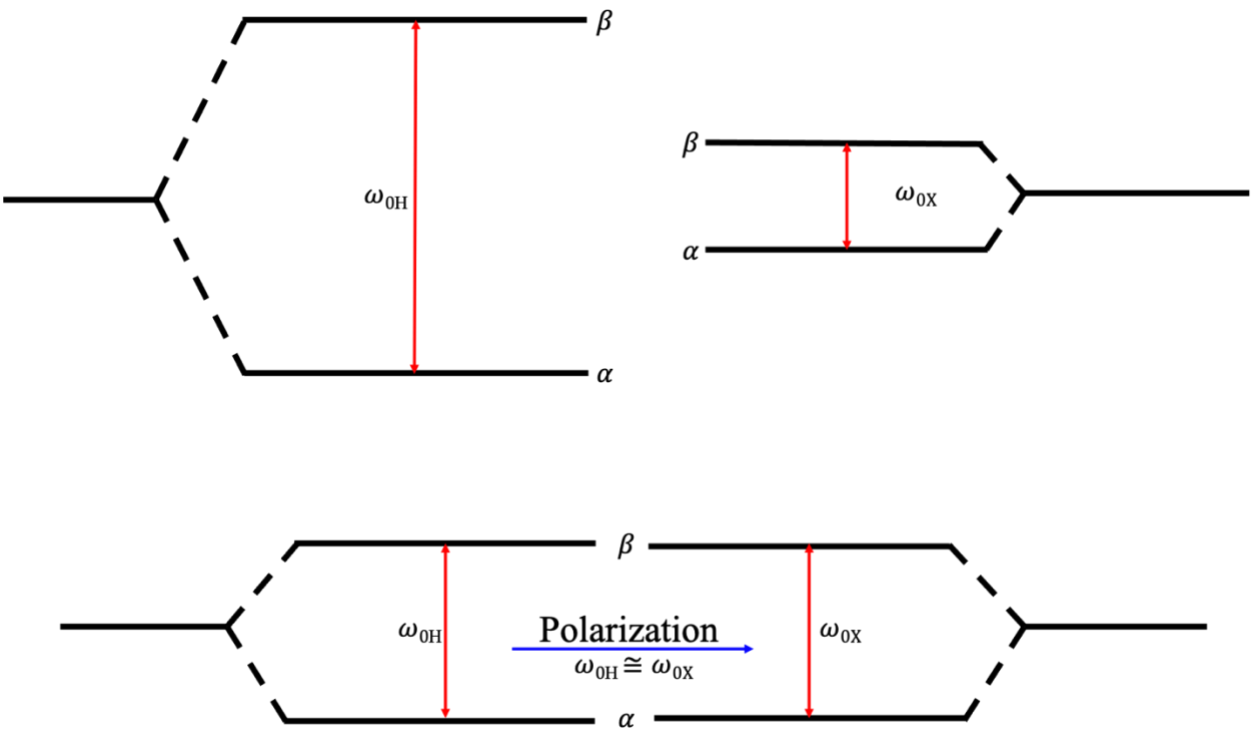


Figure 35. Schematic overview of the Hartmann-Hahn condition before (top) and after (bottom) polarization transfer. ω_H and ω_C represent the Larmor frequencies of ^{1}H and X nuclei, respectively.

2.4.2 Insensitive nuclei enhanced by polarization transfer (INEPT)

While the CPMAS experiment is strictly dipolar, experiments based on through-bond coupling are also performed to determine the covalent connectivity within a molecule, which ultimately leads to an efficient structural analysis. In the INEPT experiment, polarization is transferred from ^{1}H to any heteronuclei (^{13}C or ^{15}N), which are chemically bonded, without any dependency on the H-X bond orientation [304]. One important aspect of this experiment is that non-bonding interactions are not visible, which is extremely helpful in mapping the sequential connectivity that exists.

2.4.3 Proton driven spin diffusion (PDSD):

This is one of the standard two-dimensional experiments which is carried out on a routine basis to determine the proximity of carbons to a specific proton. This can be a powerful aid in the backbone and sidechain assignment process. In the proton-driven spin diffusion experiment, magnetization is transferred from ^{1}H to ^{13}C, which dissipates to other proximal carbon nuclei [305]. The mixing time, which allows for the magnetization exchange between the coupled spins, can be tuned to derive short-range or long-range interactions. A mixing time of 50 ms yields intra-residue information, while mixing times of 250−500 ms provide inter-residue contacts [305]. The experimental conditions and results are discussed in Chapters 3 and 4, respectively.

2.5 Goals of the Research

The Tat protein is a small nuclear protein which, upon binding to TAR RNA, triggers the viral transcription of the HIV-1 genome. It contains a small-sequence of basic amino acids (residues 48−57) which are involved in binding with TAR RNA, a 59 nucleotide-long RNA with a stem-loop structure found at the 5' end of the HIV-1 viral RNA. The Tat protein is intrinsically disordered (IDP) and similar to many IDPs, it is insoluble in aqueous media at physiological pH. However, it is highly soluble at pH 4 and has been extensively studied by NMR spectroscopy at acidic pH [121,137]. Between pH 6 and 7, the NMR peaks broaden and disappear as the protein precipitates from the solution. In order to understand the mechanism by which this protein functions and binds to its partner proteins, the protein has to be water-soluble at pH 7. Only then, can structural changes in the protein that are responsible for its functioning and its dynamics be effectively studied by NMR spectroscopy.

My thesis focuses mainly on improving the solubility of Tat protein at pH 7 so that, its physiologically relevant structure can be effectively studied. To this measure, I will explore different methodologies that could significantly contribute towards solubilizing the Tat protein. This research has a high impact as solubilizing the protein is highly important to understand its mechanistic role in viral replication and its mode of interaction with its binding partners. I will employ NMR spectroscopy, one of the powerful techniques in structural biology to study the structure of Tat. A diverse class of solution- and solid-state NMR experiments will be adopted to determine intra- and inter-residue interactions that will be extremely useful in mapping the connectivity in the protein sequence, which will ultimately provide the three-dimensional physiologically relevant structure of the Tat protein. Furthermore, once complete solubility is achieved, the interaction of Tat with its binding partners will be scrutinized, which could prove extremely beneficial in designing anti-viral drug for HIV.

3. Materials and Methods

3.1 Protein production of unlabelled Tat

The expression of histidine-tagged Tat-protein was carried out using a previously optimised protocol [121,137]. A glycerol stock containing expression plasmid was added to 25 mL of rich medium containing 0.675 grams (g) of Luria-Bertani (LB) broth powder and 30 μL of 34 mg/mL stock kanamycin in a 125 mL baffled-flask. The mixture was incubated overnight in an orbital shaking incubator at 37 °C with a fixed rotation of 300 rpm. This culture was added to 1 L of rich medium containing 25 g of LB powder and 0.5 mL of 34 mg/mL stock kanamycin in a 4 L baffled flask. The resulting mixture was further incubated at the above-mentioned conditions. Protein expression was induced by adding 60 milligrams (mg) of isopropyl β-D-1-thiogalactopyranoside (IPTG) when the OD_{600} was between 0.8−1.0. After 5 hours of expression, the cells were harvested by centrifugation at 5000 xg for 15 min at 4 °C.

The harvested cell pellets weighing *ca.* 2−3 g were resuspended in 100 mM of pH 7.2 phosphate buffer (100 mL), with 200 μg DNase, 200 μg RNase and 10 mg of lysozyme, and incubated for 30 minutes at room temperature. The suspension was frozen at -80 °C. After two cycles of freeze-thaw, the solution was subjected to three to four cycles of sonication on ice at an amplitude of 30% with 30 s between cycles using a Model-300 Sonic Dismembrator. To dissociate Tat from cellular nucleic acids and reduce its cysteines, 90 g of guanidine hydrochloride (GuHCl) (6 M) and 286.95 mg (10 mM) of tris(2-carboxyethyl) phosphine hydrochloride (TCEP-HCl) were added to the sonicated solution (100 mL) and the pH was adjusted to 7.2 using 100 mM phosphate buffer. The re-suspended solution was centrifuged at 33,000xg for 30 minutes at 4 °C (Centrifuge: Thermo scientific SORVALL LYNX 6000; Rotor: FIBERLITE F14-14x50cy). After

centrifugation, the supernatant was collected and bound to a cobalt metal-affinity chromatographic resin column as described in the next section.

3.2 Uniform ^{15}N- and ^{13}C-labelling of Tat protein for the NMR experiments

Bacterial cultures containing the Tat expression-plasmid were grown in M9 minimal medium prepared by diluting the five-times concentrated M9 medium containing 15 g of KH_2PO_4, 35 g of Na_2HPO_4, and 2.5 g of NaCl per litre. The M9 medium was autoclaved and supplemented with 34 mg of kanamycin, 1 mL of 1 M $MgSO_4$, 100 µL of 1 M $CaCl_2$, 1 mL of 1000× Trace Metals (Teknova), 1 g of $^{15}NH_4Cl$ and 2 g of [$^{13}C_6$]-D glucose. A small culture of 125 mL was grown overnight in the LB-medium and was partitioned into two flasks containing 500 mL of LB-media each. The cells were allowed to grow until the OD_{600} reached 1 and the medium was centrifuged at 2600 xg for 15 min at 4 °C. Following centrifugation, the pellet was resuspended in the M9 minimal medium which was pre-warmed at 37 °C for 30 min. Protein expression was induced by adding IPTG to the medium after 30 min. After 4−5 hours of expression, the cells were harvested by centrifuging the broth at 5000xg for 20 min at 4 °C; the cell pellets were processed as described above.

Metal affinity chromatography: 4 mL of TALON® cobalt super-flow resin (Clonetech) was packed into a 10 mL polypropylene gravity-low column (QIAGEN Inc.) and pre-equilibrated using 50 mL of extraction buffer, which contained 100 mM $NaH_2PO_4{\cdot}2H_20$ at pH 7.2 with 6 M GuHCl and 10 mM of TCEP-HCl. The supernatant prepared above was loaded onto the pre-equilibrated column followed by elution with 20 mL of extraction buffer and 30 mL of washing buffer. Washing buffer contained 6 M GuHCl, 50 mM of $NaH_2PO_4{\cdot}2H_20$, and 10 mM of TCEP-HCl at pH 6.4. Finally, the full-length Tat-protein adhering to the resin through His-tags was eluted using 20 mL of elution buffer (6 M GuHCl, 20 mM acetic acid and 10 mM of TCEP-HCl at pH 4.0). The eluent was collected as 1 mL fractions in 5 centrifuge tubes. These fractions were loaded into

water-pre-equilibrated Fisher brand dialysis tubes with a molecular weight cut-off of 3.5 kDa. Dialysis was conducted against 1 L of different concentrations of acetate buffer (0.1 M with 10 mM EDTA, and 0.1 M, 0.05 M and 0.01 M without EDTA) at pH 4 for 3−4 hours each. Each dialysis buffer was degassed and purged with argon for 5−10 min before the dialysis was carried out. The chelating agent, ethylenediaminetetraacetic acid, was added to remove any traces of cobalt metal from the resin. Post-dialysis, the protein was frozen in a dry-ice ethanol bath and lyophilized. The pure freeze-dried protein was stored at -20 °C for future experiments.

3.3 Expression and purification of sequence-specific nickel assisted cleavage (SNAC) Tat protein

SNAC-Tat gene, which was codon-optimised in *Escherichia coli* (*E.coli*) and cloned into pET28a(+) vector was purchased from Genscript. The synthetic gene (Appendix I) contains a 5' NcoI restriction cut site (C^CATGG), Tat-protein sequence, a Carboxy terminal SNAC-tag (GSHHW), a linker (GSS), a histidine tag, a linker (SSG), 3 stop codons (TAATGATAA) and a 3' EcoRI restriction site (G^AATTC) in its sequence. The plasmid was first chemically transformed into TOP10 cells (Thermo Fisher Scientific, Massachusetts, U.S. A) for storage and later into BL21(DE3) (New England Biolabs, Massachusetts, U.S. A) strain of *E.coli.* The next day, a single healthy colony was picked from the transformed cell-culture plate and added to 25 mL of LB media with kanamycin (30 μL of 34 mg/mL stock) as antibiotic. The culture was grown overnight in an orbital shaking incubator at 37 °C with a fixed rotation of 300 rpm. Next day, the glycerol stock was made by mixing 500 μL of overnight culture and 500 μL of 50% glycerol, and the mixture was flash frozen in liq. Nitrogen and stored at -80 °C.

The five-residue SNAC-tag peptide sequence [227] is marked in red in the Tat-SNAC tag-linker-His tag-linker sequence listed here:

MEPVDPRLEPWKHPGSQPKTACTNCYCKKCCFHCQVCFITKALGISYGRKKRRQRRRPP
QGSQTHQVSLSKQPASQPRGDPTGPKESKKKVERETETDPVD**GSHHW**GSSHHHHHHSS
G

As above in **Section 3.1** the overnight culture grown, and protein was expressed. The harvested cell pellets weighing around 2−3 g was resuspended in 100 mM of pH 7.2 phosphate buffer (100 mL), with 200 µg DNase, 200 µg RNase and 10 mg of lysozyme, and incubated for 30 min at room temperature. The suspension was frozen at -80 °C. After two cycles of freeze-thaw, the solution was subjected to three to four cycles of sonication on ice at an amplitude of 30% with 30 s between cycles using a Model-300 Sonic Dismembrator. To dissociate Tat from cellular nucleic acids and reduce its cysteines, 90 g of GuHCl (6 M) and 286.95 mg (10 mM) of TCEP-HCl were added to the sonicated solution (100 mL) and the pH was adjusted to 7.2 using 100 mM phosphate buffer. The resuspended solution was centrifuged at 33,000xg for 30 min at 4 °C (Centrifuge: Thermo scientific SORVALL LYNX 6000; Rotor: FIBERLITE F14-14x50cy). After centrifugation, the supernatant was collected and bound to a cobalt metal-affinity chromatographic resin column as described in the next section.

Metal affinity chromatography: 4 mL of TALON® cobalt super-flow nickel resin (Clonetech) was packed into a 10 mL polypropylene gravity column (QIAGEN Inc.) and pre-equilibrated using 50 mL of extraction buffer. Extraction buffer contained 100 mM $NaH_2PO_4 \cdot 2H_2O$ at pH 7.2 with 6 M GuHCl and 10 mM of TCEP-HCl. The supernatant prepared above was loaded onto the pre-equilibrated column followed by washing with 30 mL of washing buffer. Washing buffer contained 6 M GuHCl, 50 mM of $NaH_2PO_4 \cdot 2H_20$, and 10 mM of TCEP-HCl at pH 6.4.

The Sequence specific nickel assisted cleavage (SNAC) can be performed in two ways:

On-resin cleavage: The SNAC-tagged protein was tightly bound to the Ni^{2+}-nitrilotriacetic acid (NTA) resin by the histidine tag. The SNAC-tag from the SNAC-Tat protein was cleaved before the elution step using 10 mL of cleavage buffer containing 0.1 M 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES), 0.1−0.5 M NaCl, 0.1 M acetone oxime (sigma), 0.5 mM TCEP, 3 M GuHCl, pH 8.6 and incubated overnight at 22 °C [227]. The following day, cleavage buffer contain the Tat protein was collected.

Off-resin cleavage: In this method, the SNAC-tagged protein was eluted from the column using elution buffer as explained in **Section 2.2**. The eluted protein was dialysed, lyophilised, added (1 mg/mL) to 10 mL of cleavage buffer (pH 8.6) containing 1mM $NiCl_2$, 0.1M HEPES, 0.1 M acetone oxime, 0.1−0.5 M NaCl, 0.5 mM TCEP and 3 M GuHCl and incubated for 16 h at room temperature. The cleavage buffer containing the tagged protein was loaded onto a column filled with 0.5 mL of Ni-NTA resin the following day.

Eluents from both the methods were loaded into water-pre-equilibrated Fisher brand dialysis tubes with a molecular weight cut-off of 3.5 kDa. Dialysis was conducted as above in **Section 2.2**. Post-dialysis, the protein was frozen in a dry-ice ethanol bath and lyophilized. The pure freeze-dried protein was stored at -20 °C for future experiments. The SNAC cleavage mechanism and results are explained in later section.

3.4 Bacterial expression and purification of human Cyclin-T1 (1−266)

The synthetic human cyclin T1 (1-266) [13] gene that was codon optimised in Escherichia coli (*E.coli*) and cloned into pET28a vector (Novagen, Madison, WI) was procured from Genscript. The obtained plasmid was first transformed into TOP10 cells for reposition and later into

BL21(DE3) strain of *E.coli* for higher protein expression. The protein sequence is 300 amino acids long and contains an N-terminal histidine tag, a short linker followed by a TEV site and human cyclin-T1. The protein sequence is as follows [2]:

MGSSHHHHHHSSGLVPRGSHMASMTGGQQMGRGSDGPENLYFQGMEGERKNNNKRWYF
TREQLENSPSRRFGVDPDKELSYRQQAANLLQDMGQRLNVSQLTINTAIVYMHRFYMIQSFT
QFPGNSVAPAALFLAAKVEEQPKKLEHVIKVAHTCLHPQESLPDTRSEAYLQQVQDLVILESI
ILQTLGFELTIDHPHTHVVKCTQLVRASKDLAQTSYFMATNSLHLTTFSLQYTPPVVACVCIH
LACKWSNWEIPVSTDGKHWWEYVDATVTLELLDELTHEFLQILEKTPNRLKRIWNWRACE
AAKK

As above in **Section 3.1** the overnight grown and protein was expressed. The harvested cell-pellets were resuspended in 50 mL of resuspension buffer at pH 7.2. The resuspension buffer stock was made with 500 mL containing 50 mM HEPES (5.95 g), 300 mM NaCl (8.76 g), 2 mM TCEP (286 mg), 200 μg of DNAse and RNAse each and 10 mg of lysozyme. The suspended solution was stored at -80°C. After single freeze-thaw, the solution was sonicated on ice 10 times at an amplitude of 50% with 30 s on-off cycle using a sonic dismembrator instrument (Fisher Scientific, Model 500). The lysed cells were centrifuged at a maximum speed of 17000 xg for 30 min at 4 °C. The supernatant was treated by using a polypropylene gravity flow column (QIAGEN Inc.) of 10 mL capacity. The column was packed with 4 mL of Co^{2+} or Ni^{2+} resin (Clonetech) and pre-equilibrated with 50 mL of extraction buffer at pH 7.2, constituting 100 mM HEPES (1.1915 g), 300 mM NaCl (876 mg) and 2 mM TCEP (28.69 mg). The addition of supernatant was followed by a washing with 30 mL of washing buffer containing 20 mM HEPES (143 mg), 300 mM NaCl (525.9 mg), 15 mM imidazole (30.6 mg) and 2 mM TCEP (17.21 mg) at pH 7.2. Finally, the protein was eluted with the addition of 20 mL of elution buffer containing 20 mM HEPES (95 mg), 300 mM NaCl (350.64 mg), 200 mM imidazole (272 mg) and 2 mM TCEP (11.47 mg).

The eluted protein was dialysed against 1L of degassed dialysis buffer made of 20 mM HEPES (4.766 g), 100 mM KCl (7.453 g), 50 mM NaCl (2.92 g), 1 mM DTT (154.2 mg) and 5

mM EDTA (1.46 g) at pH 7.2. Dialysis was carried out in two steps, each with a duration of 4 h: In the first dialysis step, the dialysis buffer contained 5 mM EDTA to chelate with any cobalt or nickel metal which may have leaked from the column resin; In the second step, the dialysis buffer lacked EDTA. Post-dialysis, the solution was frozen, freeze-dried and stored at -20 °C.

3.5 Bacterial expression and purification of carboxy terminal domain of RNA polymerase II (RPB-1 hCTD 1593−1970) fusion protein.

The human RNA pol II CTD [146] (hCTD, RPB1 1593−1970) plasmid which was codon optimised in *Escherichia coli* (*E.coli*) and cloned into pET28a(+) vector was ordered from Genscript. The synthetic gene (Appendix I) contained a 5' NcoI restriction cut-site (C^CATGG), N-terminal His-tag, soluble maltose binding protein (MBP), 10X Asparagine (N) linker, tobacco etch virus (TEV) cut-site, a Carboxy terminal hCTD, RPB-1 1593−1970, stop codons (TAATGATAG) and a 3' SalI restriction site (G^TCGAC). The plasmid was first transformed into TOP10 cells for storage and later into the BL21(DE3) RIL cells (Stratagene) for higher protein expression.

The gene sequence was optimised using the Genscript Optimum Gene^{TM} codon analysis tool [306]. This algorithm optimises various parameters [**Table 1**] that influence transcription, translation and protein folding to produce a gene that can provide the highest level of protein expression. A comparison of the original and optimised gene (from Genscript) is shown in Figures 1 and 2.

Table 14. Various factors considered for the optimization of protein expression using the Genscript OptimumGene™ codon analysis tool [306].

Transcription	Translation	Protein Folding
GC content	Codon usage bias	Codon usage bias
CpG dinucleotides content*	GC content	Interaction of codon and anti-codon
Cryptic splicing sites	mRNA secondary structure	Codon-context
Negative CpG islands	Premature PolyA sites	RNA secondary structures
TATA box	RNA instability motif (ARE)	
Terminal signal	Inhibition sites	
	Stable free energy of mRNA	

*CpG dinucleotides or CpG oligodeoxynucleotides (CpG ODN): Synthetically made short single stranded DNA molecules constitute a cytosine triphosphate deoxynucleotide (C) followed by a guanine triphosphate deoxynucleotide (G).

Codon adaptation index (CAI): This parameter describes how well a codon suits the codon usage preference of a target organism. The CAI range is 0−1. The CAI=1 is a perfect match and if CAI=0.8 this is considered good [307]. Using the gene optimization tool, Genscript changed the codon bias in *E.coli* by upgrading the CAI from 0.41 to 0.69 **(Figure 36)**.

GC content adjustment: The Genscript OptimumGene™ codon analysis tool optimises this parameter to get better gene expression. The ideal range of GC content is between 30−70% [308]. The GC content in the original sequence has 53.81% and after optimization it has increased to 57.30% **(Figure 37)**.

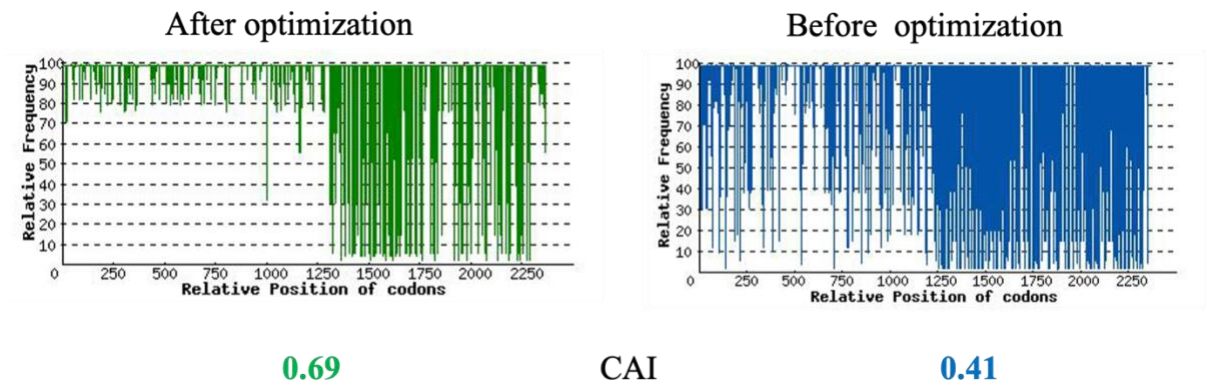


Figure 36. Comparison of CAI of RPB-1 hCTD (1593-1970) original and optimised gene [306].

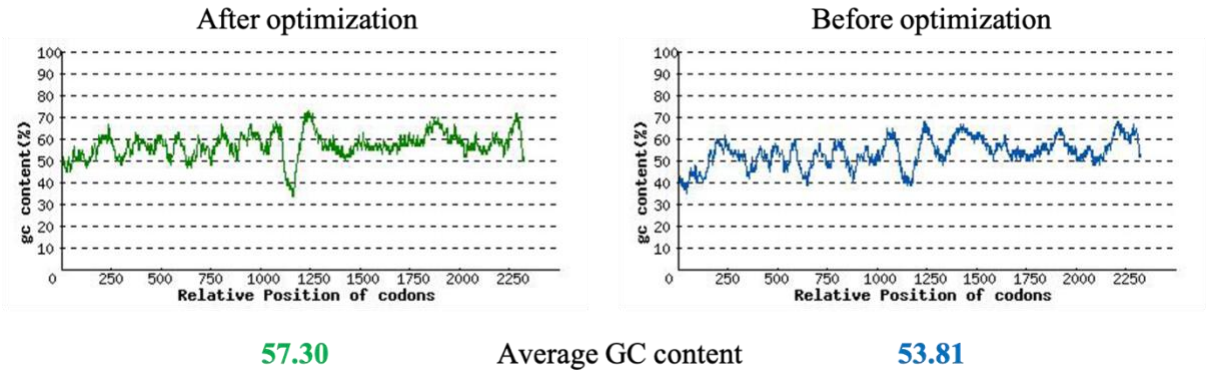


Figure 37. The GC content in the gene sequence before and after optimization [306].

The 786 amino acid long sequence of the MBP-RPB-1 hCTD (1593−1970) RNA polII fusion protein is as follows [146]:

MGSSHHHHHHGSSMKIEEGKLVIWINGDKGYNGLAEVGKKFEKDTGIKVTVEHPDKLE
EKFPQVAATGDGPDIIFWAHDRFGGYAQSGLLAEITPDKAFQDKLYPFTWDAVRYNGK
LIAYPIAVEALSLIYNKDLLPNPPKTWEEIPALDKELKAKGKSALMFNLQEPYFTWPLIAA
DGGYAFKYENGKYDIKDVGVDNAGAKAGLTFLVDLIKNKHMNADTDYSIAEAAFNKG
ETAMTINGPWAWSNIDTSKVNYGVTVLPTFKGQPSKPFVGVLSAGINAASPNKELAKEF
LENYLLTDEGLEAVNKDKPLGAVALKSYEEELAKDPRIAATMENAQKGEIMPNIPQMSA
FWYAVRTAVINAASGRQTVDEALKDAQTNSSSNNNNNNNNNNLGIEGRENLYFQSNAC

YSPTSPAYEPRSPGGYTPQSPSYSPTSPSYSPTSPSYSPTSPNYSPTSPSYSPTSPSYSPTSPS
YSPTSPSYSPTSPSYSPTSPSYSPTSPSYSPTSPSYSPTSPSYSPTSPSYSPTSPSYSPTSPSYSP
TSPSYSPTSPSYSPTSPSYSPTSPNYSPTSPNYTPTSPSYSPTSPSYSPTSPNYTPTSPNYSPTS
PSYSPTSPSYSPTSPSYSPSSPRYTPQSPTYTPSSPSYSPSSPSYSPTSPKYTPTSPSYSPSSPE
YTPTSPKYSPTSPKYSPTSPKYSPTSPTYSPTTPKYSPTSPTYSPTSPVYTPTSPKYSPTSPT
YSPTSPKYSPTSPTYSPTSPKGSTYSPTSPGYSPTSPTYSLTSPAISPDDSDEEN

The glycerol stock was added to 25 mL rich medium containing 0.675 g of LB powder and 34 mg/ml of Kanamycin in a 125 mL baffled flask. The cells were grown overnight in an orbital shaking incubator at 37°C and at a fixed rotation of 300 rpm. It was then added to 1 L of rich medium consisting of 25 g of LB powder and 0.5 mL of Kanamycin in a 4 L baffled flask. The mixture was incubated at 37 °C in an orbital shaking incubator rotating at 300 rpm. Protein expression was induced by adding 60 mg of IPTG. Once the OD_{600} was in the range of 0.6−0.8, the cells were harvested and centrifuged at a minimum of 5000 xg for 15 min at 4 °C.

The cell-pellet was resuspended in the lysis buffer (LB300) at pH 7.4, which contained 20 mM HEPES, 300 mM NaCl, 30 mM imidazole, 10% glycerol, 1 mM DTT, 0.284 μg/mL leupeptin, 1.37 μg/mL pepstatin A, 0.17 mg/mL phenylmethylsulfonyl fluoride (PMSF) and 0.33 mg/mL benzamidine. The cell-lysate was immersed in liquid nitrogen and stored at -80 °C.

The purification steps were carried out at 4 °C. The suspended cells were freeze-thawed, and the solution was subjected to five cycles of sonication on ice at an amplitude of 45% with 30 s between cycles using a Model-300 Sonic Dismembrator. The lysed cells were centrifuged at maximum speed (27000 xg) for 45 min at 4 °C (Centrifuge: Thermo scientific SORVALL LYNX 6000; Rotor: FIBERLITE F14-14x50cy). The centrifuged solution was filtered through 0.22 μm filters. The purification of the hCTD: RPB1 supernatant was done by using a gravity flow column with a capacity of 10 mL. The column was packed with 4 mL of Co^{2+} or Ni^{2+} resin and pre-equilibrated with 50 mL of LB300 lysis buffer. Then the column was washed with high salt buffer (HSB1000)

at pH 7.4 containing 20 mM HEPES, 1 M NaCl, 30 mM imidazole, 10% glycerol, 1 mM DTT, 0.284 μg/mL leupeptin, 1.37 μg/mL pepstatin A, 0.17 mg/mL PMSF, 0.33 mg/mL benzamidine. Then the proteins bound to the resin are eluted using nickel elution buffer 300 at pH 7.4 containing 20 mM HEPES, 300 mM NaCl, 500 mM imidazole, 10% glycerol, 1 mM DTT, 0.284 μg/mL leupeptin, 1.37 μg/mL pepstatin A, 0.17 mg/mL PMSF and 0.33 mg/mL benzamidine. The eluted protein solution was subjected to size-exclusion chromatography (SEC) by making use of Superdex 200 10/300 which is pre-equilibrated with size-exclusion 300 (SE300) buffer at pH 7.4 containing 20 mM HEPES, 300 mM NaCl, 10% glycerol and 1 mM TCEP-HCl. The pure fractions from SEC were assessed by SDS-PAGE and concentrated by centrifugation with 30-kDa MWCO (molecular weight cut-off) centrifugal filters [Amicon Ultra-15, Centrifugal filters, Millipore corporation]. The pure protein was aliquoted and snap frozen and preserved at -80 °C.

3.6 Bacterial expression and purification of supercharged Tat (SuperR_{10}-Tat)

The supercharged R_{10}-Tat plasmid was procured from Genscript, which was codon optimised for expression in *Escherichia coli* (*E.coli*) and cloned into pET28a(+) vector. The synthetic gene (Appendix I) **(Figure 38)** contained a 5' NcoI restriction cut site (C¯CATGG), N-terminal His-tag, TEV cut site, R_{10}-Supercharge sequence, two-residue spacers (GG), the Tat sequence, 3 stop codons (TAATGATAG) and a 3' BamHI restriction site (G↑GATCC). The plasmid was first transformed into TOP10 cells for storage and later into BL21(DE3) RIL (Stratagene) strain of *E.coli* for protein expression.

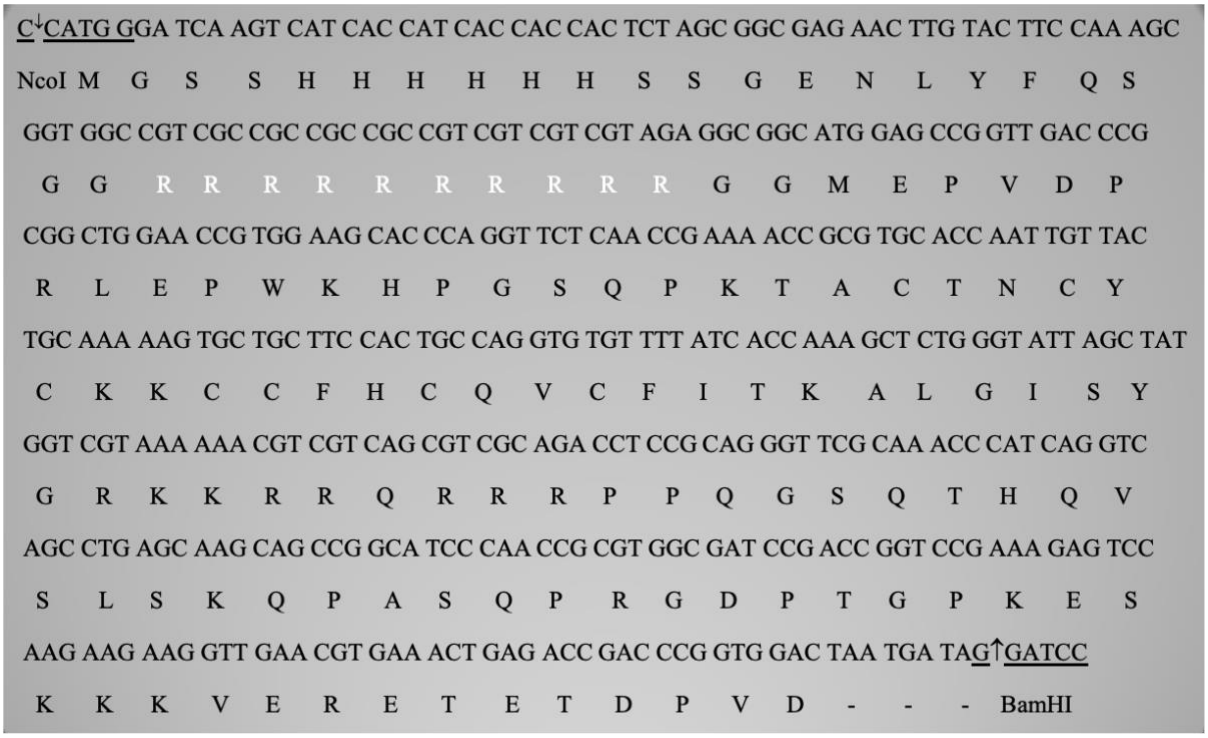


Figure 38. The optimised gene sequence and respective protein sequence of R_{10}Supercharged Tat. The stop-codons (TAATGATAG) and restriction sites NcoI (C¯CATGG) and BamHI (G↑GATCC) are also shown.

As above in **Section 3.1** the overnight culture grown, and protein was expressed. The harvested cell-pellets weighing around 2−3 g was resuspended in 100 mM of pH 7.2 phosphate buffer (100 mL), with 200 µg DNase, 200 µg RNase and 10 mg of lysozyme, and incubated for 30 min at room temperature. The suspension was frozen at -80 °C. After two cycles of freeze-thaw, the solution was subjected to three to four cycles of sonication on ice at an amplitude of 30% with 30 s between cycles using a Model-300 Sonic Dismembrator. To dissociate Tat from cellular nucleic acids and reduce its cysteines, 90 g of GuHCl (6 M) and 286.95 mg (10 mM) of TCEP-HCl were added to the sonicated solution (100 mL) and the pH was adjusted to 7.2 using 100 mM phosphate buffer. The resuspended solution was centrifuged at 33,000xg for 30 min at 4 °C (Centrifuge: Thermo

scientific SORVALL LYNX 6000; Rotor: FIBERLITE F14-14x50cy). Post-centrifugation, the supernatant was collected and bound to a cobalt metal-affinity chromatographic resin column as described above in this **Section (3.2)**

3.7 Bacterial expression and purification of Asp-Tat

The Asp Tat plasmid was procured from Genscript, which was codon optimised for expression in *Escherichia coli* (*E.coli*) and cloned into pET28a(+) vector. The plasmid was first transformed into TOP10 cells for storage and later into BL21(DE3) RIL (Stratagene) strain of *E.coli* for protein expression. All the cysteine residues are replaced with aspartic acid in the protein sequence.

As above in **Section 3.1** the overnight culture grown, and protein was expressed. The harvested cell-pellets weighing around 2−3 g was resuspended in 100 mM of pH 7.2 phosphate buffer (100 mL), with 200 μg DNase, 200 μg RNase and 10 mg of lysozyme, and incubated for 30 min at room temperature. The suspension was frozen at -80 °C. After two cycles of freeze-thaw, the solution was subjected to three to four cycles of sonication on ice at an amplitude of 30% with 30 s between cycles using a Model-300 Sonic Dismembrator. To dissociate Tat from cellular nucleic acids and reduce its cysteines, 90 g of GuHCl (6 M) and 286.95 mg (10 mM) of TCEP-HCl were added to the sonicated solution (100 mL) and the pH was adjusted to 7.2 using 100 mM phosphate buffer. The resuspended solution was centrifuged at 33,000xg for 30 min at 4 °C (Centrifuge: Thermo scientific SORVALL LYNX 6000; Rotor: FIBERLITE F14-14x50cy). Post-centrifugation, the supernatant was collected and bound to a cobalt metal-affinity chromatographic resin column as described above in this **Section (3.2)**

3.8 *In vitro* synthesis of TAR RNA

The phosphoramidite process used to chemically synthesize RNA and DNA molecules has revolutionized the fields of biochemistry, biophysics and molecular biology [309]. However, this process is limited to RNA molecules which are 50 nucleotides long as the protection of 2'-hydroxyl group of ribose sugar is complex and expensive [310]. This limitation could be overcome by synthesizing RNA molecules through transcription using a DNA template and phage T7 RNA polymerase [311]. The T7 polymerase enzyme can either be purchased or expressed and purified in *E. coli* cells. Credit to Prof. Sean McKenna's lab for providing HIV TAR RNA plasmid and T7 RNA polymerase plasmid and expression and purification directions. The *in vitro* synthesis of RNA can be divided into five different steps: (a) Plasmid linearization; (b) *In vitro* transcription; (c) Desalting; (d) Gel filtration and (e) Purity assessment.

(a) Plasmid Linearization:

The HIV TAR plasmid was obtained from Prof. McKenna's lab and the detailed procedure in synthesizing the plasmid is mentioned elsewhere [312]. The glycerol stock containing the plasmid was added to 150 mL of rich medium containing 2.5 g of LB powder and 34 mg/ml of Ampicillin in a 150 mL baffled flask. The mixture was incubated for 16 h in an orbital shaking incubator at 37 °C at a fixed rotation of 250 rpm. The following day, DNA was isolated using the Maxiprep kit (Thermofischer), which came with detailed instructions on its operation. The DNA was digested with appropriate restriction enzymes and the complete digestion of the plasmid was confirmed by conducting 1 % agarose gel electrophoresis. The digested plasmid was purified using the phenol-chloroform extraction method wherein, the digestion solution was mixed with an equal volume of phenol-chloroform and vortexed. The immiscible phases were separated by centrifuging the

mixture at 3000 xg for 5min and the upper aqueous layer was carefully placed into a new tube. To this, 1/10th volume of 3 M sodium acetate (pH 5.2) was added. The plasmid DNA was precipitated by the addition of three volumes of ice-cold 95% ethanol and centrifuged at 21000 xg for 15 min at 4 °C. The precipitated DNA was again washed with additional ice-cold 70% ethanol and centrifuged at 21000 xg for 15 minutes at 4 °C. The purified DNA pellet was dissolved in HPLC-grade water and the concentration of the plasmid was determined by measuring the absorbance at 260 nm using a nano-drop instrument (Make and Model). The final concentration of the DNA was adjusted to 500 mg/mL.

Expression and purification of T7 RNA polymerase for *In vitro* synthesis of TAR RNA

The stored glycerol stock (from Prof. Mckenna's lab) was added to 25 mL of rich medium containing 0.675 g of LB powder and 34 mg/ml of Ampicillin in a 125 mL baffled flask. This was grown overnight in an orbital shaking incubator at 37 °C and at a fixed rotation of 300 rpm. The overnight culture was added to 1 L of rich medium consisting of 25 g of LB powder and 0.5 mL of ampicillin in a 4 L baffled flask. The added culture was incubated in an orbital shaking incubator at 37 °C and a fixed rotation of 300 rpm. Protein expression was induced by adding 1 mL of 1 M IPTG when the OD_{600} reached 0.6. Later, the cells were harvested after expressing for 3 hours by centrifuging at 5000 rpm for 10 min at 4 °C. Harvested pellets were resuspended in 20 mL of cold lysis-buffer. The solution was subjected to ten cycles of sonication on ice at an amplitude of 45% with 30 s between cycles using a Model-300 Sonic Dismembrator. The solution was centrifuged at 14500 rpm for 30 min at 4 °C. The supernatant was purified using a polypropylene gravity flow column (QIAGEN Inc.) with a capacity of 10 mL. The column was packed with 4 mL of Ni^{2+} resin (Clonetech), which was pre-equilibrated with 5 mL of milli Q water and 5 mL of lysis-buffer

[**Table 15**]. The supernatant collected from centrifugation was loaded onto the column, followed by a wash with 10 mL of wash buffer-1 (defined in Table 2 below) and 10 mL of wash buffer-2. The protein was then eluted using 10 mL elution buffer. The column was cleaned by washing it with 5 mL of elution buffer and 10 mL of milli Q water and was stored in 5 mL of 20% ethanol. The eluted protein was transferred into a dialysis tube which was placed in a beaker containing storage buffer. The whole setup was stored in a cold room (4^{o}C) overnight with stirring. The protein was collected from the dialysis tube and stored at -20 °C for future experiments.

Table 15. T7 RNA polymerase buffers production protocol.

Buffers	**Lysis (100 ml)**	**Wash 1 (50 ml)**	**Wash 2 (50 ml)**	**Elution (50 ml)**	**Storage (500 ml)**
Tris-HCl (pH 8) 1 M stock	50 mM 5 ml	50 mM 2.5ml	50 mM 2.5 ml	50 mM 2.5 ml	50 mM 25 ml
NaCl 5 M stock	100 mM 2 ml	300 mM 3 ml	300 mM 3 ml	100 mM 1 ml	100 mM 10 ml
β- mercaptoethanol	5 mM 35 µL	5 mM 17.5 µL	5 mM 17.5 µL	20 mM 70 µL	20 mM 700 µL
Glycerol	5% 5 ml	5% 2.5 ml	5% 2.5 ml	20% 10 ml	50% 250 ml
Imidazole 2 M stock	1 mM 50 µL	10 mM 250 µL	25 mM 625 µL	200 mM 5 ml	
PMSF 100 mM stock	1 mM 1 ml	1 mM 0.5 ml	1 mM 0.5 ml	1 mM 0.5 ml	1 mM 5 ml
EDTA 0.5 M stock				1 mM 100 µL	1 mM 1 ml
Triton X-100				0.1% 50 µL	0.1% 500 µL
Water	86.9 ml	41.2 ml	40.8 ml	30.8 ml	207.8 ml

(b) *In vitro* Transcription

The Mg^{2+} nucleotide triphosphates (NTPs) and T7 RNA polymerase ratio should be optimised before carrying out the standard transcription. The Mg^{2+} concentration may vary for each reaction or when new plasmid, T7 RNA polymerase or NTPs reagent are prepared. Hence, the optimization should be repeated as it is known to increase the yield [312]. The concentration of T7 RNA polymerase or $MgCl_2$ are the two variable parameters in the transcription reaction, and to optimise them, the best option is to start with 50 µL of trail transcription. The composition of the 50 µL trail transcription is as follows: 5 µL of purified DNA template (500 µg/mL), 8 µL NTPs (ATP, GTP, CTP, UTP) (50 mM), 5 µL 10X transcription buffer, 2 µL T7 polymerase, 0-10 µL MgCl2 (100 mM) with 2 µL increments and nuclease-free water. The T7 RNA polymerase is the last reagent to be added to the reaction mixture.

Reagents used in *in vitro* transcription:

a) 10x transcription buffer: 400 mM Tris-HCl (pH of 8.1), 10 mM spermidine, 0.01% Triton X-100, 100 mM Dithiothreitol (DTT) and nuclease-free water.
b) 10x Tris-Borate-EDTA (TBE) buffer (1 L): 108 g of Tris, 40 mL of 0.5 M EDTA (pH of 8) and 55 g of boric acid and nuclease-free water.
c) 10 % TBE/urea denaturing gel: 19.2 g urea, 4 mL of 10x TBE buffer, 0.3 mL of ammonium persulfate (APS), 20 µL of Tetramethylethylenediamine (TEMED), 10 mL of 40 % acrylamide/bisacrylamide (29:1) and 25.6 mL of HPLC water.
d) 2x denaturing loading Dye: 24 g of urea, 2 mL of 0.5 M EDTA, 10 mL of 10x TBE buffer, 25 mg of xylene cyanol, 25 mg of bromophenol blue in 50 mL of HPLC-grade water.

e) 0.1% Toludine blue staining solution: 1 g of Toluidine blue and 10 mL of glacial acetic acid in 1000 mL of water.

The reaction mixture was placed in a water bath at 37 °C for *ca.* 1 h. The trial transcription resulted in white precipitates (pyrophosphates), which were removed by centrifugation at 3000 g for 5 min at room temperature. 10 μL of transcription mixture was mixed with 10 μL of 2X denaturing TBE loading dye and heated at 95 °C for 5 min. The mixture was assessed by electrophoresis in a 10% urea/TBE polyacrylamide denaturing gel for an hour at 100 V. The denatured gels were submerged in staining solution containing 0.1% toluidine blue for 10 min. The gel was frequently washed with deionised water until the RNA bands were visible.

Once the small-scale reaction condition was optimized, large scale transcription could be carried out at in various volumes. Typically, 10 mL of transcription reaction would yield 1−5 mg of RNA which is enough for future experiments. This mixture was incubated at 37 °C for 3 hours and the white pyrophosphate precipitates were separated by centrifugation at 3000 xg for 5 min at room temperature. The supernatant was collected into a new tube and 50 mM EDTA was added to quench the transcription reaction by chelating the Mg^{2+}, which will deactivate the T7 RNA polymerase. To remove the T7 RNA polymerase an equal volume of phenol-chloroform was added and mixed well by vortexing. The RNA was separated by centrifuging at 3000 xg for 10 min at room temperature. The upper aqueous phase was collected by pipetting and transferred into a new tube.

(c) Desalting Transcribed RNA

This is a crucial step as it serves to remove a large quantity of unused NTPs and left-over phenol from phenol-chloroform extraction. Care must be taken as the Superdex column may get damaged

by residual phenol. A gravity-flow column was used for desalting. A 10DG desalting column was equilibrated with 20 mL of RNA buffer. The RNA buffer contained a mixture of 100 mM NaCl and 10 mM sodium phosphate (pH 6.6). Later, it was loaded with 3 mL of Phenol-RNA extract and the solution was allowed to drain completely. 1 mL of RNA buffer was added to the column and it too was eluted. The RNA transcripts were eluted from the desalting column by adding 5 mL of RNA buffer. The eluted RNA could be stored at 4 °C or straightaway purified by FPLC.

(d) Gel filtration

The Fast Performance Liquid Chromatography (FPLC) technique was used to separate RNA transcripts from unsuccessful transcripts, unused NTPs, the plasmid, aggregated RNA and other contaminants. An AKTA PURIFIER 10 system was used in conjunction to a Frac-950 fraction collector, operated with UNICORN software suite kindly provided by Cytiva. Before running samples, the column was equilibrated with RNA buffers. The sample or any reagent solutions that were loaded onto the column were filtered through 0.22 μm filters. The eluted RNA sample was passed through a 50 mL superloop connected to the filtration column using RNA buffers. In total, 15 mL of sample was loaded for a typical 10 mL transcription. The RNA was eluted by adding RNA buffers from size-exclusion column at a fixed speed of 1.5−2 mL min^{-1}. The 5 mL fraction pools were collected in 15 mL glass tubes.

(e) RNA purity Assessment

Purity and aggregation of pooled RNA fractions were probed by native gel electrophoresis. 20 μL of pooled RNA fraction was loaded onto 10% native TBE gels. The electrophoresis was conducted for 3 h at 75 V, in a 0.5x TBE at 4 °C. The gel was stained with 0.1% toluidine blue solution to visualize the RNA-containing species. The concentration of purified RNA was determined

spectrophotometrically at 260 nm, by using an appropriate extinction coefficient. The detailed information on RNA extinction coefficient calculations can be found in this reference [313]. The purified RNA sample was concentrated through centrifugal ultrafiltration of suitable MWCO centrifugal filters [Amicon Ultra-15, Centrifugal filters, Millipore corporation]. Finally, the purified RNA was stored at 4 °C for short-term, and for long term storage, the RNA sample was preserved at -20 °C.

3.9 Liquid-NMR sample preparations:

After expression and purification of the protein, it was lyophilised under vacuum overnight. The lyophilized protein was dissolved in 600 μL of buffer containing 10 mM acetate, 10 mM HEPES, 75 μM 2,2-dimethyl-2-silapentane-5-sulfonate (DSS) and, 10% deuterium oxide (D_2O). The protein concentration in the NMR tube was calculated to be 500 μM using UV spectroscopy on a Nano Drop spectrophotometer (Thermo scientific). The molar extinction coefficient of Tat protein at 280 nm was predicted to be 8250 M^{-1} cm^{-1} using Protein Calculator [314]. The protein sample was filled into an NMR tube and with proper care, it was degassed by purging with argon for 5 minutes. The NMR tube was capped and sealed with Teflon tape.

All standard NMR experiments were conducted at 14.1 T on a Varian ^{INOVA} spectrometer equipped with a 5 mm HCN triple resonance probe at room temperature. The experiments were conducted on a uniformly ^{13}C- and ^{15}N-labelled reduced Tat protein. The obtained spectra were processed and visualised in NMR Pipe [315] and assignments were made using Sparky software [316].
NMR samples for direct-detection experiments were prepared in D_2O which consisted of 10 mM HEPES, pH 4, 5 mM TCEP and 75 μM DSS. The spectra were acquired at 16.4 T on a Bruker

Avance III HD Ascent spectrometer housed in the Ohio State University and equipped with 5 mm Triple-resonance observe (TXO) cryoprobe with Z gradients.

3.10 Solid-state NMR sample preparations:

After expression and purification steps, the protein was freeze-dried. The lyophilised protein was dissolved in pH 4 and pH 7 buffer containing 10 mM acetate, 10 mM HEPES and 5 mM TCEP. Again, protein was lyophilised, and the protein powder was filled in 2.5 mm rotor and the rotor capacity is 10 mg. NMR experiments of ^{13}C,^{15}N labelled Tat were conducted at 11.7 T on a Bruker spectrometer equipped with triple channel standard bore and experiments are listed in **Table 16**.

Table 16. List of experiments performed on ^{13}C, ^{15}N labelled Tat.

Sl. No	Experiments	No. of Scans	MAS rate (kHz)	Temperature (° C)
1	1D-INEPT	1024	15000	15° C, 25°C & 35 °C
2	1D-CP	1024	15000	15° C, 25°C & 35 °C
3	2D-PDSD	128	15000	25°C

3.11 Procedure for acquiring intrinsic fluorescence spectrum

The lyophilised protein was dissolved in buffer containing 3 mM 2-(N-morpholino) ethanesulfonic acid (MES), 10 mM sodium phosphate dibasic and 3 mM TCEP. The intrinsic tryptophan fluorescence spectrum was measured on a Horiba scientific spectrofluorometer. 500 μL of sample was placed in 1 cm path length quartz fluorescence cuvette and spectra were collected between 300 to 500 nm using a 1 nm data pitch, a bandwidth of 5 nm and excitation wavelength of 295 nm.

All different pH samples were equilibrated at room temperature for an hour before acquiring spectra. All spectra were corrected by subtraction of a spectrum of the buffer obtained used identical settings to those used of the sample.

3.12 Procedure for FTIR sample preparation and spectral analysis

For FTIR spectroscopy, the sample is expected to be 95% pure and the sample purity was determined by SDS-PAGE or Mass spectrometry (MS). The freeze-dried protein was dissolved in 500 μL of aqueous buffer containing 5 mM HEPES buffer, (pH 5, 6, and 7), 0.25 mM dithiothreitol (DTT) to form a protein solution of concentration 5 mg mL^{-1}.

The FTIR spectra of 50 μL of protein solution were acquired on a Bruker FT-IR spectrometer INVENIO® in attenuated total reflection (ATR) mode with a resolution of 4 cm^{-1}, a spectral range of 4000 to 1000 cm^{-1} and 1000 scans. Initially, background and buffer spectra were acquired, and their difference spectra were later subtracted from the sample spectrum to account for matrix and electronic interference effects.

The FTIR spectra were deconvoluted by second-derivative methods to determine the fraction of secondary structures (α-helix, β-sheet, β-turn, random coil and 3_{10}-helix) present in the Tat protein using their characteristic wavenumbers. The second derivative method identifies the hidden peaks, which are smoothed by applying a $7-9$ point Savitsky-Golay function with a polynomial order of 2. Deconvolution was carried out using fitting software (Peakfit and Sigma plot) that employed a combination of Gaussian and Lorentzian components [317]. Best fits were achieved by constraining the peak position to desired wave numbers and changing the peak widths and peak heights through iteration. Areas under each peak were integrated and they represent the fraction of each type of secondary structure present in the protein.

3.13 Analysis of SNAC cleaved Tat Protein by Mass Spectrometry:

The lyophilised protein powder was dissolved in 10 mM ammonium formate at pH 4. Before injecting the samples are expected to be clean and filtered. The sample was injected at the ESI source at room temperature, using an HPLC auto sampler, with a flow rate of 300 mL min-1. The mobile phase consisted of formic acid 0.1% in Milli-Q water for Channel "A", and formic acid 0.1% in acetonitrile for Channel "B". A 1-2.5 µL aliquot of the extract was injected, with no liquid chromatography (LC) column, to the ESI ion source, at 100% B from 0 to 1 min. From 1.1 to 3 min, the gradient was linearly ramped to 50% B, where it was kept for 3.9 min. Then, the gradient was linearly ramped to 100% B, and held for almost 1 min, returning to 100% A for re-equilibration for next run. The Liquid Chromatography Electron Spray Mass Spectrometry (LC-ESI-MS) used was a Bruker Compact Quadrupole Time-Of-Flight (Q-TOF) (Bruker, Billerica, MA) equipped with an electrospray source operating in positive ionization mode.

The source used 0.4 bar for the nebulizer pressure and 4 L.min-1 of N2 drying gas at 120°C; capillary voltage was 4500 V, and the scan begin from 300 m/z ending at 3000 m/z.

Data processing was performed using Bruker Compass Data Analysis (ver. 5) software, where mass of full protein was calculated using its deconvolution tool.

4. Results and Discussion:

4.1 Histidine-Tagged Tat (His-Tat)

4.1.1 Purification of His-Tat protein:

The cloning of N-terminal Histidine-tagged Tat protein into the pET28b vector and further expression and purification steps are explained in **Section 3.1**. The yields of unlabelled and isotope-labelled proteins were 15 and 10 mg/L, respectively. The expressed protein was purified by metal affinity chromatography method and the purity was determined by sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE). An image of a gel is given in Figure 3.1, which shows the bands from the protein eluted at pHs 4 and 7, and the standard DNA ladder representing different molecular weights on the right edge of the gel. A band at ca. 16 kDa (red box) is observed at both pHs, which is assigned to the histidine-tagged full-length Tat-protein based on previous observations [121]. In addition, some low-intensity bands are observed between 30 and 40 kDa, which represent aggregated Tat protein [121]. At pH 7, thick bands at the top of the stacking gel (**Figure 39a**) and at the interface of the stacking gel and the resolving gel (**Figure 39b**) are observed. These bands are attributed to aggregated protein, which is believed to be a consequence of oxidation of cysteine disulphide bonds and possible interactions between hydrophobic side chains (24). Moreover, an increase in the pH from 4 to 7 lowers the net charge of the protein, possibly leading to increased interactions between oppositely charged residues and forcing the protein to self-aggregate.

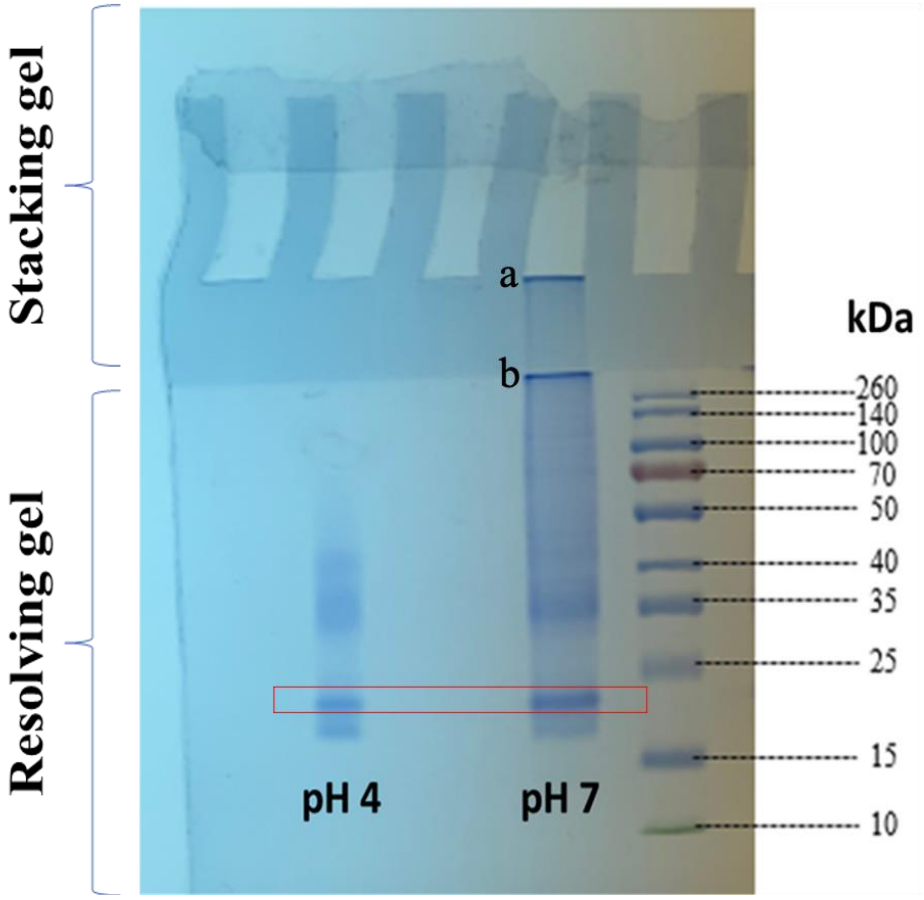


Figure 39. SDS-PAGE electropherogram of Tat-protein at pHs 4 and 7.

4.1.2 Fluorescence Spectrum of His-Tat

Changes in the fluorescence of side-chain aromatic groups can provide valuable information to understand protein folding and conformational changes. The spectral analysis of tyrosine and phenylalanine fluorescence is rarely carried out. One reason is that the fluorescence of tyrosine and phenyalanine do not change with the polarity of the environment, phenylalanine is only weakly fluorescent and tyrosine is often quenched *via* radiation-less energy transfer to tryptophan. In contrast, tryptophan fluorescence intensity and emission wavelength are highly sensitive to the charged and polar residues in its neigbourhood. Thus, the emission wavelength of tryptophan in

proteins is sensitive to small changes in the native structure brought about by changes in solution conditions such as pH, temperature and ionic strength.

The experimental parameters and sample preparation for my fluorescence studies on Tat are explained in detail in **Section 3.12**. Burstein and colleagues (Reshetnyak 2001) [318] classified the fluorescence of tryptophans in proteins into 5 classes. The emission maximum of tryptophan in denatured proteins and tryptophan fully exposed to water occurs at around 350 nm and is classified as Class III in their scheme. We were surprised to observe the emission maximum in Tat at pH 4 at 331 nm (**Figure 40**), which Burstein and colleagues would classify as Class I (331± 6 nm). Many tryptophans in Class I have reduced access to solvent water and are usually H-bonded. By comparison, fully buried tryptophans in Class S fluoresce at 322±5 nm. The fluorescence spectrum of Tat at pH 4 suggests that the tryptophan resides in a locally structured region and this interpretation is supported by NMR chemical shifts [137].

Increasing the pH of Tat a solution above 4 causes significant quenching of the signal from the single tryptophan present in the sequence. Callis and Liu (2004) [319] have shown that a wide range of tryptophan fluorescence quenching in proteins can be explained by electron transfer from the excited tryptophan to the carbonyl of a nearby amide group. Upon excitation, the electron in the highest occupied molecular orbital is excited to the lowest unoccupied molecular orbital (LUMO) of the Trp-ring, the fluorescence state ^1L_a, which is further transferred to the LUMO of an amide (π^*) provided these two states have equal energies (**Figure 41**). The charge-transfer states are very sensitive to the local electric field direction and strength near tryptophans. Thus, the observed quenching in Tat fluorescence at elevated pH may be attributed to either the presence of a negatively charged amino acid next to the indole ring of the tryptophan or a positively charged amino acid next to the amide group as this would stabilize the charge-transfer state. The intense

peak observed for His-Tat at pH 4 suggests that the protein is less quenched possibly because it is more extended, consistent with the FTIR results. As the pH increases, the protein becomes less-extended, which might result in efficient charge transfer between the Trp ring and amide quenching the fluorescence intensity. Another possible explanation is that aggregation of the protein could change the local electric field near the tryptophan because it's likely that Tat is aggregating at elevated pH values based on the observation of increased light scattering in the fluorescence and absorption spectra.

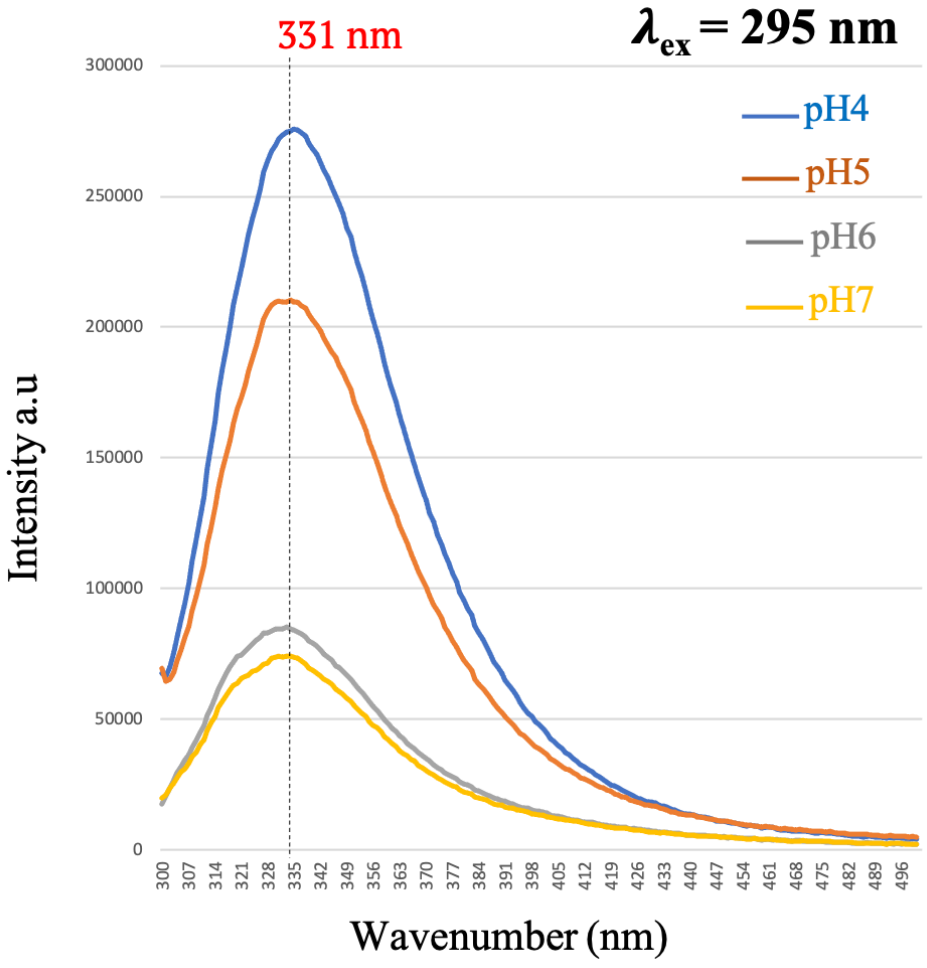


Figure 40. The baseline-subtracted fluorescence emission spectra of His-Tat at different pHs. pH4- Blue, pH5- Orange, pH6- Grey and pH7- Yellow respectively.

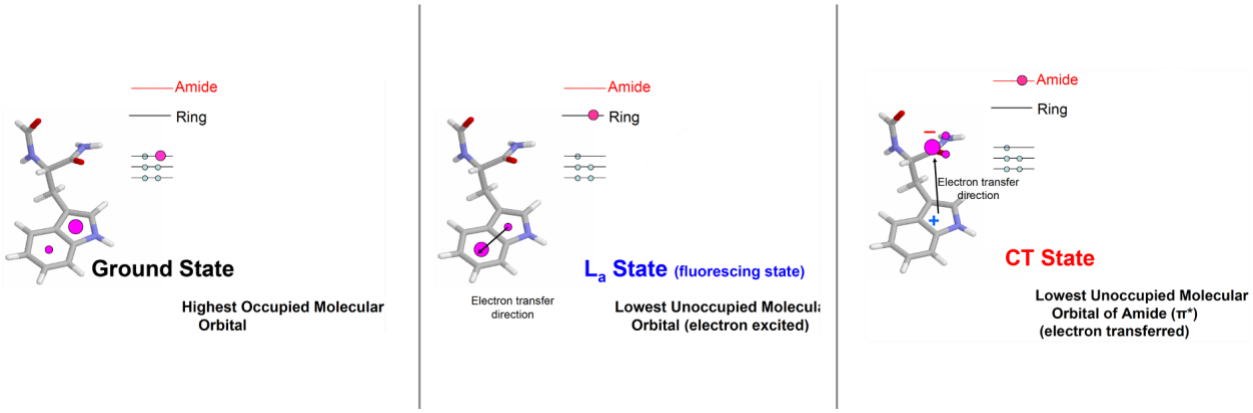


Figure 41. Schematic view of electron transfer between a tryptophan ring and an amide causing fluorescence quenching [319].

4.1.3 Quantitative analysis of Tat secondary structure by FTIR Spectroscopy

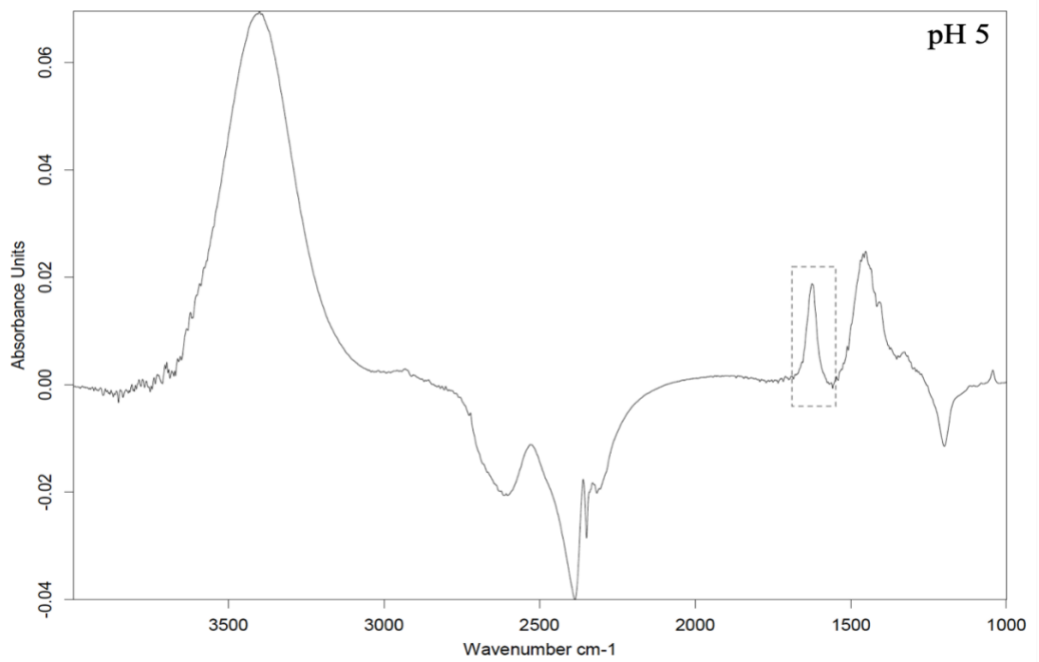
To obtain more detailed information about the conformation of Tat at pH 4, and changes in the conformation caused by raising the pH, FTIR spectroscopy was applied. Protein samples for FTIR analysis were prepared as described in **Section 3.13**. The amide-I band, ranging from 1600 to 1700 cm^{-1}, mainly represents stretching vibrations of the C=O bond with minor contributions from the stretching of C-N groups [236]. The amide-I region is focused on as it is the most intense band and is sensitive to small variations in the molecular orientation and hydrogen bonding patterns in proteins diagnostic of secondary structure. The band is deconvolved mathematically by Fourier self-deconvolution, a curve-fitting process [317]. The main features of the band that are considered during curve fitting are, the height of the band, its position (wavenumber) and the bandwidth at half-height. The values of these parameters are iterated during curve fitting until the envelope of the fit matches the FTIR spectrum of the protein. The iteration is constrained according to the characteristic bandwidth of each secondary structure as indicated in **Table 17**.

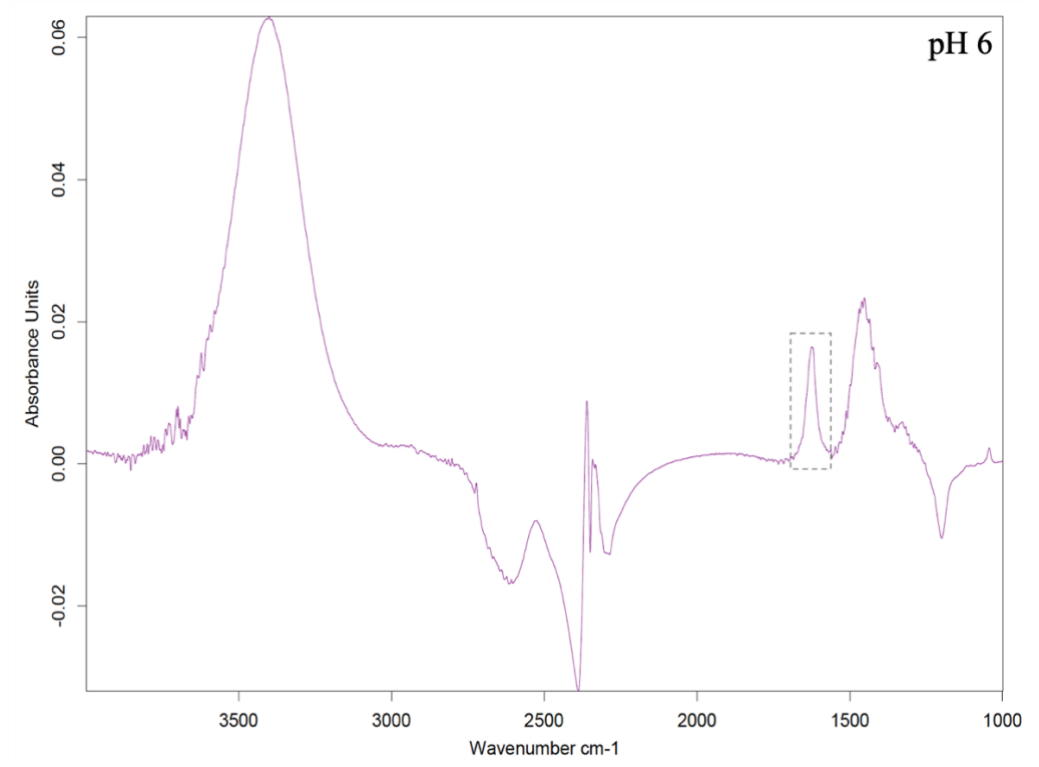
Table 17. Mean amide-I band wavenumbers corresponding to different protein secondary structures in an aqueous medium [317].

Mean Wavenumber (cm^{-1})	Secondary structure
1624 ± 1.0	β-Sheet
1627 ± 2.0	β-Sheet
1633 ± 2.0	β-Sheet
1638 ± 2.0	β-Sheet
1642 ± 1.0	β-Sheet
1648 ± 2.0	Random
1656 ± 2.0	α-Helix
1663 ± 3.0	3_{10}-Helix
1667 ± 1.0	β-Turn
1675 ± 1.0	β-Turn
1680 ± 2.0	β-Turn
1685 ± 2.0	β-Turn
1691 ± 2.0	β-Sheet
1696 ± 2.0	β-Sheet

FTIR spectra of Tat protein acquired in water at pHs 5 to 7 are given in **Figure 42**. The bandwidth-constrained deconvolution of the amide-I region is shown in **Figure 43** and the fractions of secondary structures present in the protein obtained by integrating the areas under the curve for each component used in the deconvolution are listed in **Table 18**. The fraction of β-sheet changes very little with increasing pH values whereas 3_{10}-helix content is eliminated at pH 7. On the other hand, the fraction of random coil and β-turn conformation consistently increase as the

pH is elevated. The low content of random coil is surprising but may indicate that at low pH where the protein has a high net positive charge that much of the backbone exists in an extended conformation.





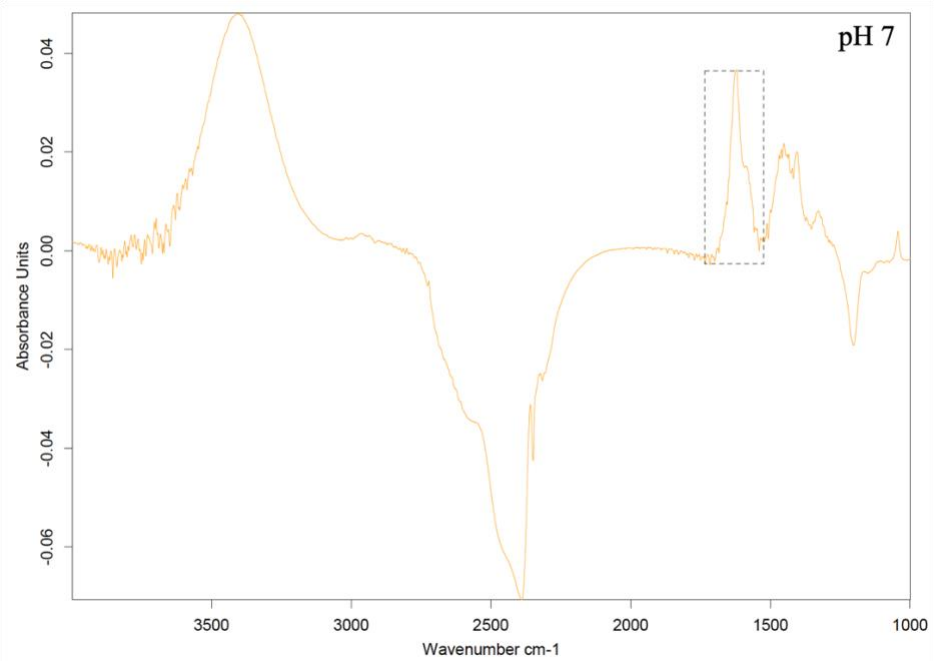


Figure 42. FTIR spectra of His-Tat at pHs 5, 6 and 7. The black dashed region is the amide I region **(**1700−1600) cm^{-1} used in deconvolution.

Table 18. Relative fractions of secondary structures in the His-tagged protein at different pH conditions as determined from FTIR spectroscopy.

	Assignment/ Wavenumber (cm^{-1})			
pH	**β-Sheet[a]**	**Random[b]**	**β-Turn[c]**	**3_{10}-Helix[d]**
5	61.78	2.18	4.36	26.55
6	58.66	4.51	5.76	24.59
7	60.25	13.47	25.69	

[a]1628/1694±2 cm^{-1}
[b]1648±2 cm^{-1}
[c]1682±3 cm^{-1}
[d]1664±3 cm^{-1}

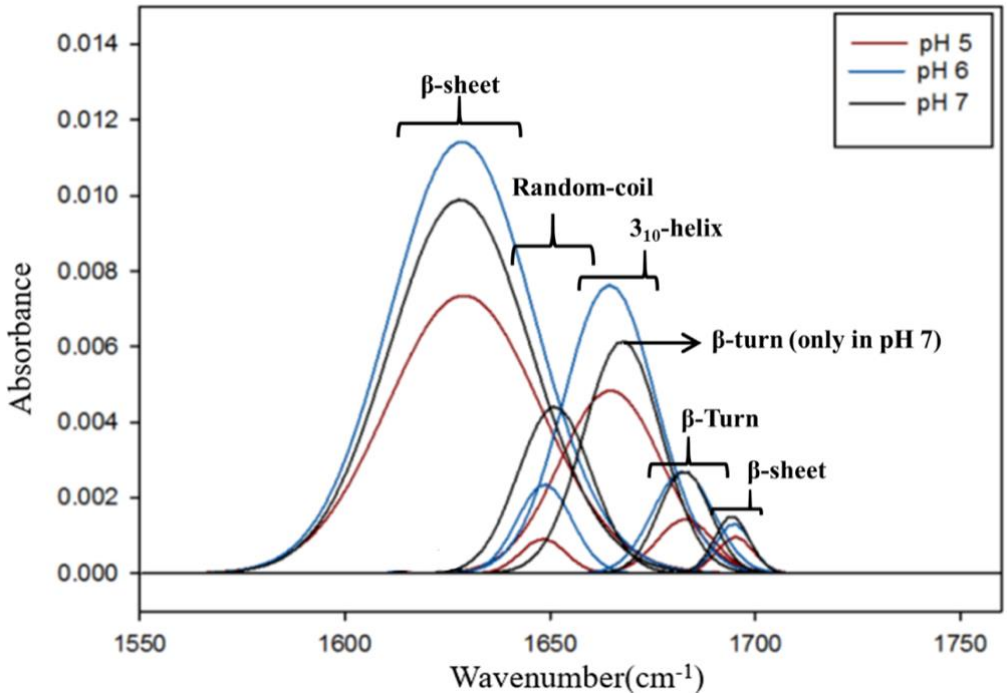


Figure 43. Deconvoluted FTIR spectra of Tat protein (5 mg/mL) at pH 5 (red), 6 (blue) and 7 (black).

4.1.4 NMR analysis of His-tagged Tat protein

4.1.4.1 Proton (^{1}H) NMR:

A water-supressed ^{1}H NMR spectrum of His-tagged Tat protein at pH 4 is shown in **Figure 44**, with resonances corresponding to different functional groups from the amino acid residues labelled. The amide region (7 to 10 ppm) is of particular interest in protein NMR as the dispersion in the chemical shift of the resonances in this region and the number of resonances observed can be diagnostic of folded and unfolded protein. While the amide region of a folded protein is populated with well-dispersed peaks and the resonances from most amino acids, for unfolded proteins the region is populated with a much smaller number of peaks, in a much narrower region and exhibiting severe overlap. Moreover, in unfolded proteins the amide protons chemically

exchange with bulk water, diminishing peak intensity. Amide peaks are visible in **Figure 43** because at pH 4 H-exchange with water is at its minimum.

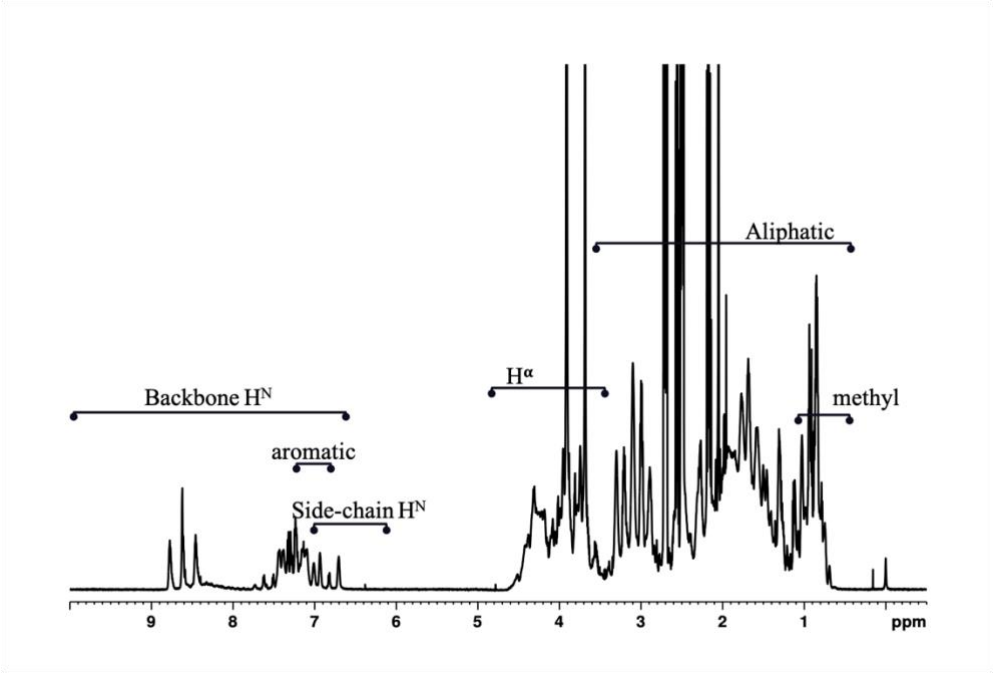


Figure 44. ^{1}H water suppressed spectrum of 400 µM His-Tat at pH 4 with 32 co-added transients at 298 K.

4.1.4.2 ^{1}H-^{15}N Heteronuclear Single Quantum Coherence (HSQC) Spectrum of His-Tat

^{15}N HSQC is a standard experiment in protein-NMR which helps in identifying the ^{15}N atoms ^{1}J-coupled to protons. This is a simple, yet an effective, experiment that generates a ^{1}H-^{15}N correlation map, which functions as a fingerprint region that is unique to every protein [134]. The Tat protein carries high net positive charge at pH 4 and hence, it is highly soluble at pH 4. The ^{1}H-^{15}N HSQC spectrum of the full-length protein at pH 4 has been assigned by To *et al.* [137] (**Figure 45**), and I used it as a template to assign spectra of Tat that I acquired. The solubility of Tat protein decreases as the pH increases and at pH 7, the protein aggregates and precipitates from solution. Among the 20 naturally occurring amino acids, only few are titratable (see **Table 19**),

i.e., they possess functional groups in their side chains, whose protonation and deprotonation is solution-pH dependent. These functional groups are responsible for changes in the net-charge of the protein and therefore, affect its solubility. By titrating the protein and tracking the chemical shift correlation changes in the ^{1}H-^{15}N HSQC NMR spectrum, a qualitative understanding of the structural changes leading to aggregation and precipitation may be obtained. His-Tat comprises 121 residues including the His-tag and 5 Asp, 6 Glu, 10 His, 9 Arg, 12 Lys, 7 Cys and 2 Tyr residues are available for titration. Since the pH range of the titration experiment was 4 to 7, the Arg, Lys and Tyr amino acids are not expected to undergo a large change in ionization state. The Lys and Arg will contribute to the protein's net positive charge as they remain positively charged over this range. To try to understand the origin of the protein's insolubility at neutral pH, a pH titration using NaOH was carried out and ^{15}N-^{1}H HSQC spectra were collected at different solution pH values for structural interpretation.

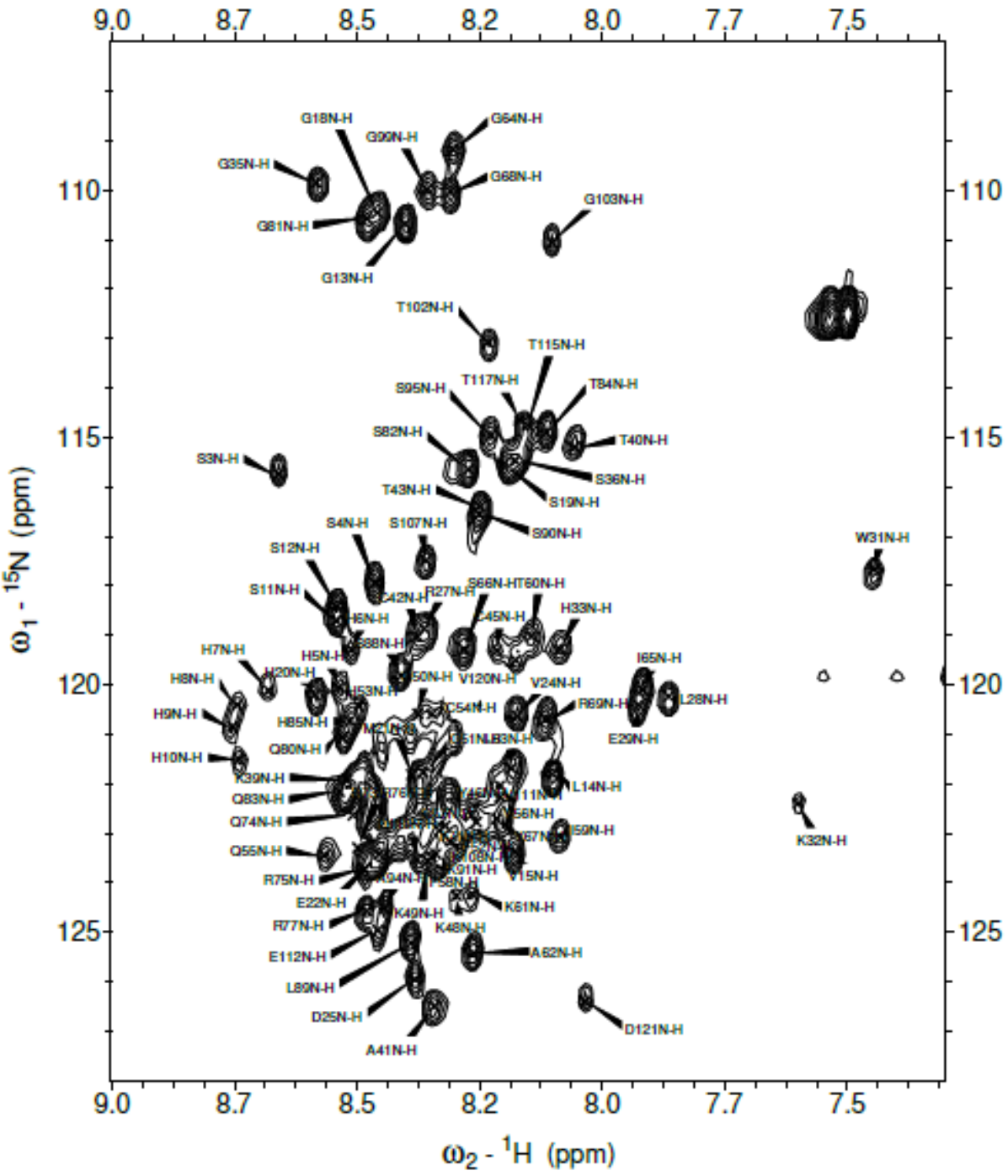


Figure 45: ^{1}H-^{15}N HSQC spectrum of ^{15}N-labelled 400 µM His-Tat protein at pH 4 with 64 coadded transients at 298 K. The cross-peak assignments are based on those deposited in the BMRB (BioMagnetic Resonance Bank) by Vu To [137].

Table 19. pKa of side chain of amino acids which affect the net charge.

Amino acids	One-letter code	Side chain pKa of amino acids at 25 ℃
Aspartic acid	D	3.9
Glutamic acid	E	4.3
Histidine	H	6.0
Lysine	L	10.8
Arginine	R	12.5
Cysteine	C	8.33
Tyrosine	Y	10

The ^{1}H-^{15}N HSQC spectra of the His-tagged Tat protein at different pH values are overlayed in **Figure 46** and are individually plotted in **Appendix-II**. The theoretical net-charge of the protein at pH 4 is +29.0, which drops to +23.1 as the pH is increased to 5. While no changes are observed in the amide (NH) correlation peaks of glycine residues, a notable change is that all of the histidine resonances in the His-tag disappeared at pH 5 (**See Figures 46 – 48**). Below the pKa's of the histidines, the His-tag will have a high net local positive charge, a higher local pH and faster rates of base-catalyzed H-exchange so it's not surprising that the His-tag resonances are among the first to disappear as the pH is elevated. Another region of high positive charge is the region from Y67 – R77 and these resonances also are lost by pH 5 as are those of the 3 lysines between residues 107 - 109. These observations support the conclusion that peak loss at pH 5 is dominated by elevated H-exchange. As the pH is further increased to pH 6, the net-charge drops to +17.6 and many cross-peaks including those of glycine and serine are lost. Two possible explanations for this observation are aggregation-mediated broadening of the peaks and chemical exchange with bulk water. Although there is modest broadening of the peaks shown in Figure 45 the dominant effect

of elevating the pH is peak intensity loss making diagnosis of aggregation difficult. At pH 7, the protein precipitates as the net-charge drops to +12.2, which severely compromises the ability to collect a quality NMR spectrum. As a result, very few peaks are observed. The tryptophan, lysine and aspartic acid residues positioned at 31, 32 and 121 in the sequence show significant changes in their chemical shifts, making them easily identifiable (**Figure 46**) even at pH 7. As mentioned above, there is fluorescence and NMR chemical shift evidence to suggest that the region around W31 is structured and the slowing of H-exchange in this region might explain the observation of W31 and K32 at pH 7. In the Cys-rich region, some peaks are lost early in the titration while others are lost later making it difficult to assign any of the changes as caused by Cys oxidation. A minimum protein concentration of 400−500 μM is required to obtain a spectrum with good signal-to-noise ratio but at these high concentrations the protein is less soluble at pH 7. Attempts were made at collecting the spectra using solutions of lower protein concentrations but they were unsuccessful. A pictorial representation of the protein sequence with amino acids that are assigned at different pHs is given in **Figure 48**. Relaxation dispersion experiments conducted by To *et al.* [121,137] on full-length Tat support the conclusion that the disappearing resonances at elevated pH is due to elevated H-exchange. Unfortunately, no insights were gained as to the residues responsible for the aggregation at pH 7.

.

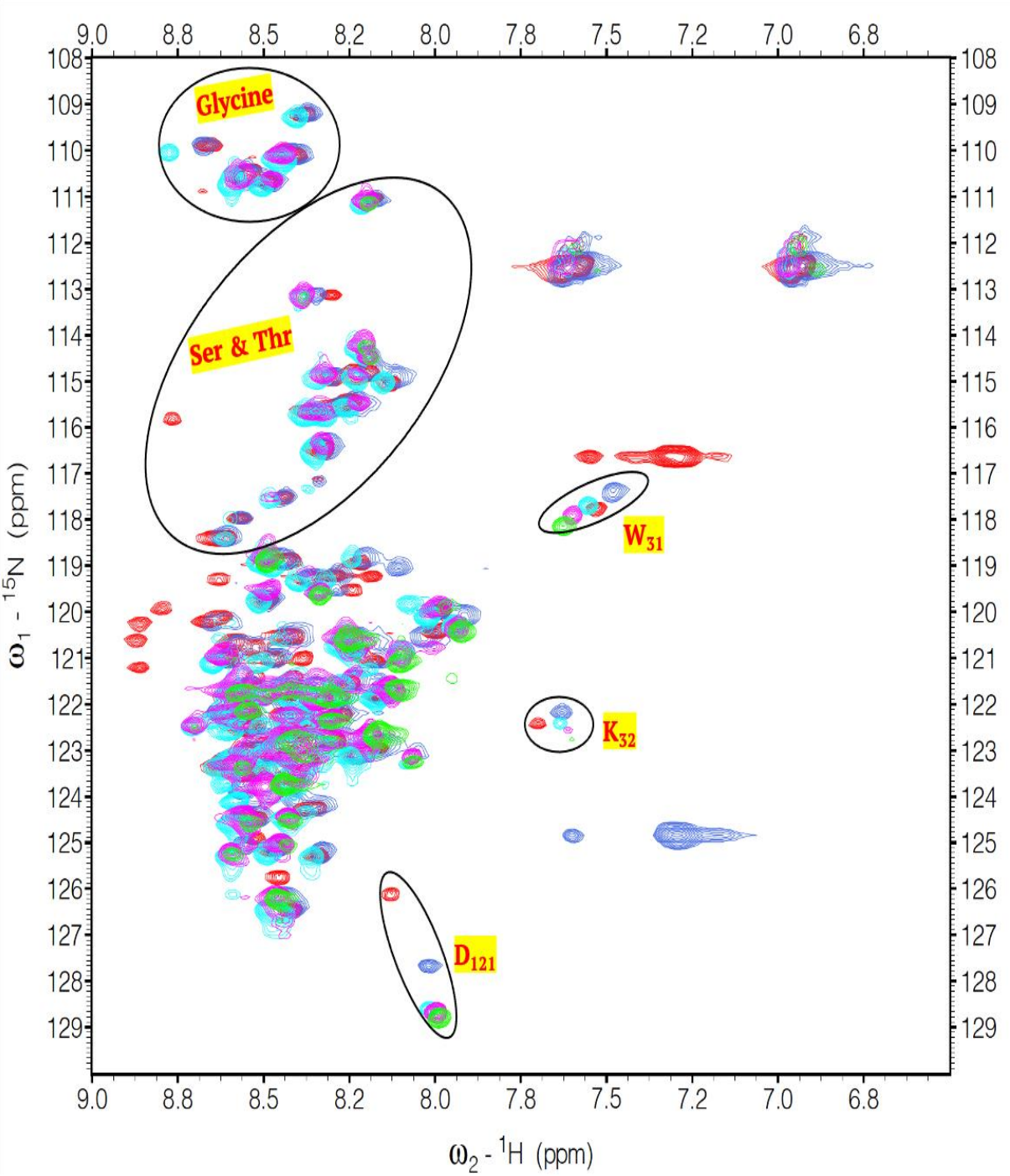


Figure 46. An overlay of the ^{1}H-^{15}N HSQC spectra of His-Tat protein at different pHs with 64 coadded transients: pH 4, Red (400 µM); pH 5, Royal blue (250 µM); pH 6, Cyan (150 µM); pH 6.5, Magenta (120 µM); pH 7, Green (100 µM).

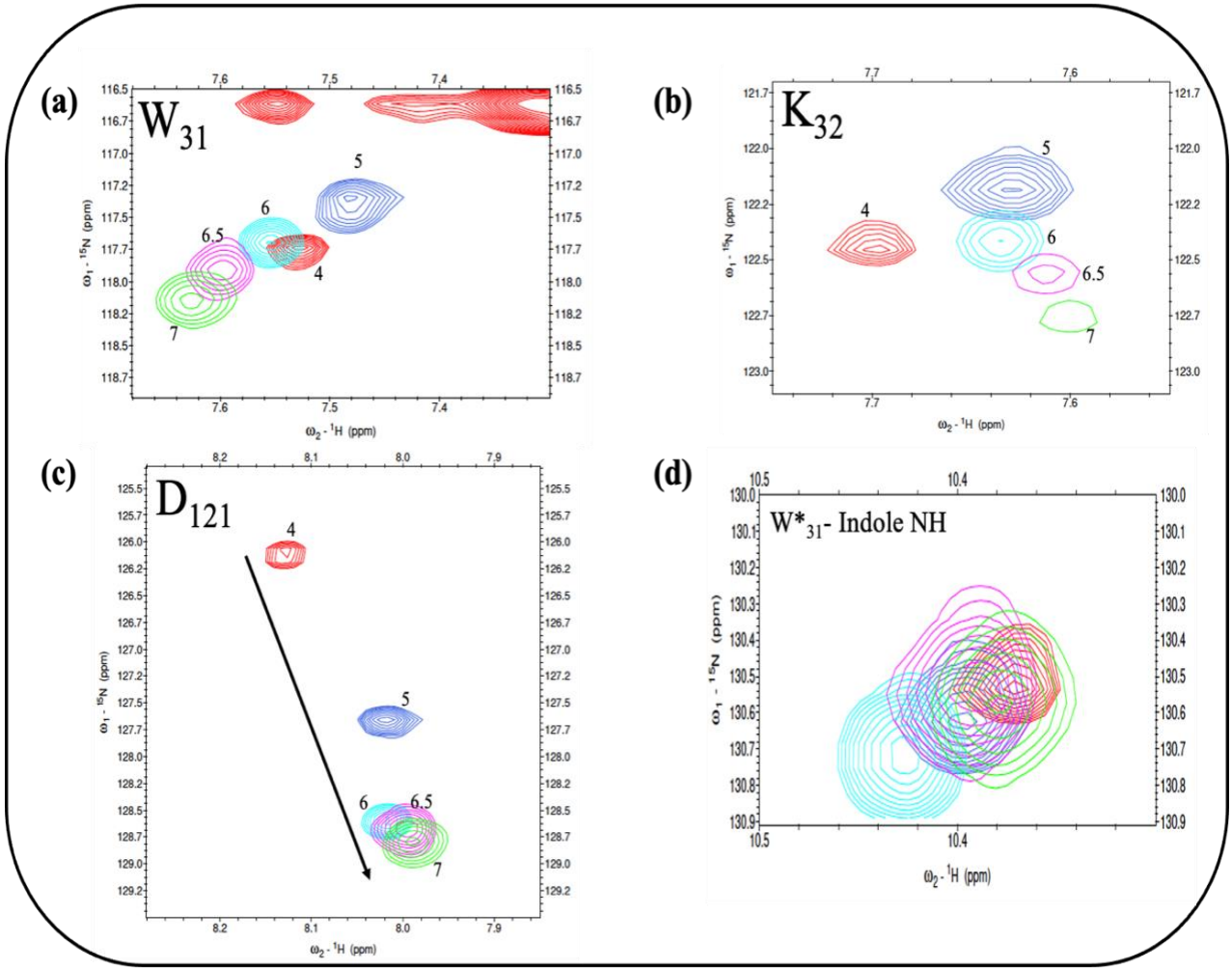


Figure 47. An overlay of the ^{1}H-^{15}N HSQC spectra of amino acid residues at different pH. (a) W31, (b) K32, (c) D121 and (d) W*31-Indole NH

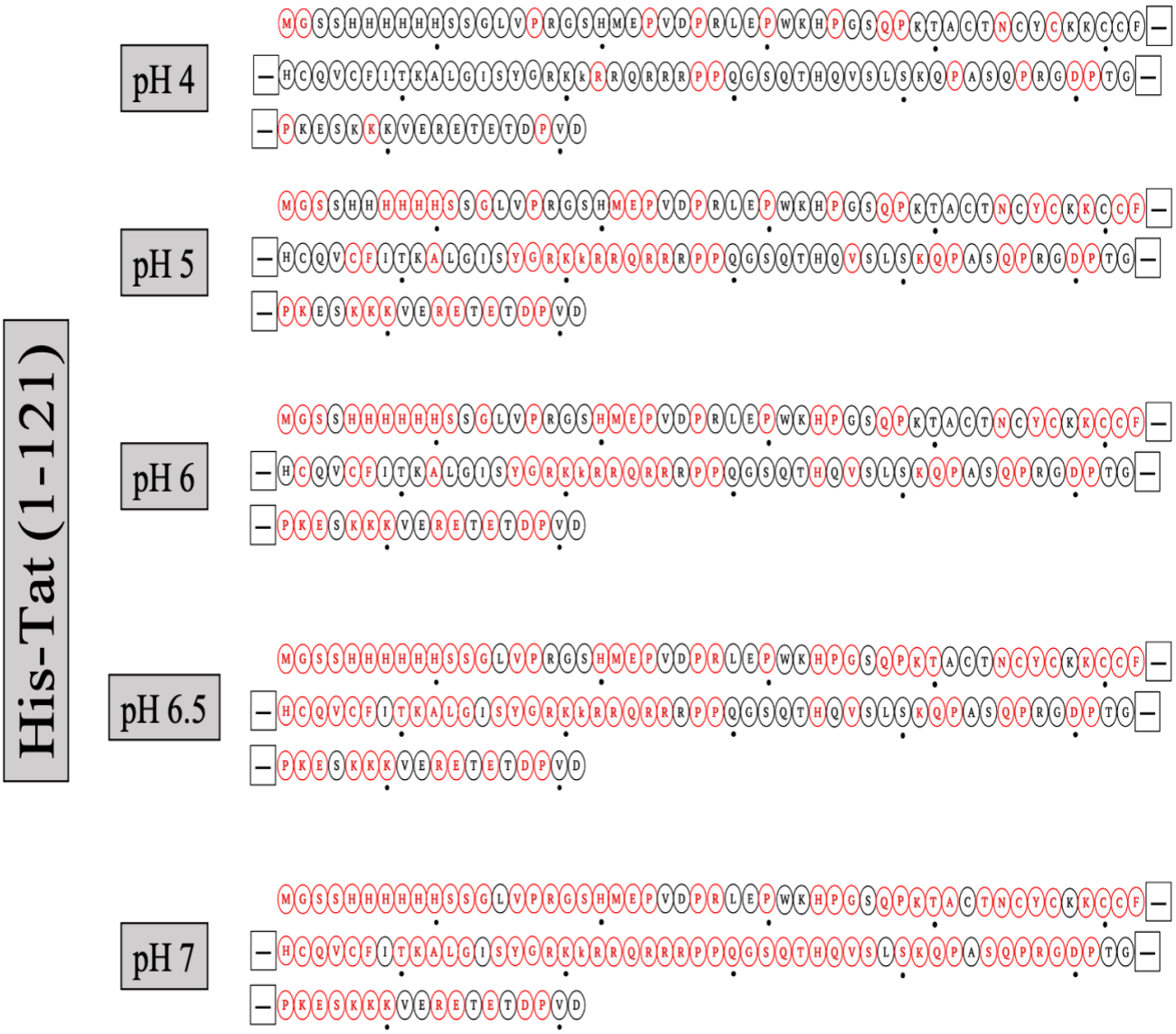


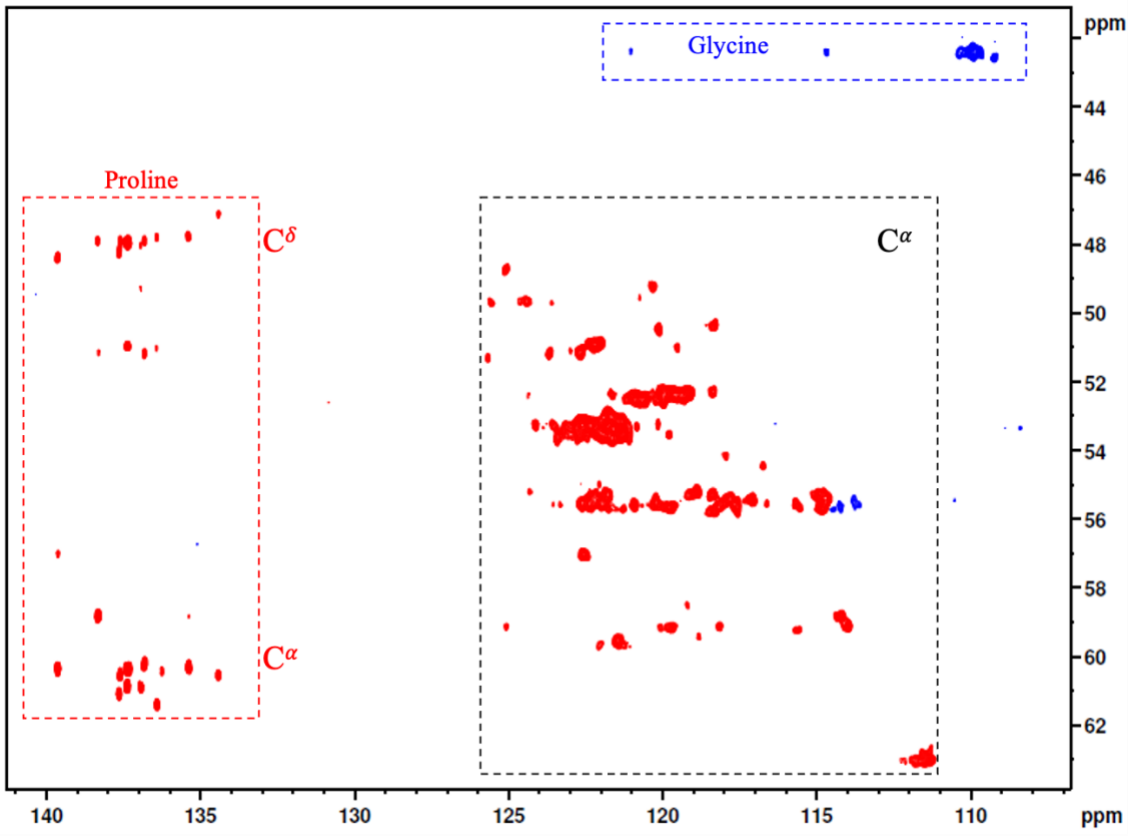
Figure 48. Graphical representation of the disappearance NMR peaks as the pH is elevated. The black coloured circles show the assigned peaks and the red coloured circles show the missing or not assigned cross peaks at different pHs. The black dots indicates every 10th residue. Note that proline cannot be assigned owing to its absence in ^{1}H-^{15}N HSQC experiments.

4.1.4.3 ^{15}N Direct-detection of His-Tat:

In biomolecular NMR, challenges such as sensitivity, resolution and severe line broadening due to the fast transverse relaxation of interested spins are often faced. Especially in unfolded proteins, this relaxation problem brought about by the chemical exchange and slow tumbling of high molecular weight systems is a serious issue. To overcome this inherent problem, new approaches have been developed using direct detection of low-γ nuclei. Direct detection of ^{15}N has the added advantage of permitting the detection and assignment of proline resonances that are not observable in the ^{1}H-^{15}N-HSQC experiment. The hCaN experiment is a ^{15}N-direct detection experiment which has been successfully adapted for observing and assigning amide ^{15}N resonances of prolines wherein.

A two-dimensional hCaN NMR spectrum of Tat protein at pH 4 is shown in **Figure 49** and consists of intense cross-peaks from the proline residues in the region between 46 and 52 ppm in the ^{13}C dimension and between 133 and 140 ppm in the ^{15}N dimension (highlighted with a red dotted-rectangle) [293]. The cross-peaks from the ^{15}N-$^{13}C_{\alpha}$ cross-peaks representing glycine residues are marked with a blue rectangle and the correlations produced by the rest of the residues are shown in black rectangle. In the first part magnetization transfer scheme (**Section 2.4.9, Figure 33**), the proton magnetization is transferred from $^{1}H^{i}_{\alpha}$ to $^{13}C_{\alpha}$. Glycine residues have C_{α} methylene protons (an IS_2 spin system) compared to other amino acids which have methine groups at the C_{α} (an IS spin system). As a consequence, at the end of the refocusing time, observable single-quantum coherences related to glycine residues are at a null, while the coherences of IS (CH) systems are at a maximum [292]. Due to this, glycine signals are opposite in phase compared to the other amino acids. This is shown with the use of a different colour (blue) in **Figure 49**. Thirteen C^{α}-N crosspeaks are observed suggesting the detection of all thirteen proline residues in the

protein. The wide range of peak intensities suggests varying backbone dynamics throughout the protein. However, the cross-peaks could not be assigned due to severe signal overlap in the rest of the spectrum. One of the drawbacks of the direct-detection hCaN experiment is that it has relatively low sensitivity. Although it would be worthwhile pursuing further hCaN experiments at pH 7 where fast H-exchange diminishes cross-peak intensity in HSQC experiments we did not pursue this any further for Tat because of its very low solubility at pH 7.



Figures 49. 2D hCaN NMR spectrum of 400 μM Tat-protein at pH 4 with 64 coadded transients at 298 K.

4.1.4.4 Solid-state NMR of His-Tat

Solid-state MAS NMR has been successfully employed to study the structure and dynamics of many biological molecules such as protein complexes [320], membrane proteins [321] and amyloids [322]. Due to non-averaging of anisotropic interactions, and, more importantly, fast relaxation rates, the peaks are often broad in ssNMR. Hence only the nuclei with slow relaxation rates can be easily studied. ^{1}H experiments cannot be employed as the resulting peaks are very broad due to severe dipolar coupling which cannot be averaged even with fast spinning conditions. ^{13}C experiments are attractive as the ^{13}C nuclei have slow relaxation rates but they suffer from inherently low gyromagnetic ratios and low natural abundances. To improve the sensitivity, the dipolar-based through-space polarization-transfer experiment was adopted wherein the polarization is transferred from ^{1}H to ^{13}C and the latter nucleus is observed. The cross-polarization (CP) MAS experiment can be used to detect static disordered regions [297,321] in proteins. The time for polarization transfer, called the contact time (τ_{CP}), is on the order of a few microseconds to milliseconds, depending on the strength of the dipolar interaction, which is affected by atomic dynamics. Thus, the CP build-up curves are heavily dependent on the internal motions of the sample. When the samples are rigid or packed together, their CP build-up curves will be significantly faster compared to molecules experiencing significant motion. Thus, rapid isotropic motions diminish the CP signals, meaning that fast tumbling averages out the dipolar interactions. The INEPT is another J-coupling (scalar) based experiment to detect dynamically disordered regions in IDPs [297,321]. When these two experiments are used in conjunction, dynamic and static regions of proteins can be determined [321]. Since the Tat protein precipitates at pH 7, solid-state NMR experiments of the precipitate or the protein in a lyophilized form could provide key structural information which would help in understanding the reason behind protein precipitation

and the structure of the precipitate. To determine the structure of the protein in the solid state, Tat was lyophilized at pH 4. The ^{1}H-^{13}C CPMAS spectrum of full-length His-tagged Tat protein lyophilized at pH 4 is shown in **Figure 50**. Peaks in the 10 to 75 ppm region correspond to aliphatic carbons whereas the peaks in the 105 to 165 ppm region are attributed to aromatic carbons. An intense peak at 180 ppm marks the carbonyl carbons.

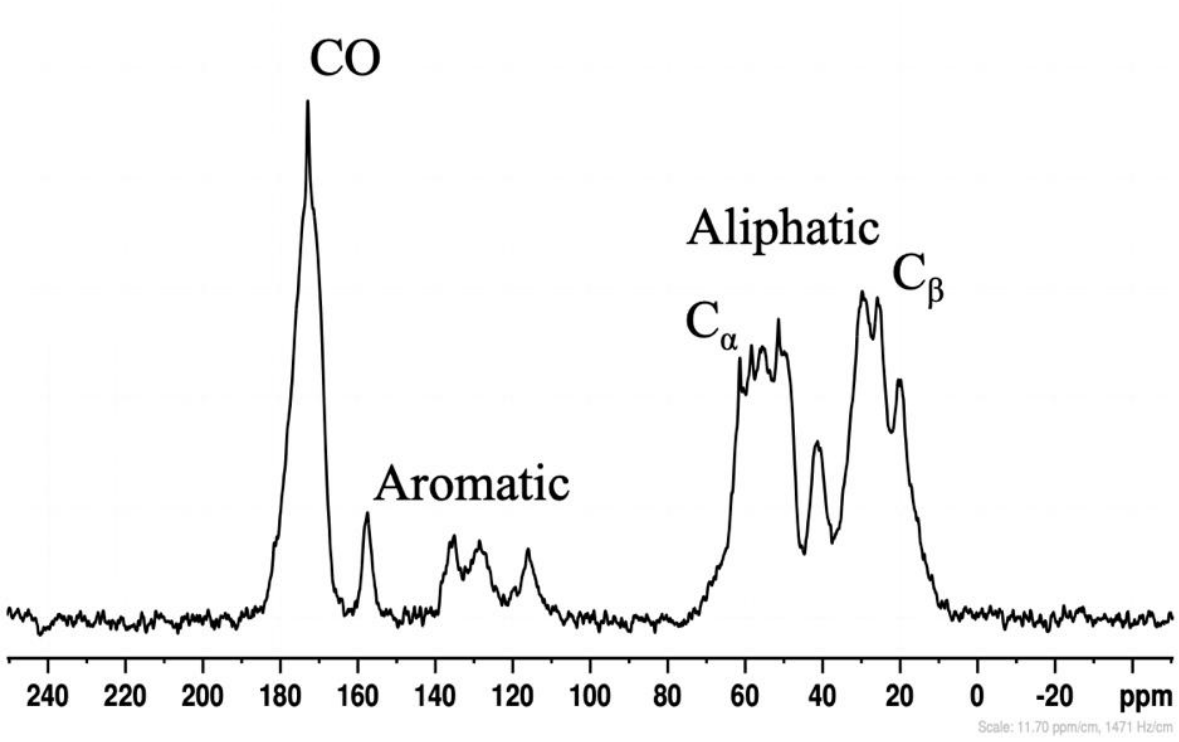


Figure 50. One-dimensional ^{13}C-CP-MAS NMR spectrum of lyophilised His-Tat at pH 4. The sample was spun at the magic-angle at 15 kHz and the spectrum was collected with 1024 co-added transients.

Figures 51 and **52** show INEPT and CPMAS NMR spectra recorded at different temperatures. The peak positions and intensities in the CPMAS spectra (Figure 49) did not change with temperature suggesting a lack of dynamic regions in the protein sequence across the range of pH values studied. The absence of peaks in the INEPT spectra (Figure 50) confirms this conclusion

that there are no flexible regions in freeze-dried Tat between pH 4 and 7. On a separate note, tight packing of the sample into the NMR rotors might have hindered the flexibility of the protein to some extent. However, since the packing method I used is similar to that in the published literature it seems likely that sample packing is not the reason why no dynamics are discernable in freeze-dried Tat.

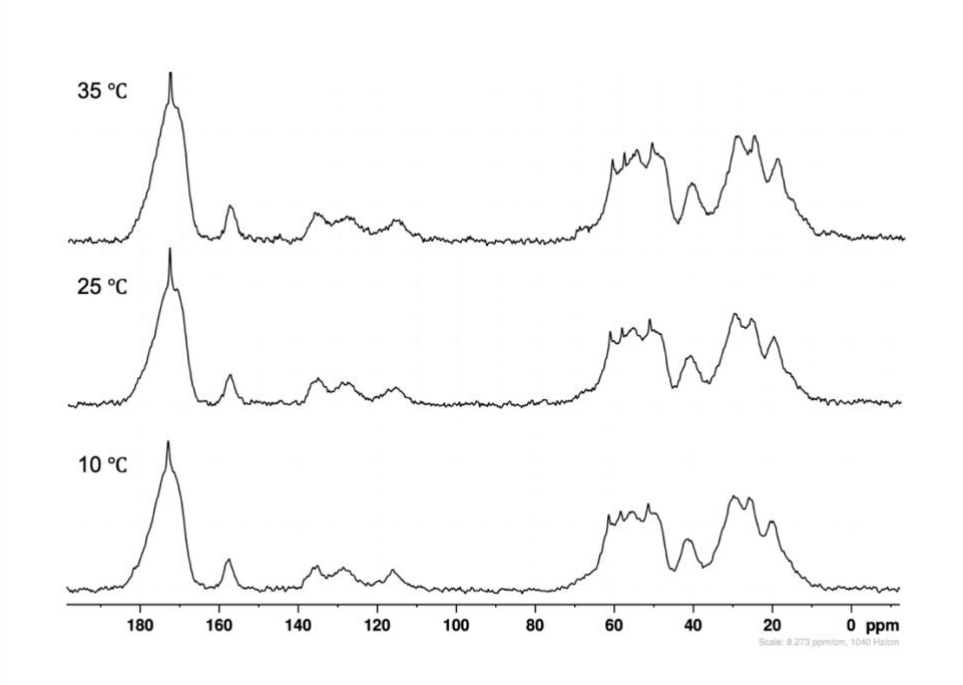


Figure 51. ^{1}H-^{13}C CPMAS spectra of ^{13}C-^{15}N labelled lyophilized Tat protein recorded at different temperatures.

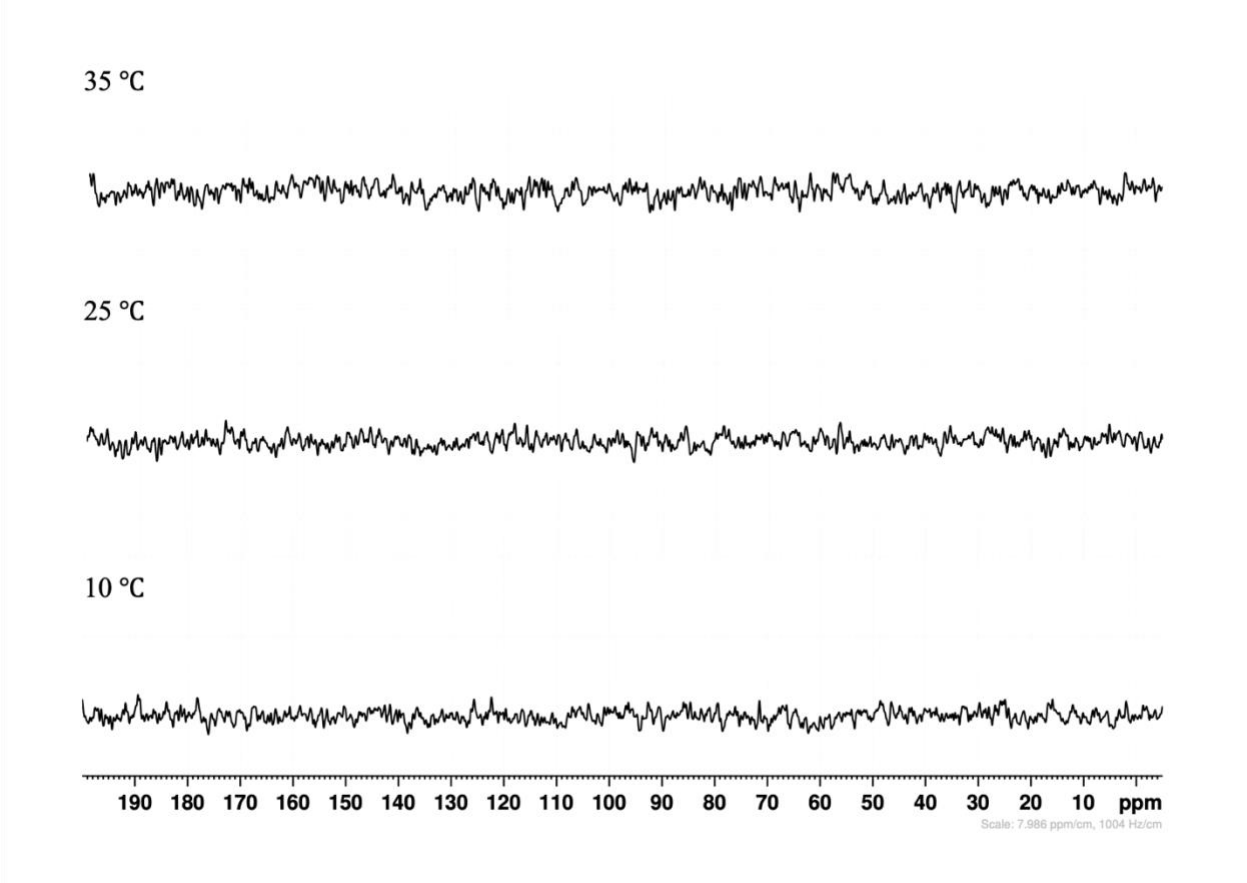


Figure 52. ^{1}H-^{13}C INEPT experiments recorded at different temperature at 10 ℃, 15 ℃ and 35 ℃ on ^{13}C-^{15}N labelled lyophilized Tat at pH 4 under MAS at 15 kHz.

A two-dimensional PDSD spectrum of Tat freeze-dried from a pH 7 solution is shown in **Figure 53**. This experiment involves the transfer of polarization from ^{1}H to the low-gamma carbon nuclei, which is further transferred to other carbons which are close in space, all through dipolar interactions. This is a standard solid-state NMR experiment wherein the observation of cross-peaks connecting intra and inter residue carbon nuclei depends on the mixing time. Short mixing times generate intra-residue correlations, while long mixing times are required to establish inter-residue connectivity. Although this is an attractive and efficient experiment with good resolution, severe signal overlap was observed in the spectrum of the Tat protein, making the assignment very

challenging and uncertain. Thus, regions with resolved resonances that could be unambiguously assigned were identified. The chemical shift range between 58-60 ppm is assigned to C_α and C_β of Thr/Ser and the peaks at ca. 40 ppm are assigned to the alpha carbons of glycine. Because of the poor resolution of the crosspeaks in this spectrum I did not pursue this avenue further.

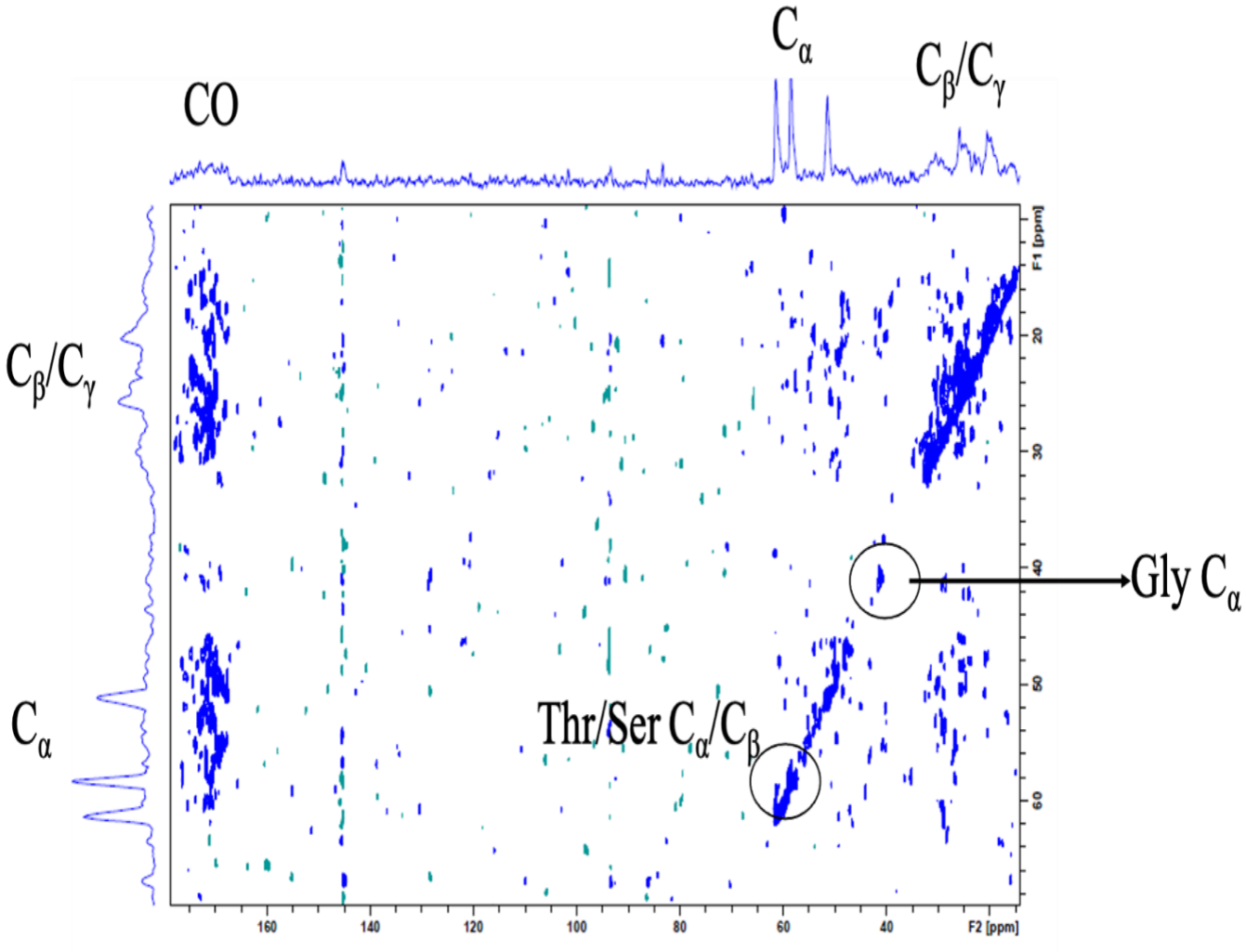


Figure 53. ^{13}C-^{13}C PDSD spectrum of freeze-dried Tat measured at 250 ms mixing time acquired with a MAS spin rate of 15 kHz and 128 scans.

4.2 NMR spectroscopy of Histidine tagged Tat-Torula RNA complex

Tat is a highly basic protein with many positively charged residues such as Arg and Lys. The high net-charge of the protein at pH 4 causes repulsion of Tat monomers and prevents protein-aggregation but, at physiological conditions, the net charge decreases leading to aggregation and precipitation.

In theory, the aggregation could be prevented by complexing the Tat protein with a macromolecule which is water-soluble at pH 7. To test this hypothesis, the Tat protein was complexed with Torula yeast RNA, a polyanionic macromolecular mixture. This may increase the solubility of the protein and give an opportunity to decipher the structural and dynamic information of Tat under physiological conditions. Since the RNA is not isotopically labelled, no interferences are expected spectroscopically. In solubility testing I was initially able to make a 100 μM (1 mg/mL) Tat solution at pH 6.5 which also contained Torula yeast RNA at 10 mg/mL in 0.1 M acetate buffer. An initial attempt of complexing Tat protein and Torula RNA in 1:1 mass ratio at pH 6.5 did not result in high-quality NMR spectra (not shown). The mass ratio was then increased to 1:3 and the NMR experiments were carried out.

An overlay of the ^{1}H-^{15}N HSQC NMR spectra of ^{15}N labelled His-Tat with and without Torula RNA at pH 6.5 is shown in **Figure 54**. The spectrum of Tat-protein alone has a significantly smaller number of cross-peaks than the spectrum at pH 4 (Figure 3.6), due to the limited solubility of the protein and possible aggregation. Also, the resonances are broad which underscores the difficulty associated with analysing aggregated systems. On the other hand, the quality of the HSQC spectrum of the protein-RNA complex is only marginally improved with more cross-peaks observed in the fingerprint region. This improved solubility may be attributed to a non-specific interaction between the poly-anionic RNA and positively charged amino acid residues in the

protein. Some of the new low-intensity cross peaks that appear in the spectrum suggest the possibility that the Tat conformation is more heterogenous in the mixture than alone and this is a significant drawback of this approach. Thus, although this approach appears to improve the solubility of Tat because no major improvements in the NMR spectra were observed, this was not pursued further.

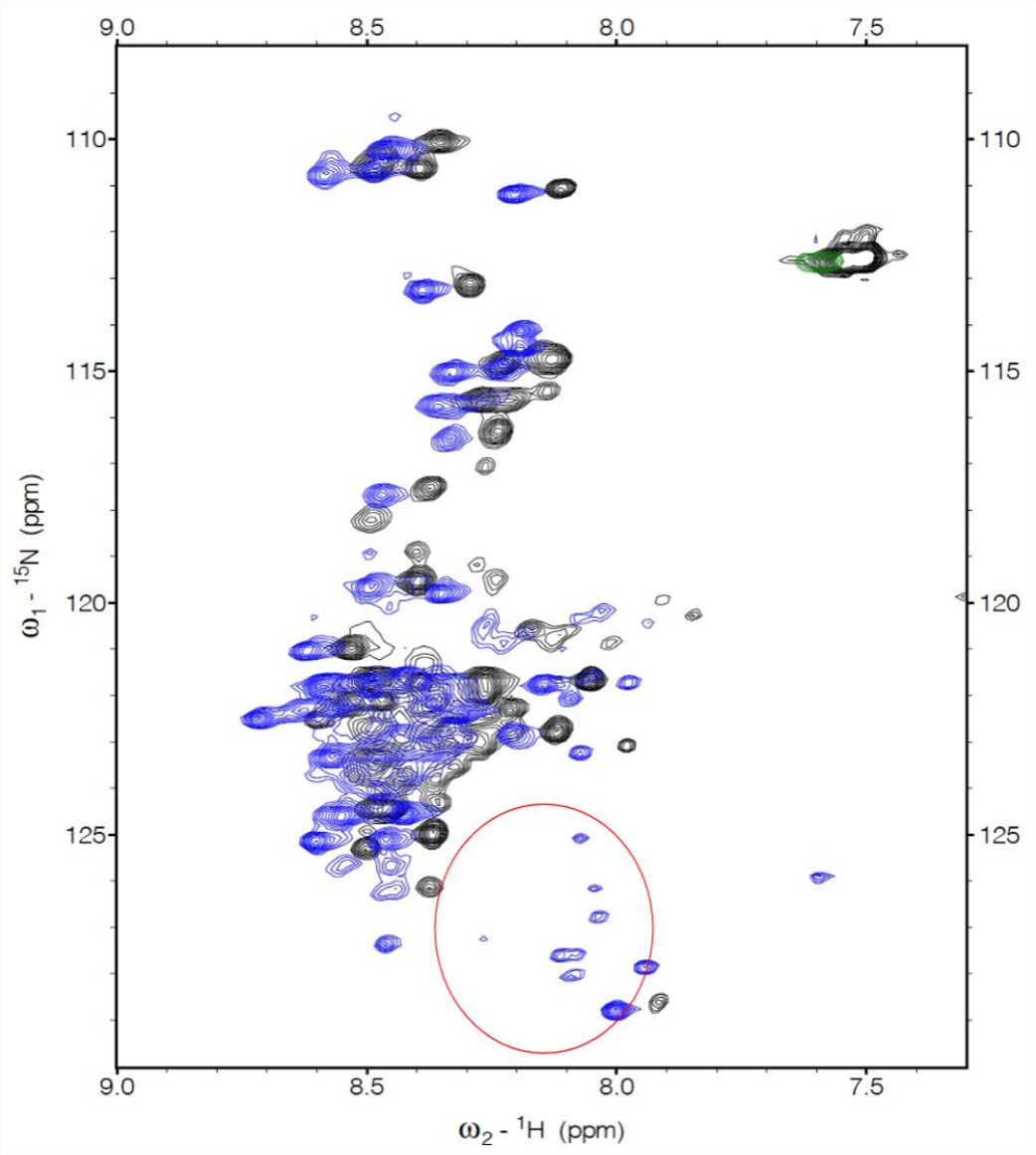


Figure 54. An overlay of the ^{1}H-^{15}N HSQC spectra of the ^{15}N-labelled 500 µM Tat-protein with RNA (blue) and without RNA (black) at pH 6.5 with 64 coadded transients. The cross peaks encircled in the red colour are new resonances that appear in the spectrum near Asp-121.

4.3 *In-vitro* Transcription of Transactivation Response (TAR) RNA

Being a disordered protein, Tat binds to a diverse class of cellular proteins and nucleic acids. TAR RNA is one such 590-nucleotide long nucleic acid, which has a stem-loop structure and is situated at the 5' ends of all HIV-1 viral transcripts (**Figure 55**). Upon binding to the Tat protein, it recruits the transcription elongation factor b complex (contains Cyclin-T1 and CDK-9 proteins), which enhances the viral replication of the HIV-1 virus by hyperphosphorylation of RNAP II.

4.3.1 Tat-TAR interaction

In TAR RNA, two helical stem regions are separated by a three-nucleotide pyrimidine bulge [323,324] (**Figure 55**). The trinucleotide bulge specifically binds to the basic region of Tat with a dissociation constant (K_D) of = 6×10^{-9} M^{-1} [325]. Tat binds to G26 and a pair of phosphates and the interaction is stabilized by formation of a base triple between U23 and A27-U38 [325,326].

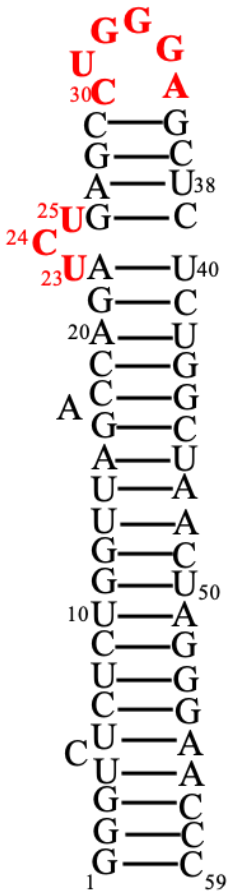


Figure 55. The sequence of the 59-nucleotide base pair TAR RNA. The major UCU trinucleotide bulge and loop regions are highlighted in red.

In the previous section, the effects of complexing Tat protein with Torula RNA on the solubility and NMR spectra were explored. This approach showed a significant improvement in the solubility of Tat at pH 6.5 but only a marginal improvement in the NMR spectra. Unlike cellular Torula RNA, TAR RNA is a single molecule known to interact specifically with Tat with high-affinity. Here, I describe the results of attempts to *in vitro* synthesize TAR and form a Tat-TAR complex for study by NMR spectroscopy. Synthesis of TAR RNA involves many steps and reagents and is explained in detail in Section 2.9. One of particular interest is the T7 RNA

polymerase II (RNAP II), an active enzyme responsible for the transcription process. Details of its expression and purification are given in **Section 3.8**. An image of the 10% SDS-PAGE gel showing the purity of the T7 RNAP II is given in **Figure 56**.

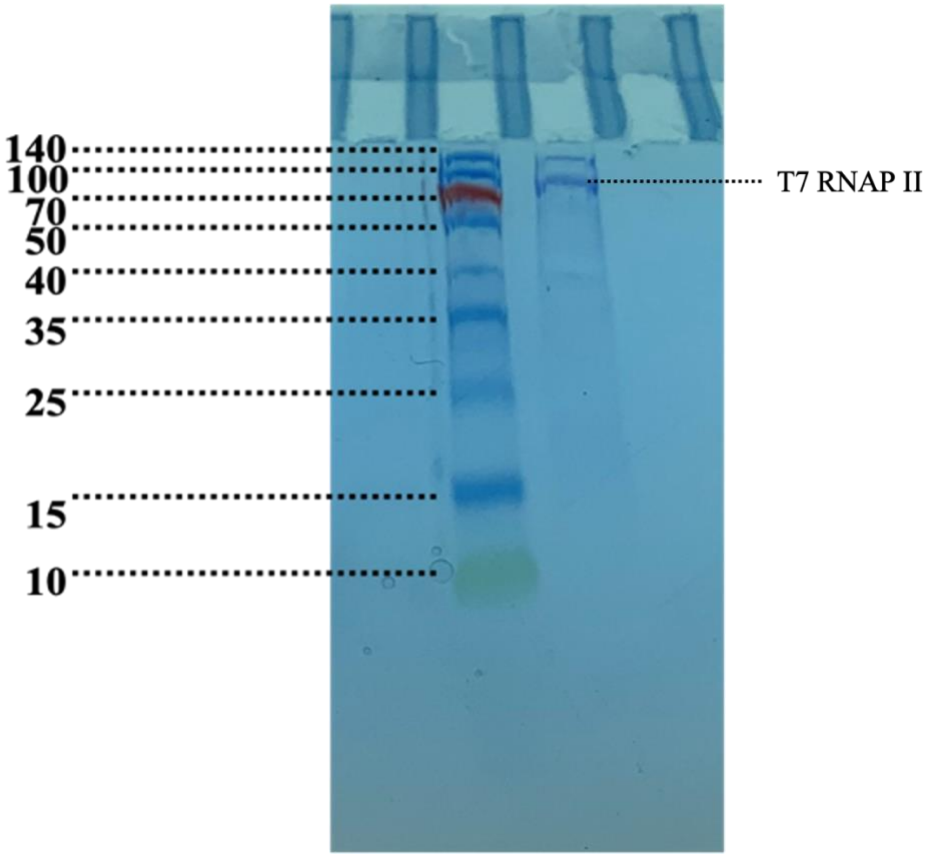


Figure 56. An image of a 10% SDS-PAGE gel showing the purity of the synthesized T7 RNAP II protein. The molecular weight of T7 RNAP II is 98 kDa [327]). The left lane shows the protein ladder and to right is the T7 RNAP II. The band between 70 and 100 kDa corresponds to T7 RNAP II.

4.3.2 $MgCl_2$ Optimization

Two active sites present in RNA polymerase II bind divalent Mg^{2+} ions, which aids in nuleotidyl transfer *via* metal ion catalysis. Optimizing the concentration of Mg^{2+} is an important step in transcription reactions including the PCR technique. The optimization trials were carried out in 3 mL reaction buffer by adding different volumes of 100 mM $MgCl_2$. Thick bands of TAR are observed when 5 and 7 μL Mg^{2+} solution are added (**Figure 57**), underscoring the minimum volume of Mg^{2+} solution required for the transcription process. Hence, 5 μL Mg^{2+} solution was used for the actual transcription process with a reaction buffer volume of 10 mL, and a thick band was observed in the gel electrophoresis (**Figure 58**). The transcribed RNA was extracted by the phenol-chloroform method and rinsed with desalting buffer to remove unused nucleotide triphosphates (NTPs) and any traces of phenol. After this step, the RNA was purified by Fast-protein liquid size-exclusion chromatography. The middle peak in the elution profile (**Figure 59**) corresponds to TAR RNA and the peaks to the left and right of it correspond to the plasmid and free NTPs, respectively. The TAR RNA was collected from a pool of fractions and the sample was concentrated using ultrafiltration centrifugation with appropriate molecular weight cut-off filters. The purified TAR RNA was used for NMR experiments.

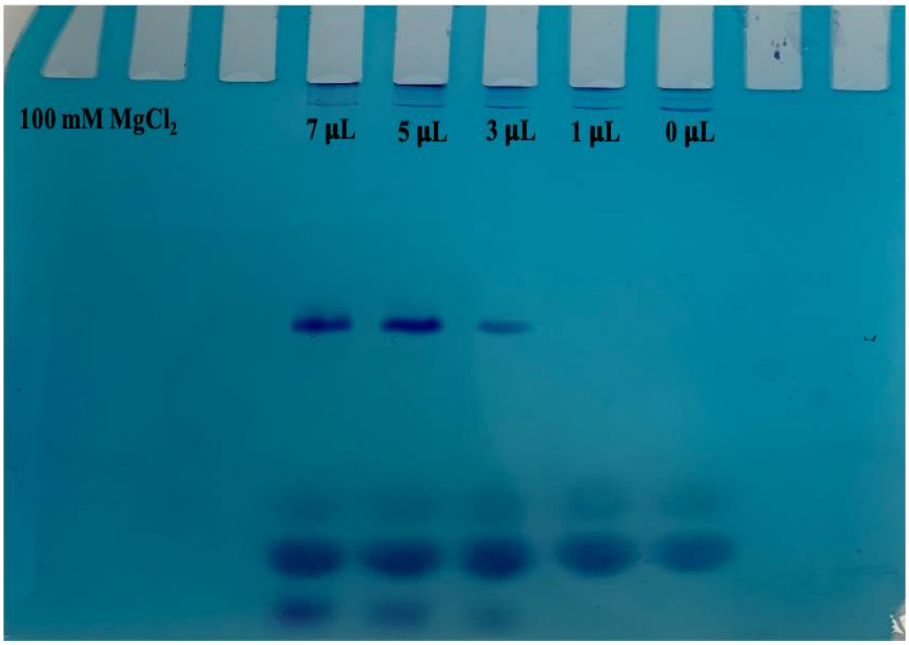


Figure 57. Results of 1-hour $MgCl_2$ trials of transcription of HIV-1 TAR RNA on a 10% denaturing gel.

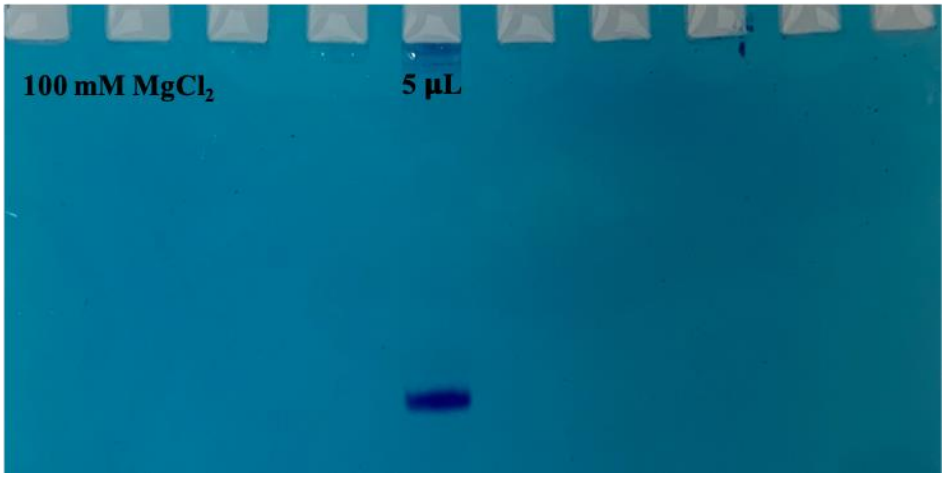


Figure 58. Results of 3 hours of transcription of HIV-1 TAR RNA on a 10% denaturing gel.

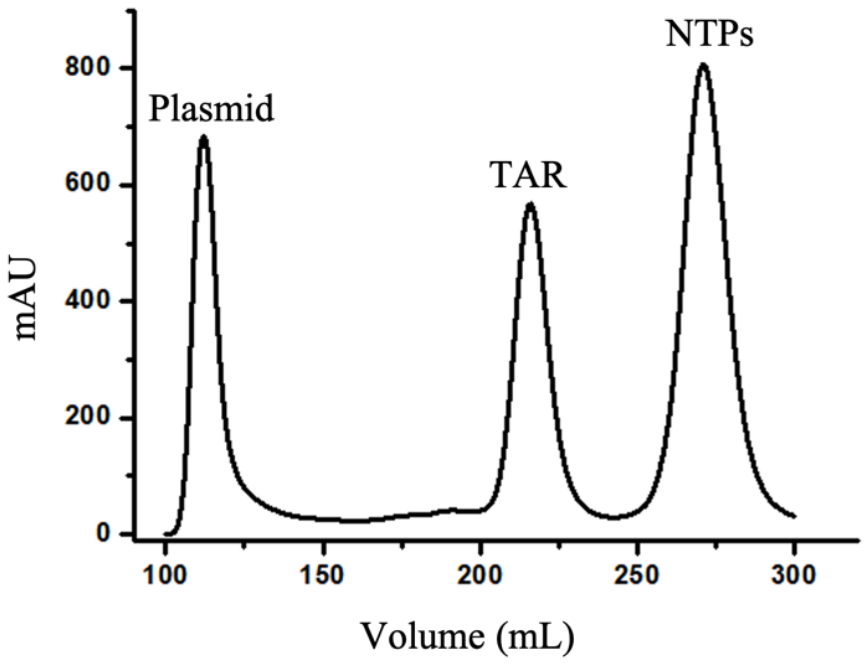


Figure 59. Size exclusion FPLC elution profile of a HIV-1 TAR RNA transcription reaction showing plasmid, TAR RNA and free NTPs.

4.3.3 NMR Spectra of Tat-TAR RNA Complex

Based on the hypothesis that TAR RNA could improve the solubility of the Tat protein, a pH titration experiment was designed wherein, TAR RNA was added to the Tat protein solution. Although the Tat protein is soluble at pH 4, precipitation was observed immediately upon adding the TAR RNA (**Figure 60**). Similar experiments at higher pH values also resulted in precipitation. This suggests that to form a stable Tat protein-TAR RNA complex, other co-factors that are involved in the transcription process are also required to bring about a positive effect on the Tat protein solubility. Since an improvement in the Tat protein's solubility was observed in the presence of Torula RNA, the possibility of improvement in the protein solubility in the presence of multiple polyanions was explored. To the solution of Tat protein, TAR RNA

and Torula RNA were added. At first, 120 μM (1.5 mg/mL) of Tat was dissolved in the presence of 4.5 mg/mL of Torula yeast RNA at pH 6.5 and a ^{1}H-^{15}N HSQC spectrum was recorded. The following day, 120 μM of TAR RNA was added in 1:1 molar ratio to the same solution. No improvements in either the solubility of the protein or the quality of the spectrum were seen (**Figure 61**). The ^{1}H-^{15}N HSQC spectrum of the Torula-Tat-TAR mixture is essentially identical to the spectrum of the Torula RNA and Tat protein mixture. Although Tat protein interacts strongly with TAR RNA *in vivo*, no changes in the spectral signature underscores the difficulty and challenge associated with translating *in vivo* phenomena into *in vitro* experiments.

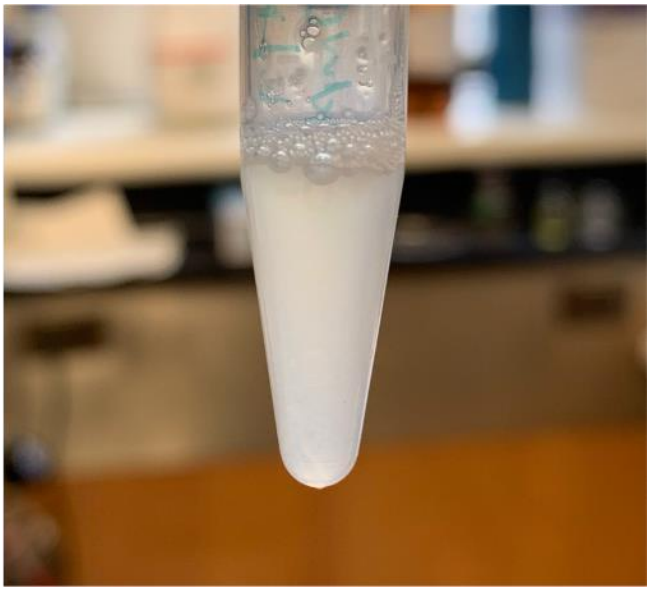


Figure 60. Solution of Tat protein and TAR RNA at pH 4 showing precipitation. A similar result was observed at higher pH values as well.

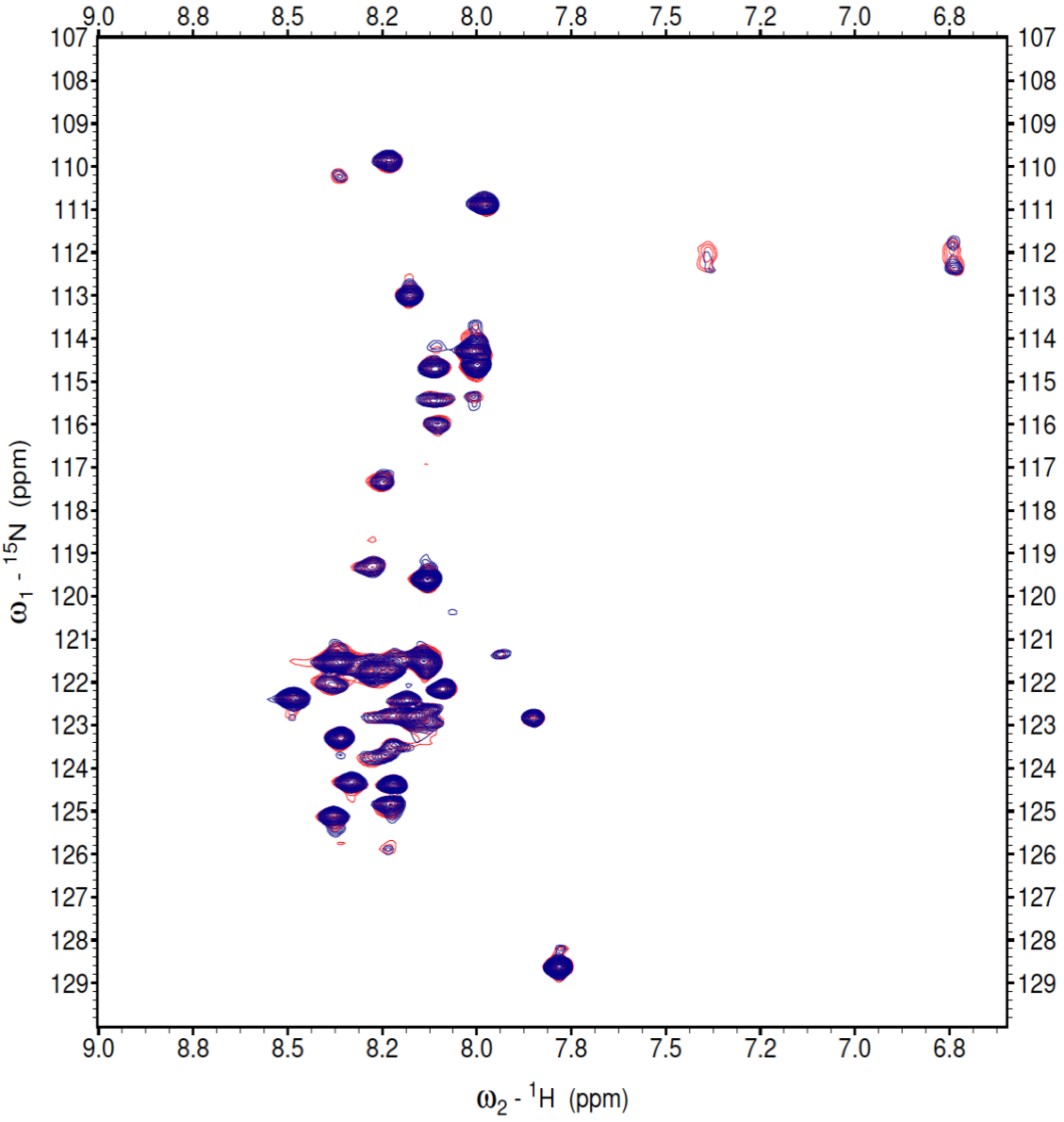


Figure 61. ^{1}H-^{15}N HSQC spectrum of the ^{15}N-labelled 200 µM Tat-protein bound to TAR RNA (1:1) and torula RNA at pH 6.5 with 64 coadded transients. Blue - Tat: Torula RNA complex, Red- TAR:Tat:Torula RNA complex.

4.4 Mass spectrometry of Sequence Specific Nickel Assisted (SNAC) Tat

The purified SNAC-Tat protein was dissolved in 10 mM ammonium formate buffer at pH 4. A detail description of sample preparation for MS analysis is given in **Section 3.13**. The obtained raw mass spectrum was deconvolved using the Maximum Entropy software [328]. **Figure 62** shows the ESI-MS/MS spectrum of uncleaved SNAC-Tat. The peak at m/z of 13395 corresponds to the SNAC-tagged Tat protein and the peak at 6698 represents the doubly charged protein (m/2z), which is exactly half the mass of the uncleaved protein.

The cleaved SNAC-Tat MS spectrum is shown in **Figure 63**. The large peak (m/z = 9882) corresponds to an unexpected internal cleavage within the sequence and the small peak (m/z = 11563) corresponds to the C-terminal cleaved Tat. Because of the unexpected cleavage at an internal cleavage site no further experiments were carried out on SNAC-Tat.

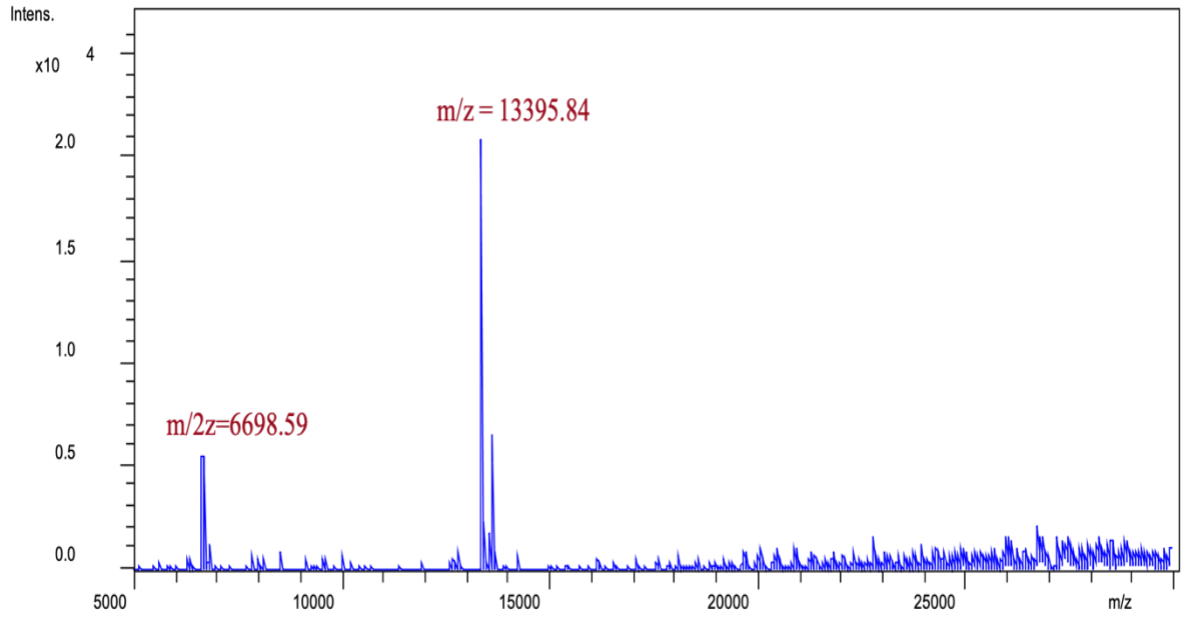


Figure 62. ESI-MS spectrum of Uncleaved SNAC-Tat protein (pH = 4; Tat protein: 100 mM; Ammonium formate buffer: 10 mM).

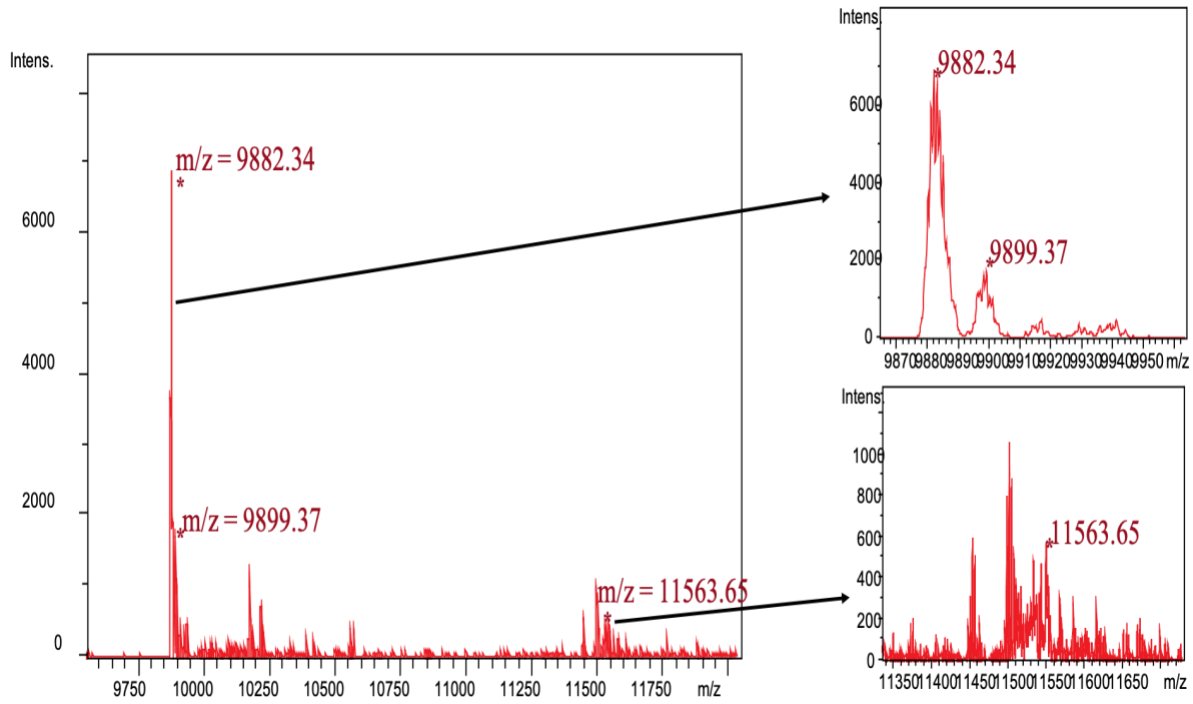


Figure 63. ESI MS spectra of cleaved SNAC-Tat acquired in pH 4 (100 μM protein concentration and 10 mM ammonium formate buffer).

4.5 Supercharged Tat (SuTat)

The net charge of the Tat protein appears to have a significant influence on its solubility as explained in previous sections. Since the protein has a high positive charge at pH 4 and is soluble, we postulated that the solubility at pH 7 can be enhanced by maintaining the high net charge even at pH 7. To accomplish this and to test the hypothesis, a supercharged protein was designed by adding a sequence containing 10 arginine residues (R_{10}) to the protein at the N-terminus (His-tag-R_{10}-Tat). The plasmid construction and sample preparation steps for NMR analysis are given in **Section 3.6**. The net charge of His-tagged Tat at pH 7 is +12.2, whereas the net charge of SuTat at pH 7 is +20.1 (**Table 20**).

The purity of the supercharged Tat protein at pH 7 was tested through SDS-PAGE analysis (**Figure 64**). The band around 17 kDa corresponds to the supercharged Tat protein, while the other bands at higher kDa are presumed to be from the aggregated Tat protein. This aggregation is believed to be a consequence of oxidation of cysteine disulphide bonds and possible interactions between hydrophobic side-chains (24). Even though the protein was charged by the addition of the polyArg sequence, the observed aggregation suggests that albeit the net charge is important, it is not the only factor responsible for the self-aggregation of the protein.

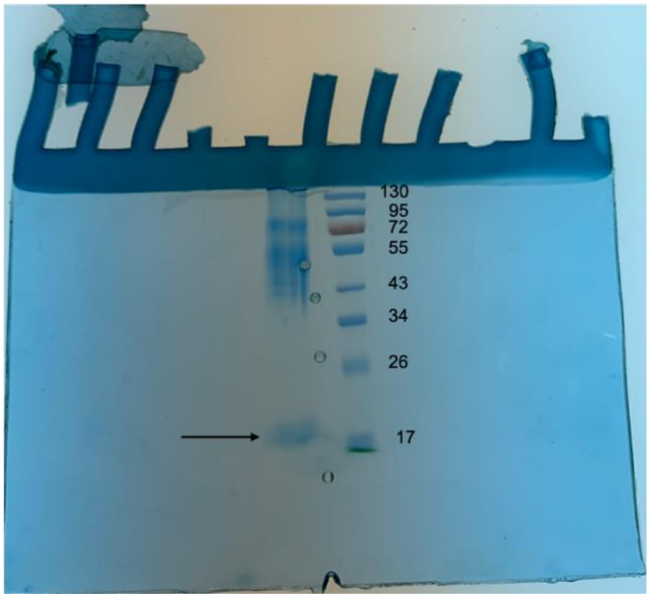


Figure 64. SDS-PAGE electropherogram of R_{10} supercharged Tat protein at pH 7.

Even though the net charge of protein was increased, the protein was poorly soluble at pHs other than 4, which led to no improvement in the quality of the ^{1}H-^{15}N HSQC NMR spectrum (**Figure 65**). The spectra of SuTat and His-tagged protein are similar in terms of the number of cross peaks and their chemical shifts so, increasing the charge of the protein did not lead to any improvement either in its solubility or in the quality of the obtained NMR spectra. Changes in the chemical shift of the D135 residue at different pH are similar in the case of His-tagged and supercharged protein, suggesting there is no effect of the supercharged tag on the solubility of the protein whatsoever. The individual ^{1}H-^{15}N HSQC NMR spectra obtained at different pHs are shown in **Appendix II**.

Table 20. Net charge of supercharge Tat at different pH.

pH	Net charge
4	+36.7
5	+30.5
6	+25.0
7	+20.1

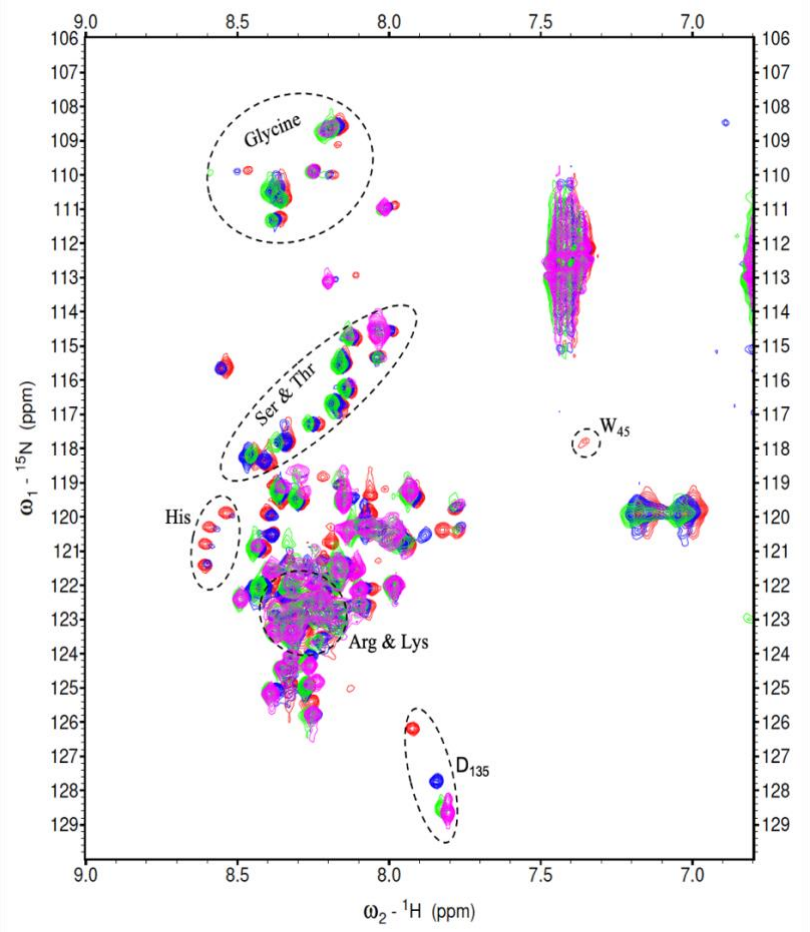


Figure 65. Overlay of ^{1}H-^{15}N HSQC spectra of Supercharged-Tat at different pHs with 90 coadded transients. pH 4 - Red (230 µM); pH 5 - Royal blue (190 µM); pH 6 - Green (150 µM); pH 7 - Magenta (100 µM).

4.6 Cysteine replaced Asp Tat mutant

The crystal structure of Tat and the p-TEFb complex reported in 2010 by Tahir *et al*. [9] revealed the interaction between the 1-49 Tat protein and the two subunits of p-TEFb, Cdk9 and Cyclin T1, which are essential for kinase activity. It was shown that 88% of Tat_{1-49} binds to Cyclin T1 and 12% interacts with Cdk9 (**Figure 66**). Cysteine plays a prominent role in this interaction and stabilizes the three-dimensional structure through the formation of two zinc binding sites or Zinc Fingers [9]. Earlier biochemical studies showed a zinc-mediated bridge forming between the cysteines of Tat and the 261st Cys residue of Cyclin T1 [329]. The crystal structure also showed the presence of two bound zinc ions, two helices (α-helix and 3_{10} helix), which are in the cysteine-rich and hydrophobic cores and the formation of a random coil tail. The zinc-binding motifs are cysteine- rich and are situated on either side of the 3_{10}-helix. The first zinc-binding motif consists of Cys-22, His-33, Cys 34 and Cys-37 residues, while the second motif includes Cys-25, Cys-27 and Cys-30 residues of Tat and Cys-261 of Cyclin T1. Several mutational studies have shown the importance of His-33 and 6/7 of the cysteine residues in functional studies of Tat [330,331].

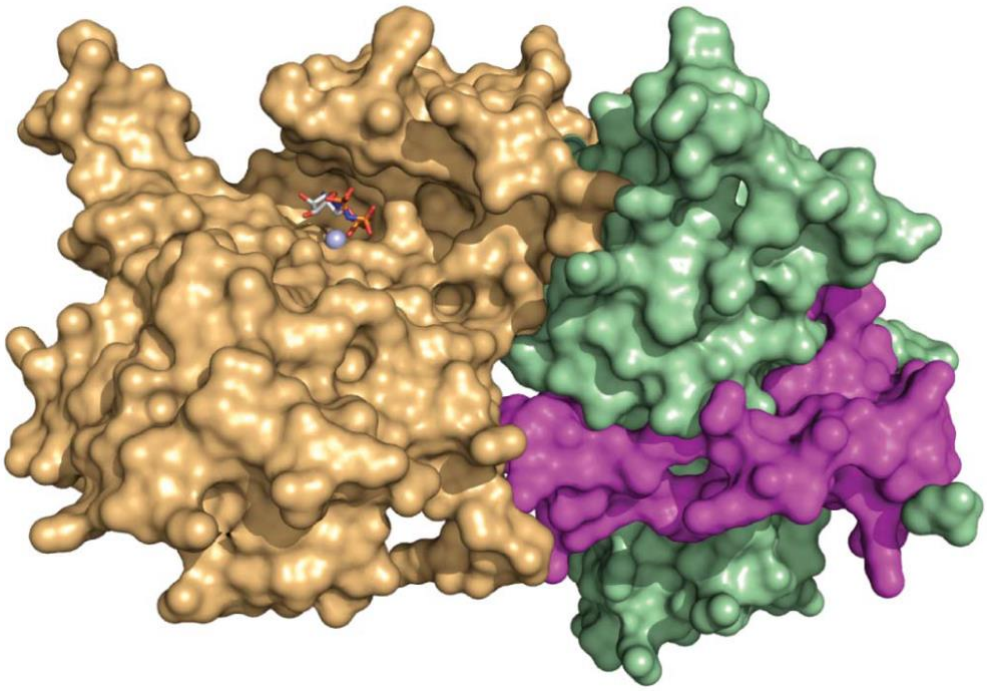


Figure 66. A pictorial representation showing interaction between Tat and positive transcription elongation factor b complex (Cyclin T1 and CDK-9). Tat is magenta, CDK-9 is light orange and Cyclin T1 is pale green. The sticks show side chain of CDK-9 interacting residues of Tat, Cys 261 of cyclin T1 and ATP analogue. *Reprinted with permission form reference* [9]. *Copyright © 2010 Springer Nature.*

The NMR studies of His-Tat and SuTat proteins containing 7 cysteine residues each resulted in poor solubility and aggregation at physiological pH. The abundance of cysteine residues could cause intermolecular disulfide bonds between the Tat monomers leading to aggregation as the pH of the solution approaches the pKa of cysteine ($pK_a = 8$). Even though a strong reducing agent (TCEP) was added to prevent the oxidation of cysteine, the efficacy of this reducing agent in preventing cysteine oxidation was never tested. Furthermore, some Cys residues may have lower pKa, especially in a locally electrostatically positive environment. Hence, to check whether the

cysteines are contributing to the aggregation, a mutant Tat sequence was designed where all of the cysteine residues are swapped with aspartic acid, referred to henceforth as *Asp-Tat* (**Figure 67**). In recent years, protein chemical modification has emerged as prominent problem-solving technique to address hurdles in biochemical research. Hypothetically, since the cysteine residues and the carboxyl side-chain of the aspartic acid are negatively charged, the substitution of Cys with Asp might lead to the same electrostatic interaction between the Zinc fingers and the Zn^{2+} ions as was the case for Cys possibly leaving the biological activity intact.

MGSSHHHHHHSSGLVPRGSHMEPVDPRLEPWKHPGSQPKTADTNDYDKKDDFHDQVDFI
TKALGISYGRKKRRQRRRPPQGSQTHQVSLSKQPASQPRGDPTGPKESKKKVERETETDPV
D

Figure 67. Amino acid sequence of Asp-Tat wherein all of the Cys residues have been replaced with Asp. The red colour-coded Asp (D) positions in the sequence represent Cys positions in the native protein.

4.6.1 UV-Vis Absorption spectrum of Asp Tat

The yield of Asp-Tat was typically about 12−15 mg per litre of culture, similar to Cys-Tat. The solubility of the Asp-Tat protein and its concentration in solution were determined from the UV-Vis absorption spectrum collected between 220 to 340 nm. The purified Asp Tat was dissolved in 10 mM HEPES, 10 mM acetate buffer at pH 4 and the pH was gradually increased to 7. The peptide bonds from the polypeptide chain absorb in the far-UV region (180-230 nm) [332] whereas, the aromatic residues such as tryptophan, tyrosine and phenylalanine absorb at 260-300 nm. From the absorption spectrum (**Figure 68**), the concentration of Asp Tat was found to be around 700 μM (ε_{280} = 8480 $M^{-1}cm^{-1}$). It should be noted that the solubility of other proteins discussed in earlier

sections was only 100 μM at similar pH conditions. Such high solubility at pH 7 is being reported for the first time, which suggests that that cysteines are in fact a major contributor to Tat aggregation. Note also in the spectrum in Figure 66 the lack of light scattering which is a hallmark of protein aggregation and the protein remain soluble more than 15 days.

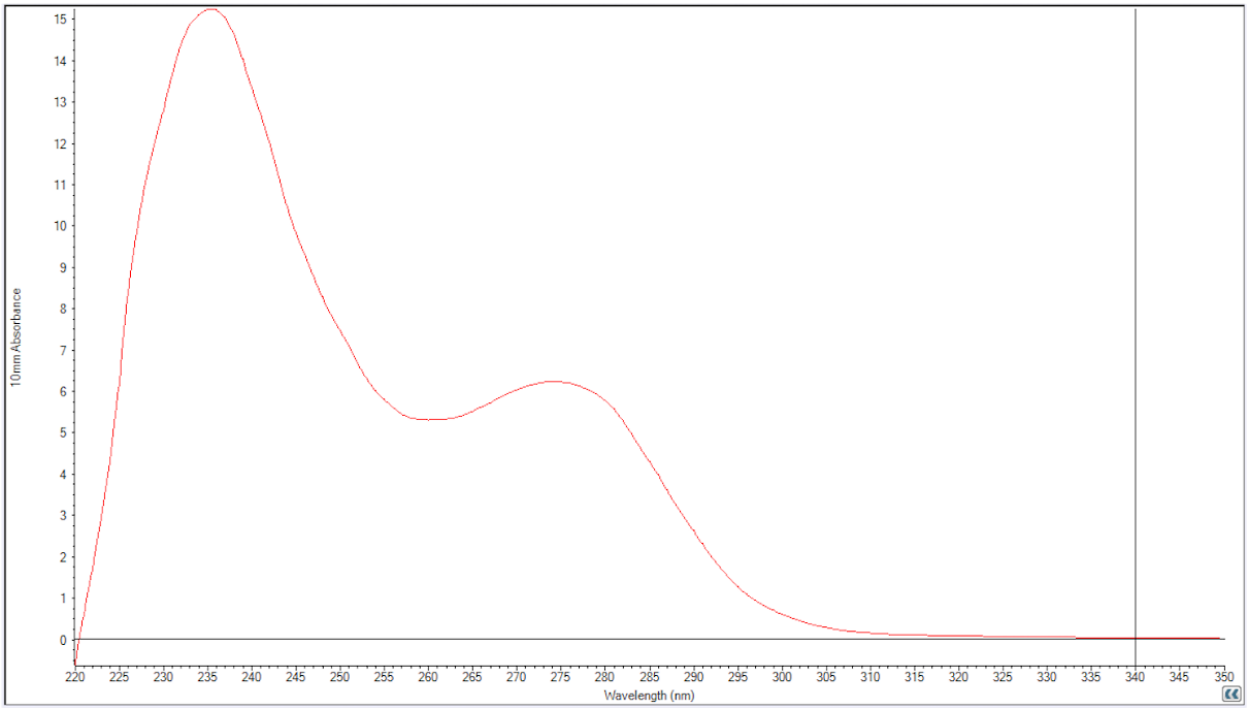


Figure 68. The UV-Vis absorption spectrum of 700 μM Asp-Tat in 10 mM HEPES; 10 mM Acetate buffer; at pH 7.

4.6.2 pH titration of Asp-Tat

NMR experiments were carried out on the highly soluble Asp-Tat protein at different pH conditions. The protocol for expression and purification of ^{15}N isotope-labelled Asp-Tat is described in **Section 3.7**. The lyophilised protein (700 μM) was dissolved at pH 4 in 10 mM acetate and 10 mM HEPES buffer. Sample preparation for NMR and the experimental details are given in

Section 3.9. The ^{1}H-^{15}N HSQC spectra of ^{15}N labelled full-length Asp Tat_{101} protein at pH 4 and 7 are given in **Figures 69** and **70**, respectively. An overlay of these two spectra is shown in **Figure 71**. The spectra show low dispersion of resonances, a characteristic of disordered and denatured proteins. They also show a congregation of cross peaks in different regions: Gly region around 106−110 ppm (black circle); Ser/Thr region between 112 and 117 ppm (green circle). Because of many repetitive residues (**see Appendix II**) and low chemical shift dispersion, severe overlap of signals from Arg and other backbone amides is observed. Unlike His-Tat and SuTat pH 7 HSQC spectra of Asp-Tat result in more cross peaks with good signal-to-noise ratios. Most importantly, all of the glycine residues are observed at pH 7, which emphasizes the superior solubility of the mutant protein due to the substitution of Cys with Asp. Many of the resonances have shifted with the change in pH and this is not surprising. Some have also lost some intensity suggesting a possible conformational change between pH 4 and 7. In the initial discussions, it was proposed that the net-charge of the protein could be a driving factor in the protein aggregation. The net charge of Asp Tat (+ 5.5) is low compared to His-Tat (+12.2) and SuT (+20.1) at pH 7. The Asp Tat sequence proves that the net charge could be one of the driving factors but not the only factor responsible. Also, the emphasis now turns to the cysteine residues, which have been downplayed so far and have not been considered to be playing a role in protein aggregation. To deduce the exact role of Cys in protein aggregation, further research is required.

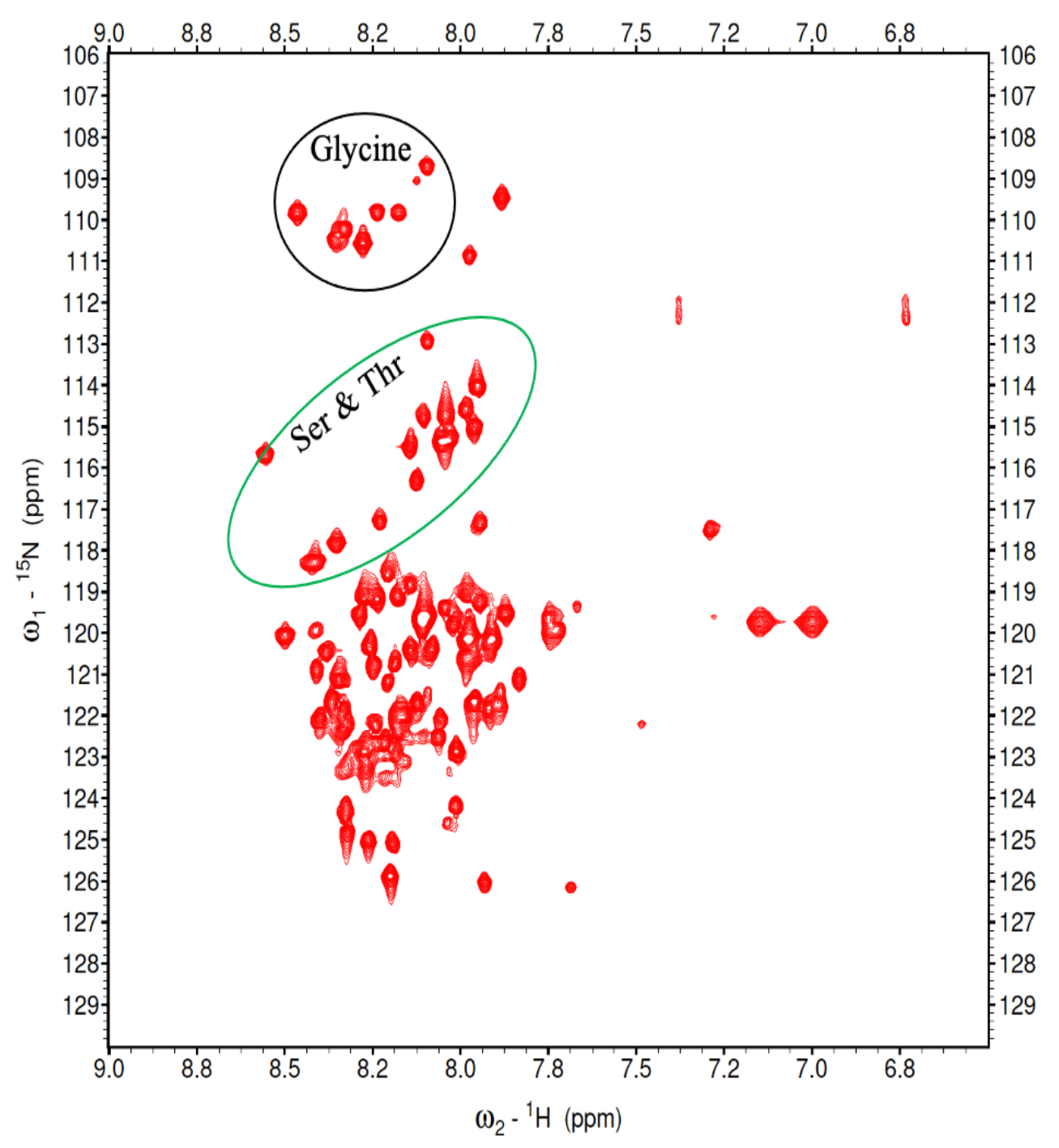


Figure 69. ^{1}H-^{15}N HSQC spectrum of 750 µM ^{15}N-labelled Asp-Tat protein at pH 4 acquired with 90 coadded transients.

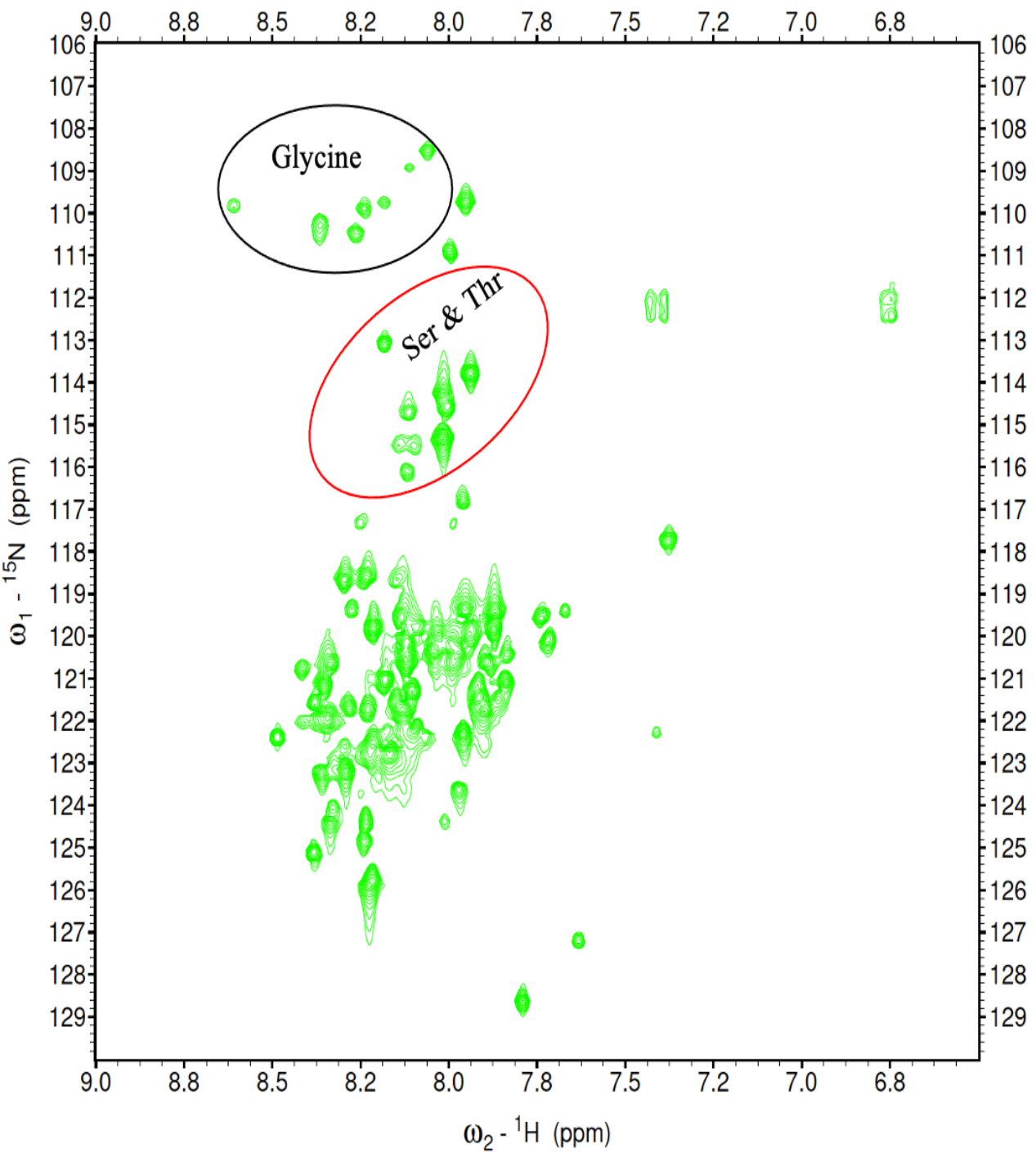


Figure 70. ^{1}H-^{15}N HSQC spectrum of 700 µM ^{15}N-labelled Asp-Tat protein at pH 7 acquired with 90 coadded transients.

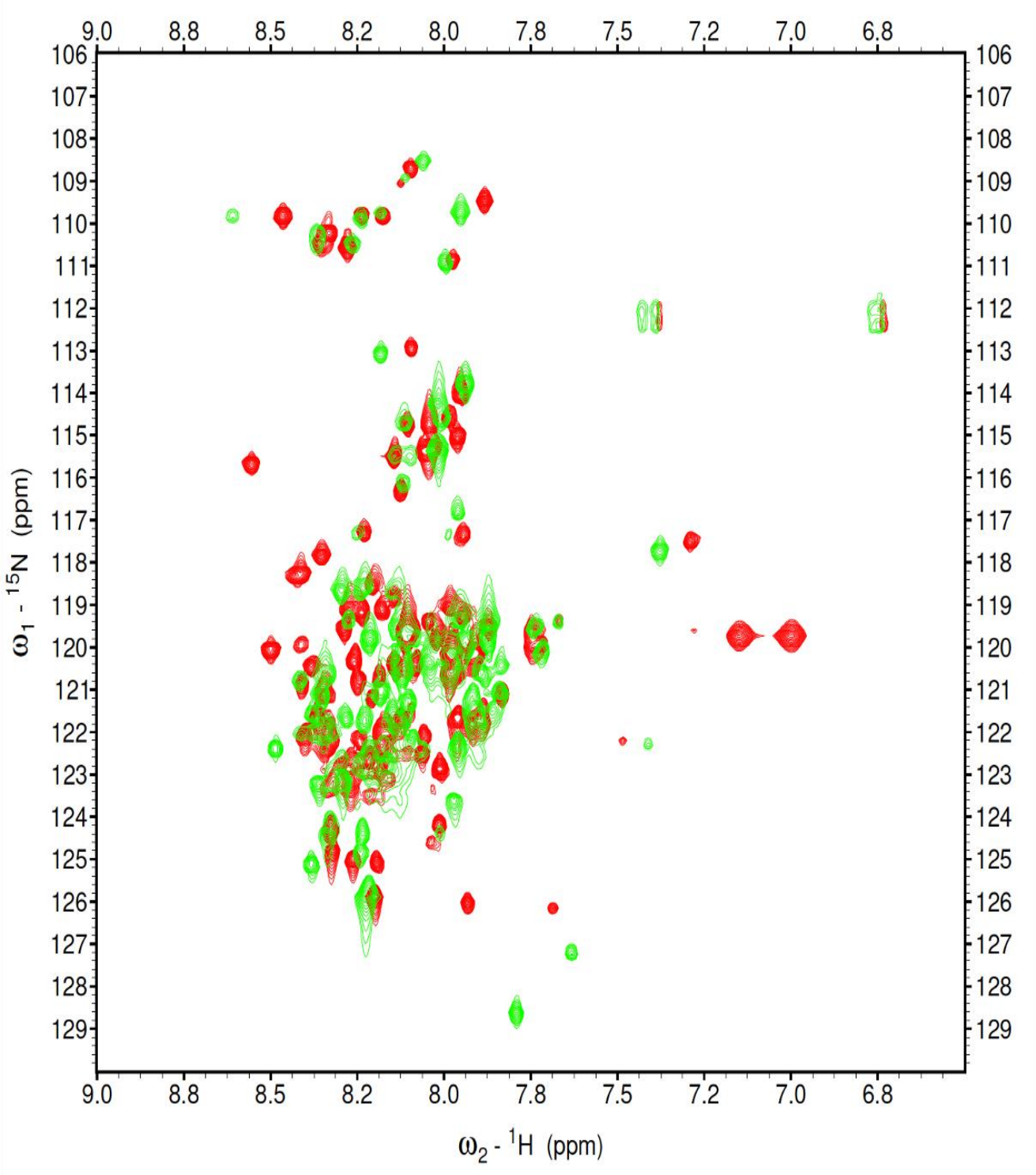


Figure 71. An overlay of the ^{1}H-^{15}N HSQC spectra of ^{15}N-labelled 700 µM Asp-Tat protein acquired at pH 4 and 7 with 90 coadded transients.

4.7 Purity check of CTD of RNAP II

The purity of purified CTD of RNAP II at pH 7.4 was checked through SDS-PAGE analysis (**Figure 72**). A faint band around 95 kDa likely correspons to the MBP-tagged full-length RNAP II CTD protein whereas the band at band at 43 kDa likely arised from truncation or proteolysis of the CTD. The mixture of proteins might be separatable by size exclusion chromatography (SEC) but due to time constraints further purification was not carried out.

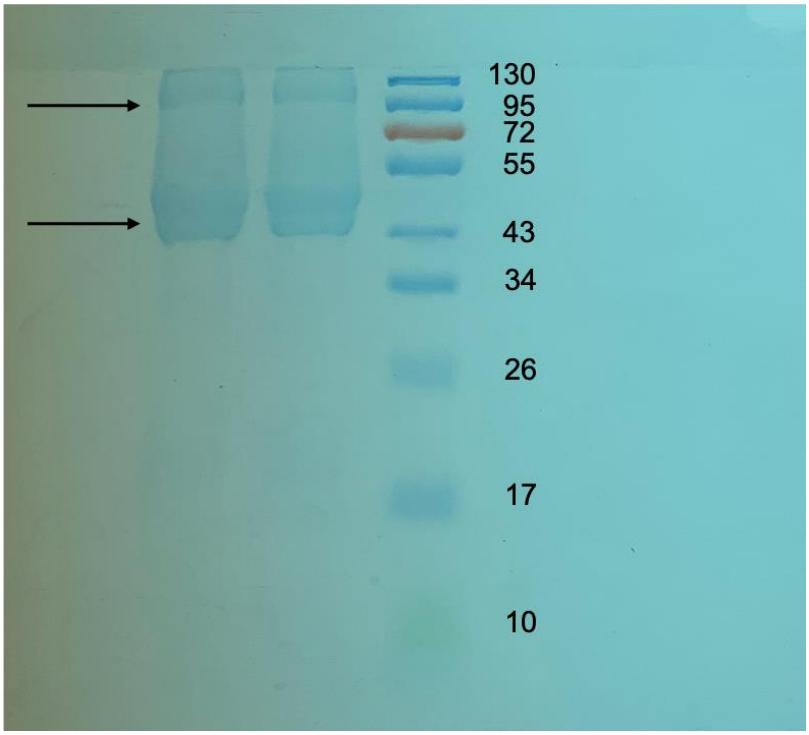


Figure 72. SDS-PAGE of RNAP II CTD 1593-1970 containing a mixture of pure and truncated protein.

5. Conclusions

The intrinisically disordered Tat protein is soluble at pH 4 and the sequence of the full length protein has been successfully analyzed by NMR spectroscopy [137]. However, with an increase in the pH, the solubility of Tat protein decreases gradually and fully precipitates at pH 7. Since, the protein is active at physiological pH, information on its structual attributes at pH 7 would drastically improve the current understanding of the protein's function and exact mechanistic role in HIV replication. Thus, several methodologies and approaches were adopted to improve the solubility of Tat, and these have been discussed in detail in this thesis.

Isotopically-labelled His-tagged full-length Tat protein was successfully expressed and purified by metal-affinity column chromatography. The ^{1}H-^{15}N HSQC experiments acquired at pH values from 4 to 7 showed a gradual deterioration in the number of cross-peaks and the signl-to-noise ratio, possibly due to a decrease in the protein's net-charge leading to the electrostatic interaction-mediated self-aggregation of Tat. Since the Tat protein consists of a single flourescing tryptophan residue at the 31st position, flourescene spectroscopy was employed to deduce the structural modifications occuring as the solution pH was increased. Changes in the emission wavelength and peak intensity can provide a qualitative understanding of the folded or unfolded nature of the protein. As the pH was increased, a quench in peak intensity was observed with no change in the emission wavelength. The measured emission wavelength of 295 nm suggested that the Trp residue is H-bonded but located in a region with lower water content than a fully exposed tryptophan. This suggests that the hydrophobic nature of Trp may result in a locally-structured protein in the vicinity of the Trp and is in agreement with NMR chemical shift measurements [137]. No structural conclusions on the protein folding or unfolding with increasing pH could be made as the emission wavelength did not change. The FTIR analysis of the His-tagged protein

suggested that the fraction of β-sheets declined with increasing pH and at pH 7, the 3_{10}-helix secondary structure was completely lost. Moreover, the fraction of random coil and β-turn secondary structures saw a continuous rise with pH. These results suggested that the protein is less extended at pH 7 than at pH 4.

CPMAS and INEPT solid-state NMR experiments were employed to determine the static and dynamically disorderered regions present in the His-tagged Tat protein. While the CPMAS NMR spectra yielded peaks representing distinct functional groups, no peaks were observed in the INEPT spectra. Moerover, the CPMAS NMR spectra recorded at different temperatures were similar. The absence of such changes in the CPMAS spectra and non-observance of peaks in the INEPT experiment point out that the Tat protein does not have any dynamic regions when prepared in a freeze-dried state.

Charges on the amino acid residues present in the sequence contribute to the protein's solubility [333]. Since Tat protein is positively charged, the concept of using a polyanionic molecule which could interact with the protein electrostatically to improve its solubility was tested. To measure this, the solubility of Tat was tested using a commercially available polyanionic Torula yeast RNA extract. Keep in mind that this RNA is a heterogenous mixture of naturally-occurring RNA molecules spanning a wide range of sizes. The Torula RNA-Tat protein mixture was titred with a NaOH solution and ^{1}H-^{15}N HSQC NMR spectra were collected. The Tat-Torula RNA mixture was optically clear with no protein aggregation over all pH's but only a marginal increase in the number of cross peaks was observed in the NMR spectra. A small improvement in the quality of the NMR spectrum was observed but the spectral information was insufficient to deduce the structure of the protein.

Since the transactivation of transcription activity of Tat is brought about by its interaction with TAR RNA, and the complex is presumably soluble under physiological conditions *in vivo* [334], the effect of TAR RNA on the solubility of Tat protein was studied. TAR RNA is also a polyanionic molecule but, interacts specifically and with high affinity with the positively charged residues of Tat protein. When TAR RNA was added to Tat in solution, aggregation followed by precipitation was observed at all pH values. Although Tat is soluble at pH 4, addition of TAR RNA resulted in protein aggregation, underscoring the requirement of additional proteins to achieve the solubility. I speculate that co-expression of Tat, TAR, Cyclin T1 and CDK9 together might facilitate the solubility of Tat at pH 7. Since the Tat-Torula RNA solution was soluble but, yielded an unassignable NMR spectrum, TAR RNA was added to see if the presence of multiple polyanionic molecules would improve the spectral features. Although all components appeared to be soluble at pH 7, the ^{1}H-^{15}N HSQC spectrum of the Tat, TAR and Torula RNA mixture at pH 6.5 was not improved over that observed with Tat and Torula RNA.

Since the net-charge of the Tat protein decreases with increasing pH, and the protein is soluble at pH 4 and has a high net-charge, a ten-residue arginine tag was attached to the N-terminus of Tat and its solubility was tested. The ^{15}N-labelled supercharged Tat protein was expressed and purified using metal affinity chromatography. Alhough the net charge of supercharged Tat (+20.1) was high compared to His-Tat (+12.2) at pH 7, the protein was not soluble, and the ^{1}H-^{15}N HSQC spectrum was of poor quality as it exhibited only a few cross peaks. The approach was not useful in enchancing the solubility of the protein. However, it may be worth noting that the net charge on His6-Tat at pH 4 is +29. This suggests that adding 9 more Arginines, perhaps at the C-terminus of the protein might overcome the solubility problems at pH 7.

In the final approach, I tested whether the Cys residues were playing a role in the solubility of the protein. As the pH is elevated and thiols deprotonate they can form disulfide bonds that might lead to irreversible aggregation. Although, all experiments were done with the strong reducing agent TCEP present, it can oxidize over time, to test this hypothesis, the cysteine residues in the Tat sequence were replaced with aspartic acid residues and the solubility was studied. The cysteine-less Tat protein was found to be highly soluble at pH 7 and the ^{1}H-^{15}N HSQC spectrum consisted of all cross peaks that were expected for the Tat protein. This experiment suggested that although the net charge on Tat could be one of the driving factors for insolubility, Cys oxidation may also play an important role. Unfortunately, transactivation assays conducted by Dr. Peter Pelka (Department of Microbiology) showed that Asp-Tat is non-functional (data not shown).

The carboxy-terminal domain of the RNA polymerase-II undergoes liquid-liquid phase separation in the presence of a crowding agent at pH 7 [146]. This phase-separated medium, in-theory, could be utilized as a solubilizing medium to dissolve the Tat protein. To test this hypothesis, the carboxy-terminal domain was expressed and purified. The purity was tested using the SDS-page, which relvealed two bands of sizes 84 and 43 kDa, corresponding the full-length CTD and truncated-CTD, respectively. The project is underway.

6. Future directions

It is extremely important to determine the structure of proteins that play a crucial roles in life-threatening viral and bacterial infections. Anti-viral drugs and vaccines for HIV-1 infection are limited and has been a hot-topic of research as the knowledge on the structure of the proteins that are involved in the replication of the HIV virus is minimal. The Tat protein is of specific interest due to its active role in the HIV replication and some crystallography studies have been carried our pertaining to its structure. However, these studies are limited to the protein domain coded from the first exon. Even in crystallographic studies where the protein was co-expressed with its binding partners, only a portion of the Tat sequence was involved. Since the function of a protein is a result of its three-dimensional physiological structure, only solution-state structural studies of the full-length Tat protein will pave the way for the development of new Tat-targeted anti-HIV drugs. The approaches adopted to improve Tat protein's solubility presented in this thesis have been unsuccessful. As a result, the NMR experiments have not borne good signal-to-noise ratio spectra containing resonances from the complete Tat sequence, which could have been used to determine the three-dimensional strucure and dynamics of the Tat protein under physiological conditions. Based on the observations made in this work and the hypotheses built on the new understanding acquired, the following experiments are proposed, that could be successful in improving the solubility of Tat protein and therefore, would facilitate the full-length structure and dynamics of the protein.

1. The CTD-mediated Tat solubility studies involving the liquid-liquid phase separation mechanism should be persued. Improved solubility of the Tat protein in the LLPS medium is anticipated.

2. Tat solubility should be tested in complex coacervates. These are liquid-liquid phase separated solutions obtained using two oppositely charged macro-ions (*e.g*., synthetic polymers, surfactant micelles and charged biomolecules like proteins and nucleic acids) [335,336]. As such media mimic the characteristics of cells, an *in vivo* condition can be created *in vitro*, which might succeed in improving the solubility of the Tat. As the phase-separation mechanism is influenced by conditions such as pH, ionic strength, temperature and ion type [337,338], these parameters would be optimized and their effect on protein's solubility would be studied.

7. References:

[1] G.M. Cooper, The central role of enzymes as biological catalysts, in: The Cell: a molecular approach. 2nd Ed., in: Sunderland (MA): Sinauer Associates, 2000.

[2] P. Forterre, The origin of DNA genomes and DNA replication proteins, Curr. Opin. Microbiol. 5 (2002) 525–532.

[3] P.M. Loriaux, A. Hoffmann, A protein turnover signaling motif controls the stimulus-sensitivity of stress response pathways, PLoS Comput Biol. 9 (2013) e1002932.

[4] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, P. Walter, Carrier proteins and active membrane transport, in: Mol. Biol. Cell. 4th Ed., Garland Science, 2002.

[5] R.U. Lemieux, U. Spohr, How Emil Fischer was led to the lock and key concept for enzyme specificity., Adv. Carbohydr. Chem. Biochem. 50 (1994) 1–20.

[6] E. Fischer, Influence of the configuration on the action of hte enzymes., Reports Ger. Chem. Soc. 27 (1894) 2985–93.

[7] A.E. Mirsky and L. Pauling, On the structure of native,denatured and coagulated proteins, Proc. Natl. Acad. Sci. 22 (1936) 439–447.

[8] L. Pauling, R.B. Corey, H.R. Branson, The structure of proteins- two hydrogen bonded helical configurations of the polypeptide chain, Proc. Natl. Acad. Sci. 37 (1951) 205–211.

[9] T.H. Tahirov, N.D. Babayeva, K. Varzavand, J.J. Cooper, S.C. Sedore, D.H. Price, Crystal structure of HIV-1 Tat complexed with human P-TEFb, Nature. 465 (2010) 747–751.

[10] M.L. Huggins, The structure of alpha keratin, Proc. Natl. Acad. Sci. U.S.A. 43 (1957) 204–209.

[11] L. Mariño-Ramírez, M.G. Kann, B.A. Shoemaker, D. Landsman, Histone structure and nucleosome stability, Expert Rev. Proteomics. 2 (2005) 719–729.

[12] K. Luger, A.W. Mäder, R.K. Richmond, D.F. Sargent, T.J. Richmond, Crystal structure of the nucleosome core particle at 2.8 Å resolution, Nature. 389 (1997) 251–260.

[13] J. Gu, N.D. Babayeva, Y. Suwa, A.G. Baranovskiy, D.H. Price, T.H. Tahirov, Crystal structure of HIV-1 Tat complexed with human P-TEFb and AFF4, Cell Cycle. 13 (2014) 1788–1797.

[14] U. Schulze-Gahmen, I. Echeverria, G. Stjepanovic, Y. Bai, H. Lu, D. Schneidman-Duhovny, J.A. Doudna, Q. Zhou, A. Sali, J.H. Hurley, Insights into HIV-1 proviral transcription from integrative structure and dynamics of the tat:AFF4:P-TEFB:TAR complex, Elife. 5 (2016) 1–21.

[15] J.A. Lukin, G. Kontaxis, V. Simplaceanu, Y. Yuan, A. Bax, C. Ho, Quaternary structure of hemoglobin in solution, Proc. Natl. Acad. Sci. U. S. A. 100 (2003) 517–520.

[16] E. Coluccio, M.Lynne, Chapter–2 Myosin Structure, in: Myosins: a superfamily molecular motors, Springer Science and Business Media. (2007) 35–54, 2007.

[17] H.W. Schroeder Jr, L. Cavacini, Structure and function of immunoglobulins, J. Allergy Clin. Immunol. 125 (2010) S41–S52.

[18] M. Weiss, D.F. Steiner, L.H. Philipson, Insulin biosynthesis , secretion , structure , and structure-activity relationships in: Endotext [internet], South Dartmouth (MA), K.R. Feingold, B.Anawalt, A. Boyce et.al., Ed., (2020) 1–51.

[19] J.C. Kendrew, G. Bodo, H.M. Dintzis, R.G. Parrish, H. Wyckoff, D.C. Phillips, A three-dimensional model of the myoglobin molecule obtained by x-ray analysis, Nature. 181 (1958) 662–666.

[20] M. Perutz, M. Rossmann, A.F. Cullis, H. Muirhead, G. Will, A.C.T. North, Structure of hoemoglobin: a three-dimensional fourier synthesis at 5.5 Å resolution, obtained by x-ray

analysis, Nature. 185 (1960) 416–422.

[21] A.R. Fersht, From the protein structures to our current knowledge of protein folding : delights and scepticisms, Nat. Rev. Mol. Cell Biol. 9 (2008) 650–654.

[22] W.T. Godbey, Chapter–2 Proteins,An introduction to biotechnology: The science, technology and medical applications, Elsevier Ltd., 2014.

[23] J. Feher, Chapter–2.3,Protein Structure, Quant. Hum. Physiol. 2nd Ed., Acad. Press. (2017) 130–141.

[24] L.R. Engelking, Chapter 4 – Protein Structure, Textb. Vet. Physiol. Chem. 3rd Ed., Elsevier Ltd, (2015) 18–25.

[25] M. Gromiha, Chapter–1 Proteins, Protein Bioinforma. From Seq. to Funct. Acad. Press. Chapter 1 (2010) 1–27.

[26] W.W. Bromer, M.E. Boucher, J.M. Patterson, A.H. Pekar, B.H. Frank, Glucagon structure and function I. purification and properties of bovine glucagon and monodesamidoglucago, J. Biol. Chem. 247 (1972) 2581–2585.

[27] V. du Vigneaud, C. Ressler, S. Trippett, The sequence of amino acids in oxytocin, with a proposal for the structure of oxytocin, J. Biol. Chem. 205 (1953) 949–957.

[28] P. Brazeau, W. Vale, R. Burgus, N. Ling, M. Butcher, J. Rivier, R. Guillemin, Hypothalamic polypeptide that inhibits the secretion of immunoreactive pituitary growth hormone, Science (80-.). 179 (1973) 77–79.

[29] G.N. Ramachandran, C. Ramakrishnan, V. Sasisekharan, Stereochemistry of polypeptide chain configurations, J. Mol. Biol. 7 (1963) 95–99.

[30] G.N. Ramachandran, V. Sasisekharan, Conformation of polypeptides and proteins, in: Adv. Protein Chem., Elsevier, 1968: pp. 283–437.

[31] L.N. David, M.C. Michael, others, Chapter– 4 The three-dimensional structure of proteins, in: Lehninger principles of biochemistry, 4th Ed., 2006.

[32] R.P. Riek, R.M. Graham, The elusive π-helix, J. Struct. Biol. 173 (2011) 153–160.

[33] R.S. Vieira-Pires, J.H. Morais-Cabral, 3_{10} helices in channels and other membrane proteins, J. Gen. Physiol. 136 (2010) 585–592.

[34] N. Bhattacharjee, P. Biswas, Position-specific propensities of amino acids in the beta strand, BMC Struct. Biol. 10 (2010).

[35] C.M. Venkatachalam, Stereochemical criteria for polypeptides and proteins. V. Conformation of a system of three linked peptide units, Biopolymers. 6 (1968) 1425–1436.

[36] A.A. Adzhubei, M.J.E. Sternberg, A.A. Makarov, Polyproline-II helix in proteins: structure and function, J. Mol. Biol. 425 (2013) 2100–2132.

[37] A.A. Adzhubei, M.J.E. Sternberg, Left-handed polyproline II helices commonly occur in globular proteins, J. Mol. Biol. 229 (1993) 472–493.

[38] Rehman, M.Farooq, S. Botelho, Biochemistry, secondary protein structure, in: Treasure Island (FL), in: StatPearls [Internet], StatPearls Publishing, 2020.

[39] B. Jirgensons, Optical rotation and viscosity of native and denatured proteins. II. Influence of temperature and concentration, Arch. Biochem. Biophys. 41 (1952) 333–344.

[40] K. Grizzuti, G.E. Perlmann, Conformation of the phosphoprotein, phosvitin., J. Biol. Chem. 245 (1970) 2573–2578.

[41] H.J. Vogel, Structure of Hen Phosvitin: A ^{31}P NMR, ^{1}H NMR, and laser photochemically induced dynamic nuclear polarization ^{1}H NMR Study, Biochemistry. 22 (1983) 668–674.

[42] B. Jirgensons, Classification of Proteins According to Conformation, Die Makromol.

Chemie Macromol. Chem. Phys. 91 (1966) 74–86.

[43] F. Meng, V.N. Uversky, L. Kurgan, Comprehensive review of methods for prediction of intrinsic disorder and its molecular functions, Cell. Mol. Life Sci. 74 (2017) 3069–3090.

[44] P.E. Wright, H.J. Dyson, Intrinsically disordered proteins in cellular signalling and regulation, Nat. Rev. Mol. Cell Biol. 16 (2015) 18–29.

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

[46] R.A. Pullen, J.A. Jenkins, I.J. Tickle, S.P. Wood, T.L. Blundell, The relation of polypeptide hormone structure and flexibility to receptor binding: The relevance of X-ray studies on insulins, glucagon and human placental lactogen, Mol. Cell. Biochem. 8 (1975) 5–20.

[47] P.D. Cary, T. Moss, E.M. Bradbury, High-Resolution Proton-Magnetic-Resonance Studies of Chromatin Core Particles, Eur. J. Biochem. 89 (1978) 475–482.

[48] P.E. Wright, H.J. Dyson, Intrinsically unstructured proteins: Re-assessing the protein structure-function paradigm, J. Mol. Biol. 293 (1999) 321–331.

[49] A.K. Dunker, J.D. Lawson, C.J. Brown, R.M. Williams, P. Romero, J.S. Oh, C.J. Oldfield, A.M. Campen, C.M. Ratliff, K.W. Hipps, J. Ausio, M.S. Nissen, R. Reeves, C.H. Kang, C.R. Kissinger, R.W. Bailey, M.D. Griswold, W. Chiu, E.C. Garner, Z. Obradovic, Intrinsically disordered protein, J. Mol. Graph. Model. 19 (2001) 26–59.

[50] O. Schweers, E. Schönbrunn-Hanebeck, A. Marx, E. Mandelkow, Structural studies of tau protein and Alzheimer paired helical filaments show no evidence for β-structure, J. Biol. Chem. 269 (1994) 24290–24297.

[51] P.H. Weinreb, W. Zhen, A.W. Poon, K.A. Conway, P.T. Lansbury, NACP, a protein

implicated in Alzheimer's disease and learning, is natively unfolded, Biochemistry. 35 (1996) 13709–13715.

[52] J. Chen, H. Liang, A. Fernández, Protein structure protection commits gene expression patterns, Genome Biol. 9 (2008) 1–11.

[53] V.N. Uversky, Dancing protein clouds: the strange biology and chaotic physics of intrinsically disordered proteins, J. Biol. Chem. 291 (2016) 6681–6688.

[54] A.K. Dunker, V.N. Uversky, Drugs for "protein clouds": Targeting intrinsically disordered transcription factors, Curr. Opin. Pharmacol. 10 (2010) 782–788.

[55] H.J. Dyson, Making Sense of Intrinsically Disordered Proteins, Biophys. J. 110 (2016) 1013–1016.

[56] R.W. Kriwacki, L. Hengst, L. Tennant, S.I. Reed, P.E. Wright, Structural studies of p21Waf1/Cip1/Sdi1 in the free and Cdk2-bound state: conformational disorder mediates binding diversity, Proc. Natl. Acad. Sci. 93 (1996) 11504–11509.

[57] G.W. Daughdrill, M.S. Chadsey, J.E. Karlinsey, K.T. Hughes, F.W. Dahlquist, The C-terminal half of the anti-sigma factor, FlgM, becomes structured when bound to its target, σ28, Nat. Struct. Biol. 4 (1997) 285–291.

[58] K. Sugase, H.J. Dyson, P.E. Wright, Mechanism of coupled folding and binding of an intrinsically disordered protein, Nature. 447 (2007) 1021–1025.

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

[60] R. Mohana-Borges, N.K. Goto, G.J.A. Kroon, H.J. Dyson, P.E. Wright, Structural characterization of unfolded states of apomyoglobin using residual dipolar couplings, J.

Mol. Biol. 340 (2004) 1131–1142.

[61] V.N. Uversky, Intrinsically disordered proteins and their “Mysterious” (meta)physics, Front. Phys. 7 (2019) 8–23.

[62] V.N. Uversky, Unusual biophysics of intrinsically disordered proteins, Biochim. Biophys. Acta - Proteins Proteomics. 1834 (2013) 932–951.

[63] V.N. Uversky, J.R. Gillespie, A.L. Fink, Why are “natively unfolded” proteins unstructured under physiologic conditions?, Proteins Struct. Funct. Genet. 41 (2000) 415–427.

[64] V.N. Uversky, Intrinsically disordered proteins from A to Z, Int. J. Biochem. Cell Biol. 43 (2011) 1090–1103.

[65] V.N. Uversky, A.K. Dunker, Understanding protein non-folding, Biochemica et Biophysica Acta (BBA)-proteins and proteomics. 1804 (2010) 1231-1264.

[66] A. Campen, R. Williams, C. Brown, J. Meng, V. Uversky, A. Dunker, TOP-IDP-scale: a new amino acid scale measuring propensity for intrinsic disorder, Protein Pept. Lett. 15 (2008) 956–963.

[67] Y.H. Sung, J. Shin, J. Shin, W. Lee, Solution structure of p21Waf1/Cip1/Sdi1 C-terminal domain bound to Cdk4, J. Biomol. Struct. Dyn. 19 (2001) 419–427.

[68] A.M. Abukhdeir, B.H. Park, P21 and p27: roles in carcinogenesis and drug resistance, Expert Rev. Mol. Med. 10 (2008) e19.

[69] A.C. Joerger, A.R. Fersht, The tumor suppressor p53: from structures to drug discovery, Cold Spring Harb. Perspect. Biol. 2 (2010) a000919.

[70] S.L. Clark, A.M. Rodriguez, R.R. Snyder, G.D. V Hankins, D. Boehning, Structure-function of the tumor suppressor BRCA1, Comput. Struct. Biotechnol. J. 1 (2012)

e201204005.

[71] P. Yasar, G. Ayaz, S.D. User, G. Güpür, M. Muyan, Molecular mechanism of estrogen-estrogen receptor signaling, Reprod. Med. Biol. 16 (2017) 4–20.

[72] N. Sugitani, R.M. Sivley, K.E. Perry, J.A. Capra, W.J. Chazin, XPA: A key scaffold for human nucleotide excision repair, DNA Repair (Amst). 44 (2016) 123–135.

[73] F.N. Emamzadeh, Alpha-synuclein structure, functions, and interactions, J. Res. Med. Sci. 21 (2016) 21–29.

[74] P.M. Mishra, N.C. Verma, C. Rao, V.N. Uversky, C.K. Nandi, Intrinsically disordered proteins of viruses: Involvement in the mechanism of cell regulation and pathogenesis, Prog. Mol. Biol. Transl. Sci. 174 (2020) 1–78.

[75] A.B. Sigalov, A. V. Zhuravleva, V.Y. Orekhov, Binding of intrinsically disordered proteins is not necessarily accompanied by a structural transition to a folded form, Biochimie. 89 (2007) 419–421.

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

[77] Z. Liu, Y. Huang, Advantages of proteins being disordered, Protein Sci. 23 (2014) 539–550.

[78] A.B. Sigalov, W.M. Kim, M. Saline, L.J. Stern, The intrinsically disordered cytoplasmic domain of the T cell receptor ζ chain binds to the Nef protein of simian immunodeficiency virus without a disorder-to-order transition, Biochemistry. 47 (2008) 12942–12944.

[79] B. He, K. Wang, Y. Liu, B. Xue, V.N. Uversky, A.K. Dunker, Predicting intrinsic disorder in proteins: an overview, Cell Res. 19 (2009) 929–949.

[80] F. Meng, V.N. Uversky, L. Kurgan, Comprehensive review of methods for prediction of

intrinsic disorder and its molecular functions, Cell. Mol. Life Sci. 74 (2017) 3069–3090.

[81] P. Romero, Z. Obradovic, C. Kissinger, J.E. Villafranca, A.K. Dunker, Identifying disordered regions in proteins from amino acid sequence, IEEE Int. Conf. Neural Networks - Conf. Proc. 1 (1997) 90–95.

[82] J. Prilusky, C.E. Felder, T. Zeev-Ben-Mordehai, E.H. Rydberg, O. Man, J.S. Beckmann, I. Silman, J.L. Sussman, FoldIndex©: A simple tool to predict whether a given protein sequence is intrinsically unfolded, Bioinformatics. 21 (2005) 3435–3438.

[83] R. Linding, R.B. Russell, V. Neduva, T.J. Gibson, GlobPlot: Exploring protein sequences for globularity and disorder, Nucleic Acids Res. 31 (2003) 3701–3708.

[84] Z. Dosztányi, V. Csizmok, P. Tompa, I. Simon, IUPred: Web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content, Bioinformatics. 21 (2005) 3433–3434.

[85] P. Romero, Z. Obradovic, X. Li, E.C. Garner, C.J. Brown, A.K. Dunker, Sequence complexity of disordered protein, Proteins Struct. Funct. Genet. 42 (2001) 38–48.

[86] R. Linding, L.J. Jensen, F. Diella, P. Bork, T.J. Gibson, R.B. Russell, Protein disorder prediction: Implications for structural proteomics, Structure. 11 (2003) 1453–1459.

[87] D.T. Jones, J.J. Ward, Prediction of disordered regions in proteins from position specific score matrices, Proteins Struct. Funct. Genet. 53 (2003) 573–578.

[88] K. Peng, S. Vucetic, P. Radivojac, C.J. Brown, K. Dunker, Z. Obradovic, Optimizing long intrinsic disorder predictors with protein evolutionary information, J. Bioinform. Comput. Biol. 3 (2005) 35–60.

[89] J.J. Ward, J.S. Sodhi, L.J. McGuffin, B.F. Buxton, D.T. Jones, Prediction and functional analysis of native disorder in proteins from the three kingdoms of life, J. Mol. Biol. 337

(2004) 635–645.

[90] A. Schlessinger, G. Yachdav, B. Rost, PROFbval: Predict flexible and rigid residues in proteins, Bioinformatics. 22 (2006) 891–893.

[91] J. Cheng, M.J. Sweredoski, P. Baldi, Accurate prediction of protein disordered regions by mining protein structure data, Data Min. Knowl. Discov. 11 (2005) 213–222.

[92] J. Liu, B. Rost, NORSp: Predictions of long regions without regular secondary structure, Nucleic Acids Res. 31 (2003) 3833–3835.

[93] M.J. Mizianty, W. Stach, K. Chen, K.D. Kedarisetti, F.M. Disfani, L. Kurgan, Improved sequence-based prediction of disordered regions with multilayer fusion of multiple information sources, Bioinformatics. 27 (2011) i489–i496.

[94] L.P. Kozlowski, J.M. Bujnicki, MetaDisorder: a meta-server for the prediction of intrinsic disorder in proteins, BMC Bioinformatics. 13 (2012) 1–11.

[95] M.J. Mizianty, Z. Peng, L. Kurgan, MFDp2: accurate predictor of disorder in proteins by fusion of disorder probabilities, content and profiles, Intrinsically Disord. Proteins. 1 (2013) e24428.

[96] Y. Chen, Structural genomics: general applications, in: Methods in molecular biology, Springer Protoc. Humana Press New York City. 1091 (2014) 179–186.

[97] D. Boehr, R. Nussinov, P.E. Wright, The role of dynamic conformational ensembles in biomolecular recognition, Nat Chem Biol. 5 (2009) 789–796.

[98] M. Arai, Unified understanding of folding and binding mechanisms of globular and intrinsically disordered proteins, Biophys. Rev. 10 (2018) 163–181.

[99] J.M. Rogers, V. Oleinikovas, S.L. Shammas, C.T. Wong, D. De Sancho, C.M. Baker, J. Clarke, Interplay between partner and ligand facilitates the folding and binding of an

intrinsically disordered protein, Proc. Natl. Acad. Sci. U. S. A. 111 (2014) 15420–15425.

[100] M. Arai, K. Sugase, H.J. Dyson, P.E. Wright, Conformational propensities of intrinsically disordered proteins influence the mechanism of binding and folding, Proc. Natl. Acad. Sci. 112 (2015) 9614–9619.

[101] P. Tompa, M. Fuxreiter, Fuzzy complexes: polymorphism and structural disorder in protein-protein interactions, Trends Biochem. Sci. 33 (2008) 2–8.

[102] M. Fuxreiter, Fuzziness: Linking regulation to protein dynamics, Mol. Biosyst. 8 (2012) 168–177.

[103] S. Rauscher, R. Pomès, Structural disorder and protein elasticity, Adv. Exp. Med. Biol. 725 (2012) 159–183.

[104] Q. Shen, J. Shi, D. Zeng, B. Zhao, P. Li, W. Hwang, J.H. Cho, Molecular Mechanisms of Tight Binding through Fuzzy Interactions, Biophys. J. 114 (2018) 1313–1320.

[105] R. Sharma, Z. Raduly, M. Miskei, M. Fuxreiter, Fuzzy complexes: Specific binding without complete folding, FEBS Lett. 589 (2015) 2533–2542.

[106] M. Arbesú, G. Iruela, H. Fuentes, J.M.C. Teixeira, M. Pons, Intramolecular Fuzzy Interactions Involving Intrinsically Disordered Domains, Front. Mol. Biosci. 5 (2018) 1–7.

[107] P. Tompa, P. Csermely, The role of structural disorder in the function of RNA and protein chaperones, FASEB J. 18 (2004) 1169–1175.

[108] R. Van Der Lee, M. Buljan, B. Lang, R.J. Weatheritt, G.W. Daughdrill, A.K. Dunker, M. Fuxreiter, J. Gough, J. Gsponer, D.T. Jones, P.M. Kim, R.W. Kriwacki, C.J. Oldfield, R. V. Pappu, P. Tompa, V.N. Uversky, P.E. Wright, M.M. Babu, Classification of intrinsically disordered regions and proteins, Chem. Rev. 114 (2014) 6589–6631.

[109] M.P. Sang, Y.J. Han, T.D. Kim, H.P. Jeon, C.H. Yang, J. Kim, Distinct roles of the N-

terminal-binding domain and the C-terminal-solubilizing domain of α-synuclein, a molecular chaperone, J. Biol. Chem. 277 (2002) 28512–28520.

[110] J. Bhattacharyya, K.P. Das, Molecular chaperone-like properties of an unfolded protein, α(s)- casein, J. Biol. Chem. 274 (1999) 15505–15509.

[111] P.A. Chong, J.D. Forman-Kay, Liquid–liquid phase separation in cellular signaling systems, Curr. Opin. Struct. Biol. 41 (2016) 180–186.

[112] D.M. Mitrea, R.W. Kriwacki, Phase separation in biology; functional organization of a higher order, Cell Commun. Signal. 14 (2016) 1–20

[113] F. Barré-Sinoussi, A.L. Ross, J.F. Delfraissy, Past, present and future: 30 years of HIV research, Nat. Rev. Microbiol. 11 (2013) 877–883.

[114] CDC, Current Trends update on Acquired Immune Deficiency Syndrome (AIDS), Morb. Mortal. Wkly. Rep. 31 (1982) 507–508.

[115] R.C. Gallo, HIV—the cause of AIDS: An overview on its biology, mechanisms of disease induction, and our attempts to control it, JAIDS J. Acquir. Immune Defic. Syndr. 1 (1988) 521–535.

[116] S. Schmid, The discovery of HIV-1, Nat. Commun. 1981 (2018) 1. www.nature.com/collections/hiv-milestone.

[117] H.I. V Global, AIDS statistics, 2020 fact sheet United Nations Political Declaration on HIV and AIDS, Geneva: UNAIDS. (2020).

[118] P.A. Luciw, Human immunodeficiency viruses and their replication, Fields Virol. 2 (1996) 1881–1952.

[119] R. Seitz, Human Immunodeficiency Virus (HIV), Transfus. Med. Hemotherapy. 43 (2016) 203–222.

[120] E. Fanales-Belasio, M. Raimondo, B. Suligoi, S. Buttò, HIV virology and pathogenetic mechanisms of infection: a brief overview, Ann. Ist. Super. Sanita. 46 (2010) 5–14.

[121] V. To, Backbone Dynamics of the Intrinsically Disordered HIV-1 Tat Protein, (2017).

[122] A. Engelman, P. Cherepanov, The structural biology of HIV-1: mechanistic and therapeutic insights, Nat. Rev. Microbiol. 10 (2012) 279–290.

[123] E.O. Freed, HIV-1 assembly, release and maturation, Nat. Rev. Microbiol. 13 (2015) 484–496.

[124] M.D. F.J. Palella, Jr., M.D. Kathleen, Declining morbidity and mortality among patients with advanced human immunodeficiency virus infection, N. Engl. J. Med. 338 (1998) 853–860.

[125] F. Esposito, A. Corona, E. Tramontano, HIV-1 reverse transcriptase still remains a new drug target: structure, function, classical inhibitors, and new inhibitors with innovative mechanisms of actions, Mol. Biol. Int. 2012 (2012).

[126] X. Wei, J.M. Decker, H. Liu, Z. Zhang, R.B. Arani, J.M. Kilby, M.S. Saag, X. Wu, G.M. Shaw, J.C. Kappes, Emergence of resistant human immunodeficiency virus type 1 in patients receiving fusion inhibitor (T-20) monotherapy, Antimicrob. Agents Chemother. 46 (2002) 1896–1905.

[127] Deutsche AIDS-Gesellschaft (DAIG): German -Austrian guidelines for antiretroviral thereapy of HIV-infection.

[128] S. Shojania, J.D. O'Neil, HIV-1 Tat is a natively unfolded protein: The solution conformation and dynamics of reduced HIV-1 Tat-(1-72) by NMR spectroscopy, J. Biol. Chem. 281 (2006) 8347–8356.

[129] M.K. Johri, R. Mishra, C. Chhatbar, S.K. Unni, S.K. Singh, Tits and bits of HIV Tat

protein, Expert Opin. Biol. Ther. 11 (2011) 269–283.

[130] C. Spector, A.R. Mele, B. Wigdahl, M.R. Nonnemacher, Genetic variation and function of the HIV-1 Tat protein, Springer Berlin Heidelberg, 2019.

[131] J.A. Garcia, D. Harrich, L. Pearson, R. Mitsuyasu, R.B. Gaynor, Functional domains required for tat-induced transcriptional activation of the HIV-1 long terminal repeat., EMBO J. 7 (1988) 3143–3147.

[132] G. Marzio, M. Tyagi, M.I. Gutierrez, M. Giacca, HIV-1 tat transactivator recruits p300 and CREB-binding protein histone acetyltransferases to the viral promoter, Proc. Natl. Acad. Sci. 95 (1998) 13519–13524.

[133] J.M. Kim, H.S. Choi, B.L. Seong, The folding competence of HIV-1 Tat mediated by interaction with TAR RNA, RNA Biol. 14 (2017) 926–937.

[134] E. Clark, B. Nava, M. Caputi, Tat is a multifunctional viral protein that modulates cellular gene expression and functions, Oncotarget. 8 (2017) 27569–27581.

[135] U. Mahlknecht, I. Dichamp, A. Varin, C. Van Lint, G. Herbein, NF-$κ$B-dependent control of HIV-1 transcription by the second coding exon of Tat in T cells, J. Leukoc. Biol. 83 (2008) 718–727.

[136] K.L. Shaw, G.R. Grimsley, G.I. Yakovlev, A.A. Makarov, C.N. Pace, The effect of net charge on the solubility, activity, and stability of ribonuclease Sa, Protein Sci. 10 (2001) 1206–1215.

[137] V. To, E. Dzananovic, S.A. McKenna, J. O'Neil, The dynamic landscape of the full-length HIV-1 transactivator of transcription, Biochemistry. 55 (2016) 1314–1325.

[138] Prot pi | Protein Tool, (n.d.). https://www.protpi.ch/Calculator/ProteinTool.

[139] A.T. Das, A. Harwig, B. Berkhout, The HIV-1 Tat Protein Has a Versatile Role in

Activating Viral Transcription, J. Virol. 85 (2011) 9506–9516.

[140] K. Jeang, HIV-1 Tat : Structure and Function, Hum. Retroviruses AIDS 1996 A Compil. Anal. Nucleic Acid Amin. Acid Seq. (1996) 11–26.

[141] H. Zhang, X. Ji, P. Li, C. Liu, J. Lou, Z. Wang, W. Wen, Y. Xiao, M. Zhang, X. Zhu, Liquid-liquid phase separation in biology: mechanisms, physiological functions and human diseases, Sci. China Life Sci. 63 (2020) 953–985.

[142] P.A. Chong, J.D. Forman-Kay, Liquid--liquid phase separation in cellular signaling systems, Curr. Opin. Struct. Biol. 41 (2016) 180–186.

[143] S.F. Banani, H.O. Lee, A.A. Hyman, M.K. Rosen, Biomolecular condensates: organizers of cellular biochemistry, Nat. Rev. Mol. Cell Biol. 18 (2017) 285–298.

[144] X. Su, J.A. Ditlev, E. Hui, W. Xing, S. Banjade, J. Okrut, D.S. King, J. Taunton, M.K. Rosen, R.D. Vale, Phase separation of signaling molecules promotes T cell receptor signal transduction, Science (80-.). 352 (2016) 595–599.

[145] G. Zhang, Z. Wang, Z. Du, H. Zhang, mTOR regulates phase separation of PGL granules to modulate their autophagic degradation, Cell. 174 (2018) 1492–1506.

[146] M. Boehning, C. Dugast-Darzacq, M. Rankovic, A.S. Hansen, T. Yu, H. Marie-Nelly, D.T. McSwiggen, G. Kokic, G.M. Dailey, P. Cramer, others, RNA polymerase II clustering through carboxy-terminal domain phase separation, Nat. Struct. Mol. Biol. 25 (2018) 833–840.

[147] M. Kato, T.W. Han, S. Xie, K. Shi, X. Du, L.C. Wu, H. Mirzaei, E.J. Goldsmith, J. Longgood, J. Pei, others, Cell-free formation of RNA granules: low complexity sequence domains form dynamic fibers within hydrogels, Cell. 149 (2012) 753–767.

[148] T.W. Han, M. Kato, S. Xie, L.C. Wu, H. Mirzaei, J. Pei, M. Chen, Y. Xie, J. Allen, G.

Xiao, others, Cell-free formation of RNA granules: bound RNAs identify features and components of cellular assemblies, Cell. 149 (2012) 768–779.

[149] M.A.M. Reijns, R.D. Alexander, M.P. Spiller, J.D. Beggs, A role for Q/N-rich aggregation-prone regions in P-body localization, J. Cell Sci. 121 (2008) 2463–2472.

[150] T.J. Nott, E. Petsalaki, P. Farber, D. Jervis, E. Fussner, A. Plochowietz, T.D. Craggs, D.P. Bazett-Jones, T. Pawson, J.D. Forman-Kay, others, Phase transition of a disordered nuage protein generates environmentally responsive membraneless organelles, Mol. Cell. 57 (2015) 936–947.

[151] S.L. Crick, K.M. Ruff, K. Garai, C. Frieden, R. V Pappu, Unmasking the roles of N-and C-terminal flanking sequences from exon 1 of huntingtin as modulators of polyglutamine aggregation, Proc. Natl. Acad. Sci. 110 (2013) 20075–20080.

[152] S.B. Weiss, L. Gladstone, A Mammalian System for The Incorporation of Cytidine Triphosphate into Ribonucleic Acid, J. Am. Chem. Soc. 81 (1959) 4118–4119.

[153] S.M. Tan-Wong, J.B. Zaugg, J. Camblong, Z. Xu, D.W. Zhang, H.E. Mischo, A.Z. Ansari, N.M. Luscombe, L.M. Steinmetz, N.J. Proudfoot, Gene loops enhance transcriptional directionality, Science (80-.). 338 (2012) 671–675.

[154] A. Sentenac, M. Sawadogo, RNA Polymerase B (II) and General Transcription Factors, Annu. Rev Biochem. 59 (1990) 711–754.

[155] R. A.Young, RNA polymerase II, Annu. Rev Biochem. 60 (1991) 689–715.

[156] L.A. Allison, M. Moyle, M. Shales, C.J. Ingles, Extensive homology among the largest subunits of eukaryotic and prokaryotic RNA polymerases, Cell. 42 (1985) 599–610.

[157] J.L. Corden, D.L. Cadena, J.M. Ahearn, M.E. Dahmus, A unique structure at the carboxyl terminus of the largest subunit of eukaryotic RNA polymerase II, Proc. Natl. Acad. Sci. 82

(1985) 7934–7938.

[158] A. Meinhart, T. Kamenski, S. Hoeppner, S. Baumli, P. Cramer, A structural perspective of CTD function, Genes Dev. 19 (2005) 1401–1415.

[159] D. Eick, M. Geyer, The RNA polymerase II carboxy-terminal domain (CTD) code, Chem. Rev. 113 (2013) 8456–8490.

[160] J.M. Ahearn, M.S. Bartolomei, M.L. West, L.J. Cisek, J.L. Corden, Cloning and sequence analysis of the mouse genomic locus encoding the largest subunit of RNA polymerase II., J. Biol. Chem. 262 (1987) 10695–10705.

[161] Y.K. Kim, C.F. Bourgeois, C. Isel, M.J. Churcher, J. Karn, Phosphorylation of the RNA polymerase II carboxyl-terminal domain by CDK9 is directly responsible for human immunodeficiency virus type 1 Tat-activated transcriptional elongation, Mol. Cell. Biol. 22 (2002) 4622–4637.

[162] B. Palancade, O. Bensaude, Investigating RNA polymerase II carboxyl-terminal domain (CTD) phosphorylation, Eur. J. Biochem. 270 (2003) 3859–3870.

[163] T. Riedl, J.-M. Egly, Phosphorylation in transcription: the CTD and more, Gene Expr. J. Liver Res. 9 (2001) 3–13.

[164] Y. Hirose, J.L. Manley, RNA polymerase II and the integration of nuclear events, Genes Dev. 14 (2000) 1415–1429.

[165] G. Wang, G.T. Cantin, J.L. Stevens, A.J. Berk, Characterization of mediator complexes from HeLa cell nuclear extract, Mol. Cell. Biol. 21 (2001) 4604–4613.

[166] H. Lu, D. Yu, A.S. Hansen, S. Ganguly, R. Liu, A. Heckert, X. Darzacq, Q. Zhou, Phase-separation mechanism for C-terminal hyperphosphorylation of RNA polymerase II, Nature. 558 (2018) 318–323.

[167] I. Kwon, M. Kato, S. Xiang, L. Wu, P. Theodoropoulos, H. Mirzaei, T. Han, S. Xie, J.L. Corden, S.L. McKnight, Phosphorylation-regulated binding of RNA polymerase II to fibrous polymers of low-complexity domains, Cell. 155 (2013) 1049–1060.

[168] C. Ma, A. Malessa, A.J. Boersma, K. Liu, A. Herrmann, Supercharged Proteins and Polypeptides, Adv. Mater. 32, 1905309, (2020) 1–21.

[169] A. Kolbe, Fabrication and Applications of Supercharged, Unfolded Proteins, 2012.

[170] D.B. Thompson, J.J. Cronican, D.R. Liu, Engineering and identifying supercharged proteins for macromolecule delivery into mammalian cells, in: Methods Enzymol., Elsevier, 2012: pp. 293–319.

[171] M. Kornreich, E. Malka-Gibor, B. Zuker, A. Laser-Azogui, R. Beck, Neurofilaments function as shock absorbers: compression response arising from disordered proteins, Phys. Rev. Lett. 117 (2016) 148101.

[172] T.J. Nott, T.D. Craggs, A.J. Baldwin, Membraneless organelles can melt nucleic acid duplexes and act as biomolecular filters, Nat. Chem. 8 (2016) 569–575.

[173] K. Liu, D. Pesce, C. Ma, M. Tuchband, M. Shuai, D. Chen, J. Su, Q. Liu, J.Y. Gerasimov, A. Kolbe, others, Solvent-Free Liquid Crystals and Liquids Based on Genetically Engineered Supercharged Polypeptides with High Elasticity, Adv. Mater. 27 (2015) 2459–2465.

[174] M.C. Huber, A. Schreiber, P. Von Olshausen, B.R. Varga, O. Kretz, B. Joch, S. Barnert, R. Schubert, S. Eimer, P. Kele, others, Designer amphiphilic proteins as building blocks for the intracellular formation of organelle-like compartments, Nat. Mater. 14 (2015) 125–132.

[175] R.R. Costa, A.M. Testera, F.J. Arias, J.C. Rodriguez-Cabello, J.F. Mano, Layer-by-layer

film growth using polysaccharides and recombinant polypeptides: a combinatorial approach, J. Phys. Chem. B. 117 (2013) 6839–6848.

[176] I. Gitlin, J.D. Carbeck, G.M. Whitesides, Why are proteins charged? Networks of charge-charge interactions in proteins measured by charge ladders and capillary electrophoresis, Angew. Chemie - Int. Ed. 45 (2006) 3022–3060.

[177] D. Stigter, D.O. Alonso, K.A. Dill, Protein stability: electrostatics and compact denatured states., Proc. Natl. Acad. Sci. 88 (1991) 4176–4180.

[178] K.L. Shaw, G.R. Grimsley, G.I. Yakovlev, A.A. Makarov, C.N. Pace, The effect of net charge on the solubility, activity, and stability of ribonuclease Sa, Protein Sci. 10 (2001) 1206–1215.

[179] Y. Jiang, A. Lee, J. Chen, V. Ruta, M. Cadene, B.T. Chait, R. MacKinnon, X-ray structure of a voltage-dependent K+ channel, Nature. 423 (2003) 33–41.

[180] I. Muegge, P.X. Qi, A.J. Wand, Z.T. Chu, A. Warshel, The reorganization energy of cytochrome c revisited, J. Phys. Chem. B. 101 (1997) 825–836.

[181] C. Park, R.T. Raines, Quantitative analysis of the effect of salt concentration on enzymatic catalysis, J. Am. Chem. Soc. 123 (2001) 11472–11479.

[182] S.J. Davis, E.A. Davies, M.G. Tucknott, E.Y. Jones, P.A. Van Der Merwe, The role of charged residues mediating low affinity protein--protein recognition at the cell surface by CD2, Proc. Natl. Acad. Sci. 95 (1998) 5490–5494.

[183] I. Moarefi, D. Jeruzalmi, J. Turner, M. O'Donnell, J. Kuriyan, Crystal structure of the DNA polymerase processivity factor of T4 bacteriophage, J. Mol. Biol. 296 (2000) 1215–1223.

[184] A.S. Yang, B. Honig, On the pH dependence of protein stability, J. Mol. Biol. 231 (1993)

459–474.

[185] A.H. Elcock, D. Sept, J.A. McCammon, Computer simulation of protein- protein interactions, (2001).

[186] A.H. Elcock, M.J. Potter, D.A. Matthews, D.R. Knighton, J.A. McCammon, Electrostatic channeling in the bifunctional enzyme dihydrofolate reductase-thymidylate synthase, J. Mol. Biol. 262 (1996) 370–374.

[187] A. Warshel, Computer simulations of enzyme catalysis: methods, progress, and insights, Annu. Rev. Biophys. Biomol. Struct. 32 (2003) 425–443.

[188] M.S. Lawrence, K.J. Phillips, D.R. Liu, Supercharging proteins can impart unusual resilience, J. Am. Chem. Soc. 129 (2007) 10110–10112.

[189] B.R. McNaughton, J.J. Cronican, D.B. Thompson, D.R. Liu, Mammalian cell penetration, siRNA transfection, and DNA transfection by supercharged proteins, Proc. Natl. Acad. Sci. 106 (2009) 6111–6116.

[190] S.J. Remington, Green fluorescent protein: a perspective, Protein Sci. 20 (2011) 1509–1519.

[191] M. Zimmer, Green fluorescent protein (GFP): applications, structure, and related photophysical behavior, Chem. Rev. 102 (2002) 759–782.

[192] E.N. Lee, Y.M. Kim, H.J. Lee, S.W. Park, H.Y. Jung, J.M. Lee, Y.-H. Ahn, J. Kim, Stabilizing peptide fusion for solving the stability and solubility problems of therapeutic proteins, Pharm. Res. 22 (2005) 1735–1746.

[193] H.J.P. Ryser, R. Hancock, Histones and basic polyamino acids stimulate the uptake of albumin by tumor cells in culture, Science. 150 (1965) 501–503.

[194] A. Ziegler, P. Nervi, M. Dürrenberger, J. Seelig, The cationic cell-penetrating peptide

CPPTAT derived from the HIV-1 protein TAT is rapidly transported into living fibroblasts: Optical, biophysical, and metabolic evidence, Biochemistry. 44 (2005) 138–148.

[195] M. Rizzuti, M. Nizzardo, C. Zanetta, A. Ramirez, S. Corti, Therapeutic applications of the cell-penetrating HIV-1 Tat peptide, Drug Discov. Today. 20 (2015) 76–85.

[196] A.K. Dotiwala, C. McCausland, N.S. Samra, Anatomy, Head and Neck, Blood Brain Barrier, in: StatPearls [Internet], StatPearls Publishing, 2019.

[197] S. Aroui, A. Kenani, Cell-Penetrating Peptides: A Challenge for Drug Delivery, in: Cheminformatics Its Appl., IntechOpen, 2020.

[198] T. Lehto, K. Kurrikoff, Ü. Langel, Cell-penetrating peptides for the delivery of nucleic acids, Expert Opin. Drug Deliv. 9 (2012) 823–836.

[199] B.G. Bitler, J.A. Schroeder, Anti-cancer therapies that utilize cell penetrating peptides, Recent Pat. Anticancer. Drug Discov. 5 (2010) 99–108.

[200] A. Baoum, D. Ovcharenko, C. Berkland, Calcium condensed cell penetrating peptide complexes offer highly efficient, low toxicity gene silencing, Int. J. Pharm. 427 (2012) 134–142.

[201] C. De Coupade, A. Fittipaldi, V. Chagnas, M. Michel, S. Carlier, E. Tasciotti, A. Darmon, D. Ravel, J. Kearsey, M. Giacca, others, Novel human-derived cell-penetrating peptides for specific subcellular delivery of therapeutic biomolecules, Biochem. J. 390 (2005) 407–418.

[202] E.J.B. Nielsen, S. Yoshida, N. Kamei, R. Iwamae, E.-S. Khafagy, J. Olsen, U.L. Rahbek, B.L. Pedersen, K. Takayama, M. Takeda-Morishita, In vivo proof of concept of oral insulin delivery based on a co-administration strategy with the cell-penetrating peptide

penetratin, J. Control. Release. 189 (2014) 19–24.

[203] D. Chu, W. Xu, R. Pan, Y. Ding, W. Sui, P. Chen, Rational modification of oligoarginine for highly efficient siRNA delivery: structure--activity relationship and mechanism of intracellular trafficking of siRNA, Nanomedicine Nanotechnology, Biol. Med. 11 (2015) 435–446.

[204] E. Vives, P. Brodin, B. Lebleu, A truncated HIV-1 Tat protein basic domain rapidly translocates through the plasma membrane and accumulates in the cell nucleus, J. Biol. Chem. 272 (1997) 16010–16017.

[205] E.P. Loret, P.S. Ho, W.C. Johnson, E. Vives, H. Rochat, J. Van Rietschoten, Activating Region of HIV-1 Tat Protein: Vacuum UV Circular Dichroism and Energy Minimization, Biochemistry. 30 (1991) 6013–6023.

[206] E. Virès, C. Granier, P. Prevot, B. Lebleu, Structure-activity relationship study of the plasma membrane translocating potential of a short peptide from HIV-1 Tat protein, Lett. Pept. Sci. 4 (1997) 429–436.

[207] R. Tréhin, H.P. Merkle, Chances and pitfalls of cell penetrating peptides for cellular drug delivery, Eur. J. Pharm. Biopharm. 58 (2004) 209–223.

[208] V. Polyakov, V. Sharma, J.L. Dahlheimer, C.M. Pica, G.D. Luker, D. Piwnica-Worms, Novel Tat-peptide chelates for direct transduction of technetium-99m and rhenium into human cells for imaging and radiotherapy, Bioconjug. Chem. 11 (2000) 762–771.

[209] R. Bhorade, R. Weissleder, T. Nakakoshi, A. Moore, C.-H. Tung, Macrocyclic chelators with paramagnetic cations are internalized into mammalian cells via a HIV-tat derived membrane translocation peptide, Bioconjug. Chem. 11 (2000) 301–305.

[210] D.C. Anderson, E. Nichols, R. Manger, D. Woodle, M. Barry, A.R. Fritzberg, Tumor cell

retention of antibody Fab fragments is enhanced by an attached HIV TAT protein-derived peptide, Biochem. Biophys. Res. Commun. 194 (1993) 876–884.

[211] S. Stein, A. Weiss, K. Adermann, P. Lazarovici, J. Hochman, H. Wellhöner, A disulfide conjugate between anti-tetanus antibodies and HIV (37--72) Tat neutralizes tetanus toxin inside chromaffin cells, Febs Lett. 458 (1999) 383–386.

[212] L. Guelen, H. Paterson, J. Gäken, M. Meyers, F. Farzaneh, M. Tavassoli, TAT-apoptin is efficiently delivered and induces apoptosis in cancer cells, Oncogene. 23 (2004) 1153–1165.

[213] D.T. Kim, D.J. Mitchell, D.G. Brockstedt, L. Fong, G.P. Nolan, C.G. Fathman, E.G. Engleman, J.B. Rothbard, Introduction of soluble proteins into the MHC class I pathway by conjugation to an HIV tat peptide., J. Immunol. 159 (1997) 1666–1668.

[214] N. Shibagaki, M.C. Udey, Dendritic cells transduced with protein antigens induce cytotoxic lymphocytes and elicit antitumor immunity, J. Immunol. 168 (2002) 2393–2401.

[215] V.P. Torchilin, T.S. Levchenko, TAT-liposomes: a novel intracellular drug carrier, Curr. Protein Pept. Sci. 4 (2003) 133–140.

[216] M. Lewin, N. Carlesso, C.-H. Tung, X.-W. Tang, D. Cory, D.T. Scadden, R. Weissleder, Tat peptide-derivatized magnetic nanoparticles allow in vivo tracking and recovery of progenitor cells, Nat. Biotechnol. 18 (2000) 410–414.

[217] A. Nori, K.D. Jensen, M. Tijerina, P. Kopecková, J. Kopeček, Tat-conjugated synthetic macromolecules facilitate cytoplasmic drug delivery to human ovarian carcinoma cells, Bioconjug. Chem. 14 (2003) 44–50.

[218] A. Kato, K. Maki, T. Ebina, K. Kuwajima, K. Soda, Y. Kuroda, Mutational analysis of protein solubility enhancement using short peptide tags, Biopolym. Orig. Res. Biomol. 85

(2007) 12–18.

[219] C. di Guana, P. Lib, P.D. Riggsa, H. Inouyeb, Vectors that facilitate the expression and purification of foreign peptides in Escherichia coli by fusion to maltose-binding protein, Gene. 67 (1988) 21–30.

[220] D.B. Smith, K.S. Johnson, Single-step purification of polypeptides expressed in Escherichia coli as fusions with glutathione S-transferase, Gene. 67 (1988) 31–40.

[221] E. Bianchi, S. Venturini, A. Pessi, A. Tramontano, M. Sollazzo, High level expression and rational mutagenesis of a designed protein, the minibody: from an insoluble to a soluble molecule, J. Mol. Biol. 236 (1994) 649–659.

[222] M.M. Islam, M.A. Khan, Y. Kuroda, Analysis of amino acid contributions to protein solubility using short peptide tags fused to a simplified BPTI variant, Biochim. Biophys. Acta (BBA)-Proteins Proteomics. 1824 (2012) 1144–1150.

[223] M.M. Islam, S. Nakamura, K. Noguchi, M. Yohda, S. Kidokoro, Y. Kuroda, Analysis and control of protein crystallization using short peptide tags that change solubility without affecting structure, thermal stability, and function, Cryst. Growth Des. 15 (2015) 2703–2711.

[224] C.L. Young, Z.T. Britton, A.S. Robinson, Recombinant protein expression and purification: a comprehensive review of affinity tags and microbial applications, Biotechnol. J. 7 (2012) 620–634.

[225] M.E. Kimple, A.L. Brill, R.L. Pasker, Overview of affinity tags for protein purification, Curr. Protoc. Protein Sci. 73 (2013) 9.

[226] D.S. Waugh, An overview of enzymatic reagents for the removal of affinity tags, Protein Expr. Purif. 80 (2011) 283–293.

[227] B. Dang, M. Mravic, H. Hu, N. Schmidt, B. Mensa, W.F. DeGrado, SNAC-tag for sequence-specific chemical protein cleavage, Nat. Methods. 16 (2019) 319–322.

[228] A. Krezel, E. Kopera, A.M. Protas, J. Poznański, A. Wysłouch-Cieszyńska, W. Bal, Sequence-specific Ni (II)-dependent peptide bond hydrolysis for protein engineering. Combinatorial library determination of optimal sequences, J. Am. Chem. Soc. 132 (2010) 3355–3366.

[229] G.L. Glish, R.W. Vachet, The basics of mass spectrometry in the twenty-first century, Nat. Rev. Drug Discov. 2 (2003) 140–150.

[230] D.A. Williams, Fundamentals of mass spectrometry, in: Pharmacochem. Libr., Elsevier, 1997: 19–45.

[231] A.B.T. Ghisaidoobe, S.J. Chung, Intrinsic tryptophan fluorescence in the detection and analysis of proteins: a focus on Förster resonance energy transfer techniques, Int. J. Mol. Sci. 15 (2014) 22518–22538.

[232] M. Arik, N. Çelebi, Y. Onganer, Fluorescence quenching of fluorescein with molecular oxygen in solution, J. Photochem. Photobiol. A Chem. 170 (2005) 105–111.

[233] M.R. Eftink, The use of fluorescence methods to monitor unfolding transitions in proteins, Biophys. J. 66 (1994) 482–501.

[234] R. Swaminathan, G. Krishnamoorthy, N. Periasamy, Similarity of fluorescence lifetime distributions for single tryptophan proteins in the random coil state, Biophys. J. 67 (1994) 2013–2023.

[235] A.G. Szabo, T.M. Stepanik, D.M. Wayner, N.M. Young, Conformational heterogeneity of the copper binding site in azurin. A time-resolved fluorescence study, Biophys. J. 41 (1983) 233–244.

[236] D.A. and S.M.D. Natalello, Antonino, Fourier Transform infrared spectroscopy of IDP measurements procedures and data analysis, Silicon Agric. 8 (2001) 85–113.

[237] A. Barth, Infrared spectroscopy of proteins, Biochim. Biophys. Acta - Bioenerg. 1767 (2007) 1073–1101.

[238] D. Usoltsev, V. Sitnikova, A. Kajava, M. Uspenskaya, Systematic FTIR spectroscopy study of the secondary structure changes in human serum albumin under various denaturation conditions, Biomolecules. 9 (2019) 1–17.

[239] S. Cai, B.R. Singh, Identification of β-turn and random coil amide III infrared bands for secondary structure estimation of proteins, Biophys. Chem. 80 (1999) 7–20.

[240] L.M. Miller, M.W. Bourassa, R.J. Smith, FTIR spectroscopic imaging of protein aggregation in living cells, Biochim. Biophys. Acta - Biomembr. 1828 (2013) 2339–2346.

[241] E. Goormaghtigh, V. Cabiaux, J.M. Ruysschaert, Determination of soluble and membrane protein structure by Fourier transform infrared spectroscopy. II. Experimental aspects, side chain structure, and H/D exchange., Sub-Cellular Biochem. 23 (1994) 363-403

[242] F. Bloch, Nuclear induction, Phys. Rev. 70 (1946) 460.

[243] E.M. Purcell, H.C. Torrey, R. V Pound, Resonance absorption by nuclear magnetic moments in a solid, Phys. Rev. 69 (1946) 37.

[244] D. Marion, An introduction to biological NMR spectroscopy, Mol. Cell. Proteomics. 12 (2013) 3006–3025.

[245] D. Ban, C.A. Smith, B.L. de Groot, C. Griesinger, D. Lee, Recent advances in measuring the kinetics of biomolecules by NMR relaxation dispersion spectroscopy, Arch. Biochem. Biophys. 628 (2017) 81–91.

[246] M.D. Mantle, NMR and MRI studies of drug delivery systems, Curr. Opin. Colloid

Interface Sci. 18 (2013) 214–227.

[247] E.U. Saritas, P.W. Goodwill, L.R. Croft, J.J. Konkle, K. Lu, B. Zheng, S.M. Conolly, Magnetic particle imaging (MPI) for NMR and MRI researchers, J. Magn. Reson. 229 (2013) 116–126.

[248] A.J. Baldwin, L.E. Kay, NMR spectroscopy brings invisible protein states into focus, Nat. Chem. Biol. 5 (2009) 808–814.

[249] H. Yu, Extending the size limit of protein nuclear magnetic resonance, Proc. Natl. Acad. Sci. U. S. A. 96 (1999) 332–334.

[250] PDB Statistics: Growth of Structures from NMR Experiments Released per Year, (n.d.). https://www.rcsb.org/stats/growth/growth-nmr.

[251] S.A. Goudsmit, Pauli and nuclear spin, Phys. Today. 14 (1961) 18.

[252] J. Cavanagh, W.J. Fairbrother, A.G. Palmer III, N.J. Skelton, Protein NMR spectroscopy: principles and practice, Elsevier, 1995.

[253] J. Keeler, Understanding NMR spectroscopy, John Wiley & Sons, 2011.

[254] V.A. Jarymowycz, M.J. Stone, Fast time scale dynamics of protein backbones: NMR relaxation methods, applications, and functional consequences, Chem. Rev. 106 (2006) 1624–1671.

[255] D.S. Wishart, A.M. Nip, Protein chemical shift analysis: A practical guide, Biochem. Cell Biol. 76 (1998) 153–163.

[256] D.S. Wishart, B.D. Sykes, F.M. Richards, Relationship between nuclear magnetic resonance chemical shift and protein secondary structure, J. Mol. Biol. 222 (1991) 311–333.

[257] D.S. Wishart, B.D. Sykes, F.M. Richards, The Chemical Shift Index : A Fast and Simple

Method for the Assignment of Protein Secondary Structure Through NMR Spectroscopy, Biochemistry. 31 (1992) 1647–1651.

[258] J. Li, K.C. Bennett, Y. Liu, M. V. Martin, T. Head-Gordon, Accurate prediction of chemical shifts for aqueous protein structure on “real World” data, Chem. Sci. 11 (2020) 3180–3191.

[259] A. Cavalli, X. Salvatella, C.M. Dobson, M. Vendruscolo, Protein structure determination from NMR chemical shifts, Proc. Natl. Acad. Sci. U. S. A. 104 (2007) 9615–9620.

[260] M. Kjaergaard, F.M. Poulsen, Disordered proteins studied by chemical shifts, Prog. Nucl. Magn. Reson. Spectrosc. 60 (2012) 42–51.

[261] S. Spera, A. Bax, Empirical Correlation between Protein Backbone Conformation and Cα and Cβ 13C Nuclear Magnetic Resonance Chemical Shifts, J. Am. Chem. Soc. 113 (1991) 5490–5492.

[262] A.C. De Dios, J.G. Pearson, E. Oldfield, Secondary and tertiary structural effects on protein NMR chemical shifts: An ab initio approach, Science. 260 (1993) 1491–1496.

[263] N.E. Hafsa, D. Arndt, D.S. Wishart, CSI 3.0: A web server for identifying secondary and super-secondary structure in proteins using NMR chemical shifts, Nucleic Acids Res. 43 (2015) W370–W377.

[264] J.A. Marsh, V.K. Singh, Z. Jia, J.D. Forman-Kay, Sensitivity of secondary structure propensities to sequence differences between α- and γ-synuclein: Implications for fibrillation, Protein Sci. 15 (2006) 2795–2804.

[265] D.S. Wishart, C.G. Bigam, A. Holm, R.S. Hodges, B.D. Sykes, 1 H, 13 C and 15 N random coil NMR chemical shifts of the common amino acids. I. Investigations of nearest-neighbor effects, J. Biomol. NMR. 5 (1995) 67–81.

[266] J.M. Schmidt, M.J. Howard, M. Maestre-Martínez, C.S. Pérez, F. Löhr, Variation in protein Cα-related one-bond J couplings, Magn. Reson. Chem. 47 (2009) 16–30.

[267] L.J. Smith, K.A. Bolin, H. Schwalbe, M.W. MacArthur, J.M. Thornton, C.M. Dobson, Analysis of main chain torsion angles in proteins: Prediction of NMR coupling constants for native and random coil conformations, J. Mol. Biol. 255 (1996) 494–506.

[268] R.R. Gil, Residual Dipolar Couplings in Small-Molecule NMR, In Encyclopedia of Spectroscopy and Spectrometry, J.C. Lindon, G.E. Tranter, D.W. Koppenaal, Eds., Academic press Ltd, Elsevier Science Ltd: London, UK, (2017) 946–955.

[269] A. Annila, P. Permi, Weakly aligned biological macromolecules in dilute aqueous liquid crystals, Concepts Magn. Reson. Part A An Educ. J. 23 (2004) 22–37.

[270] M.R. Gryk, J.C. Hoch, Local knowledge helps determine protein structures, Proc. Natl. Acad. Sci. U. S. A. 105 (2008) 4533–4534.

[271] S. Pal, Structure analysis and visualization, In Fundam. Mol. Struct. Biol. Academic press Ltd, Elsevier Science Ltd: London, UK, (2020) 119–147.

[272] S. Kosol, S. Contreras-Martos, C. Cedeño, P. Tompa, Structural characterization of intrinsically disordered proteins by NMR spectroscopy, Molecules. 18 (2013) 10802–10828.

[273] H. Saitô, Dynamic pictures of proteins by NMR, In Annual reports on NMR spectroscopy, A.W. Graham., Eds., Academic press Ltd, Elsevier Sceince Ltd: London, UK, 83 (2014) 1–66.

[274] I.R. Kleckner, M.P. Foster, An introduction to NMR-based approaches for measuring protein dynamics, Biochim. Biophys. Acta - Proteins Proteomics. 1814 (2011) 942–968.

[275] J.G. Kempf, J.P. Loria, Protein dynamics from solution NMR: Theory and applications,

Cell Biochem. Biophys. 37 (2002) 187–211.

[276] L.E. Kay, Protein dynamics from NMR, Biochem Cell Biol. 76 (1998) 145–152.

[277] C. Göbl, N. Tjandra, Application of solution NMR spectroscopy to study protein dynamics, Entropy. 14 (2012) 581–598.

[278] P. Neudecker, P. Lundström, L.E. Kay, Relaxation dispersion NMR spectroscopy as a tool for detailed studies of protein folding, Biophys. J. 96 (2009) 2045–2054.

[279] D.D. Boehr, H.J. Dyson, P.E. Wright, An NMR perspective on enzyme dynamics, Chem. Rev. 106 (2006) 3055–3079.

[280] V. Kharchenko, M. Nowakowski, M. Jaremko, A. Ejchart, Ł. Jaremko, Dynamic $^{15}N\{^{1}H\}$ NOE measurements: a tool for studying protein dynamics, J. Biomol. NMR. 74 (2020) 707–716.

[281] A.G. Palmer, NMR characterization of the dynamics of biomacromolecules, Chem. Rev. 104 (2004) 3623–3640.

[282] D. Eleisar, Biophysical characterization of intrinsically disordered proteins, Bone. 23 (2008) 1–7.

[283] N. Rezaei-Ghaleh, M. Blackledge, M. Zweckstetter, Intrinsically Disordered Proteins: From Sequence and Conformational Properties toward Drug Discovery, ChemBioChem. 13 (2012) 930–950.

[284] S. Kim, K.P. Wu, J. Baum, Fast hydrogen exchange affects ^{15}N relaxation measurements in intrinsically disordered proteins, J. Biomol. NMR. 55 (2013) 249–256.

[285] A.G. Palmer III, Probing molecular motion by NMR, Curr. Opin. Struct. Biol. 7 (1997) 732–737.

[286] A. Kumar, J. Balbach, Real-time protein NMR spectroscopy and investigation of assisted

protein folding, Biochim. Biophys. Acta (BBA)-General Subj. 1850 (2015) 1965–1972.

[287] G. Ortega, M. Pons, O. Millet, Protein functional dynamics in multiple timescales as studied by NMR spectroscopy, Adv. Protein Chem. Struct. Biol. 92 (2013) 219–251.

[288] P. Giraudeau, Quantitative 2D liquid-state NMR, Magn. Reson. Chem. 52 (2014) 259–272.

[289] A. Herrera, E. Fernández-Valle, R. Martinez-Álvarez, D. Molero, Z.D. Pardo, E. Sáez, M. Gal, Real-Time Monitoring of Organic Reactions with Two-Dimensional Ultrafast TOCSY NMR Spectroscopy, Angew. Chemie Int. Ed. 48 (2009) 6274–6277.

[290] I.C. Felli, R. Pierattelli, Intrinsically disordered proteins studied by NMR spectroscopy, Springer International Publishing. 870 (2015).

[291] I.C. Felli, A. Piai, R. Pierattelli, Recent advances in solution NMR studies: ^{13}C direct detection for biomolecular NMR applications, 1st Ed., Elsevier Ltd., 2013.

[292] M. Gal, K.A. Edmonds, A.G. Milbradt, K. Takeuchi, G.Wagner., Speeding up direct ^{15}N detection- hCaN 2D NMR experiment, J. Biomol. NMR. 51 (2011) 497–504.

[293] S. Chhabra, P. Fischer, K. Takeuchi, A. Dubey, J.J. Ziarek, A. Boeszoermenyi, D. Mathieu, W. Bermel, N.E. Davey, G. Wagner, H. Arthanari, 15N detection harnesses the slow relaxation property of nitrogen: Delivering enhanced resolution for intrinsically disordered proteins, Proc. Natl. Acad. Sci. U. S. A. 115 (2018) E1710–E1719.

[294] K. Takeuchi, G. Heffron, Z.Y.J. Sun, D.P. Frueh, G. Wagner, Nitrogen-detected CAN and CON experiments as alternative experiments for main chain NMR resonance assignments, J. Biomol. NMR. 47 (2010) 271–282.

[295] Y. Nekooki-Machida, M. Kurosawa, N. Nukina, K. Ito, T. Oda, M. Tanaka, Distinct conformations of in vitro and in vivo amyloids of huntingtin-exon1 show different

cytotoxicity, Proc. Natl. Acad. Sci. 106 (2009) 9679–9684.

[296] V. Ladizhansky, Applications of solid-state NMR to membrane proteins, Biochim. Biophys. Acta (BBA)-Proteins Proteomics. 1865 (2017) 1577–1586.

[297] I. Matlahov, P.C.A. van der Wel, Hidden motions and motion-induced invisibility: dynamics-based spectral editing in solid-state NMR, Methods. 148 (2018) 123–135.

[298] A.B. Siemer, Advances in studying protein disorder with solid-state NMR, Solid State Nucl. Magn. Reson. 106 (2020) 101643.

[299] E. Delaforge, T. Cordeiro, P. Bernadó, N. Sibille, Conformational characterization of intrinsically disordered proteins and its biological significance, Webb G.A., Ed., Modern Magnetic Resonance. Cham, Switzerland: Springer International Publishing. (2017) 1–20.

[300] D. Kruschel, B. Zagrovic, Conformational averaging in structural biology: issues, challenges and computational solutions, Mol. Biosyst. 5 (2009) 1606–1616.

[301] M.J. Duer (Ed.,) Introduction to solid-state NMR spectroscopy, Oxford, UK, Malden, MA: Blackwell (2004) 116–125.

[302] M.J. Duer, Solid state NMR spectroscopy: Principles and Applications. Blackwell Sci. Ltd., Oxford, UK (2002)

[303] S.R. Hartmann, E.L. Hahn, Nuclear double resonance in the rotating frame, Phys. Rev. 128 (1962) 2042.

[304] T.M. Alam, G.P. Holland, 1H--13C INEPT MAS NMR correlation experiments with ^{1}H-^{1}H mediated magnetization exchange to probe organization in lipid biomembranes, J. Magn. Reson. 180 (2006) 210–221.

[305] N. Bloembergen, On the interaction of nuclear spins in a crystalline lattice, Physica. 15 (1949) 386–426.

[306] OptimumGene - Codon Optimization, https://www.genscript.com/codon_opt_pr.html

[307] G.A. Gutman, G.W. Hatfield, Nonrandom utilization of codon pairs in Escherichia coli., Proc. Natl. Acad. Sci. U. S. A. 86 (1989) 3699–3703.

[308] S. Boycheva, G. Chkodrov, I. Ivanov, Codon pairs in the genome of Escherichia coli, Bioinformatics. 19 (2003) 987–998.

[309] M.H. Caruthers, A brief review of DNA and RNA chemical synthesis, Biochem. Soc. Trans. 39 (2011) 575–580.

[310] Á. Somoza, Protecting groups for RNA synthesis: an increasing need for selective preparative methods, Chem. Soc. Rev. 37 (2008) 2668–2675.

[311] J.F. Milligan, D.R. Groebe, G.W. Witherell, O.C. Uhlenbeck, Oligoribonucleotide synthesis using T7 RNA polymerase and synthetic DNA templates, Nucleic Acids Res. 15 (1987) 8783–8798.

[312] E.P. Booy, H. Meng, S.A. Mckenna, Chapter 6 Native RNA Puri fi cation by Gel Filtration Chromatography, Recomb. Vitr. RNA Synth. Methods Protoc. 941 (2012) 69–81.

[313] J.D. Puglisi, I. Tinoco, Absorbance melting curves of RNA, Methods Enzymol. 180 (1989) 304–325.

[314] ExPASy - ProtParam tool, https://web.expasy.org/protparam.

[315] F. Delaglio, S. Grzesiek, G.W. Vuister, G. Zhu, J. Pfeifer, A. Bax, NMRPipe: A multidimensional spectral processing system based on UNIX pipes, J. Biomol. NMR. 6 (1995) 277–293.

[316] W. Lee, M. Tonelli, J.L. Markley, NMRFAM-SPARKY: Enhanced software for biomolecular NMR spectroscopy, Bioinformatics. 31 (2015) 1325–1327.

[317] H. Yang, S. Yang, J. Kong, A. Dong, S. Yu, Obtaining information about protein secondary structures in aqueous solution using Fourier transform IR spectroscopy, Nat. Protoc. 10 (2015) 382–396.

[318] Y.K. Reshetnyak, Y. Koshevnik, E.A. Burstein, Decomposition of protein tryptophan fluorescence spectra into log-normal components. III. Correlation between fluorescence and microenvironment parameters of individual tryptophan residues, Biophys. J. 81 (2001) 1735–1758.

[319] P.R. Callis, T. Liu, Quantitative prediction of fluorescence quantum yields for tryptophan in proteins, J. Phys. Chem. B. 108 (2004) 4248–4259.

[320] S. Sun, Y. Han, S. Paramasivam, S. Yan, A.E. Siglin, J.C. Williams, I.-J.L. Byeon, J. Ahn, A.M. Gronenborn, T. Polenova, Solid-state NMR spectroscopy of protein complexes, in: Protein NMR Tech., Springer, (2012) 303–331.

[321] O.C. Andronesi, S. Becker, K. Seidel, H. Heise, H.S. Young, M. Baldus, Determination of membrane protein structure and dynamics by magic-angle-spinning solid-state NMR spectroscopy, J. Am. Chem. Soc. 127 (2005) 12965–12974.

[322] R. Tycko, Solid-state NMR studies of amyloid fibril structure, Annu. Rev. Phys. Chem. 62 (2011) 279–299.

[323] U. Schulze-Gahmen, J.H. Hurley, Structural mechanism for HIV-1 TAR loop recognition by Tat and the super elongation complex, Proc. Natl. Acad. Sci. 115 (2018) 12973–12978.

[324] T.M. Rana, K.-T. Jeang, Biochemical and functional interactions between HIV-1 Tat protein and TAR RNA, Arch. Biochem. Biophys. 365 (1999) 175–185.

[325] M.J. Churcher, C. Lamont, F. Hamy, C. Dingwall, S.M. Green, A.D. Lowe, P.J.G. Butler, M.J. Gait, J. Karn, High affinity binding of TAR RNA by the human immunodeficiency

virus type-1 tat protein requires base-pairs in the RNA stem and amino acid residues flanking the basic region, J. Mol. Biol. 230 (1993) 90–110.

[326] K.M. Weeks, D.M. Crothers, RNA recognition by Tat-derived peptides: interaction in the major groove?, Cell. 66 (1991) 577–588.

[327] B.A. Moffatt, J.J. Dunn, F.W. Studier, Nucleotide sequence of the gene for bacteriophage T7 RNA polymerase, J. Mol. Biol. 173 (1984) 265–269.

[328] J.C. Cottrell, B.N. Green, MaxEnt: An Essential Maximum Entropy Based Tool for Interpreting Multiply-Charged Electrospray Data, (1998).

[329] M.E. Garber, P. Wei, V.N. KewalRamani, T.P. Mayall, C.H. Herrmann, A.P. Rice, D.R. Littman, K.A. Jones, The interaction between HIV-1 Tat and human cyclin T1 requires zinc and a critical cysteine residue that is not conserved in the murine CycT1 protein, Genes Dev. 12 (1998) 3512–3527.

[330] S. Ruben, A. Perkins, R. Purcell, K. Joung, R. Sia, R. Burghoff, W.A. Haseltine, C.A. Rosen, Structural and functional characterization of human immunodeficiency virus tat protein., J. Virol. 63 (1989) 1–8.

[331] M.R. Sadaei, R. Mukhopadhyaya, Z.N. Benaissa, G.N. Pavlakis, F. Wong-Staal, Conservative mutations in the putative metal-binding region of human immunodeficiency virus tat disrupt virus replication, AIDS Res. Hum. Retroviruses. 6 (1990) 1257–1263.

[332] F.X. Schmid, Biological macromolecules: UV-visible spectrophotometry, Encyclopedia of Life Sciences. Macmillan Publishers, New York (2001).

[333] F. Chiti, M. Stefani, N. Taddei, G. Ramponi, C.M. Dobson, Rationalization of the effects of mutations on peptide andprotein aggregation rates, Nature. 424 (2003) 805–808.

[334] S. Il Choi, K.S. Han, C.W. Kim, K.-S. Ryu, B.H. Kim, K.-H. Kim, S.-I. Kim, T.H. Kang,

H.-C. Shin, K.-H. Lim, others, Protein solubility and folding enhancement by interaction with RNA, PLoS One. 3 (2008) e2677.

[335] K.A. Black, D. Priftis, S.L. Perry, J. Yip, W.Y. Byun, M. Tirrell, Protein encapsulation via polypeptide complex coacervation, ACS Macro Lett. 3 (2014) 1088–1091.

[336] W.C.B. McTigue, S.L. Perry, Design rules for encapsulating proteins into complex coacervates, Soft Matter. 15 (2019) 3089–3103.

[337] D. Priftis, K. Megley, N. Laugel, M. Tirrell, Complex coacervation of poly (ethylene-imine)/polypeptide aqueous solutions: Thermodynamic and rheological characterization, J. Colloid Interface Sci. 398 (2013) 39–50.

[338] S.L. Perry, Y. Li, D. Priftis, L. Leon, M. Tirrell, The effect of salt on the complex coacervation of vinyl polyelectrolytes, Polymers (Basel). 6 (2014) 1756–1772.

Appendix I

Gene sequence of SNAC-tagged Tat protein

CCATGGAACCGGTCGACCCGCGTCTGGAACCGTGGAAACACCCCGGGTCCCAGCCG
AAAACCGCGTGCACCAACTGCTACTGCAAAAAATGCTGCTTCCACTGCCAGGTTTGC
TTCATCACCAAAGCCCTAGGTATCTCTTACGGCCGTAAAAAACGTCGTCAGCGACGT
CGTCCGCCGCAGGGATCCCAGACTCATCAAGTTTCCTTGTCCAAGCAACCGGCGTCT
CAGCCGCGTGGTGACCCGACCGGTCCGAAAGAATCTAAAAAAAAAAGTTGAACGTGA
AACCGAAACCGACCCGGTTGACGGTAGCCACCATTGGGGCAGCAGCCATCACCATC
ACCATCATAGCAGCGGTTAATGATAAGAATTC

Gene sequence of SuperCharged R_{10} tagged Tat protein

CCATGGGATCAAGTCATCACCATCACCACCACTCTAGCGGCGAGAACTTGTACTTCC
AAAGCGGTGGCCGTCGCCGCCGCCGCCGTCGTCGTCGTAGAGGCGGCATGGAGCCG
GTTGACCCGCGGCTGGAACCGTGGAAGCACCCAGGTTCTCAACCGAAAACCGCGTG
CACCAATTGTTACTGCAAAAAGTGCTGCTTCCACTGCCAGGTGTGTTTTATCACCAA
AGCTCTGGGTATTAGCTATGGTCGTAAAAAACGTCGTCAGCGTCGCAGACCTCCGCA
GGGTTCGCAAACCCATCAGGTCAGCCTGAGCAAGCAGCCGGCATCCCAACCGCGTG
GCGATCCGACCGGTCCGAAAGAGTCCAAGAAGAAGGTTGAACGTGAAACTGAGACC
GACCCGGTGGACTAATGATAGGATCC

Gene Sequence of RPB1 CTD of RNAP II

CCATGGGTAGCAGCCATCATCATCATCACCACGGTAGCAGCATGAAAATCGAAGAG
GGCAAACTGGTTATCTGGATCAACGGCGACAAAGGTTACAACGGCCTGGCGGAAGT
GGGTAAGAAATTCGAGAAAGACACCGGCATCAAGGTGACCGTTGAACACCCGGATA
AACTGGAGGAAAAGTTTCCGCAAGTTGCGGCGACCGGTGATGGTCCGGACATCATT
TTCTGGGCGCACGACCGTTTTGGTGGCTACGCGCAGAGCGGTCTGCTGGCGGAAATT
ACCCCGGACAAAGCGTTCCAAGATAAGCTGTATCCGTTTACCTGGGATGCGGTGCGT
TACAACGGCAAACTGATCGCGTATCCGATTGCGGTTGAGGCGCTGAGCCTGATCTAC
AACAAGGACCTGCTGCCGAACCCGCCGAAAACCTGGGAGGAAATTCCGGCGCTGGA
TAAGGAACTGAAGGCGAAAGGCAAGAGCGCGCTGATGTTCAACCTGCAGGAGCCGT
ACTTTACCTGGCCGCTGATTGCGGCGGATGGTGGCTACGCGTTCAAGTACGAAAACG
GCAAGTACGACATTAAGGATGTGGGCGTTGACAACGCGGGTGCGAAGGCGGGCCTG
ACCTTCCTGGTGGATCTGATCAAAAACAAGCACATGAACGCGGACACCGATTACAG
CATTGCGGAAGCGGCGTTTAACAAAGGTGAAACCGCGATGACCATCAACGGCCCGT
GGGCGTGGAGCAACATTGATACCAGCAAGGTTAACTACGGTGTGACCGTTCTGCCG
ACCTTCAAAGGCCAACCGAGCAAGCCGTTTGTGGGTGTTCTGAGCGCGGGTATCAAC
GCGGCGAGCCCGAACAAAGAGCTGGCGAAGGAATTTCTGGAGAACTACCTGCTGAC
CGACGAAGGTCTGGAGGCGGTGAACAAAGATAAGCCGCTGGGCGCGGTTGCGCTGA
AGAGCTACGAGGAAGAGCTGGCGAAAGACCCGCGTATCGCGGCGACGATGGAGAA
CGCGCAGAAAGGCGAGATCATGCCGAACATTCCGCAAATGAGCGCGTTCTGGTATG
CGGTGCGTACCGCGGTTATTAACGCGGCGAGCGGCCGTCAGACCGTGGACGAAGCG
CTGAAGGATGCGCAAACCAACAGCAGCAGCAATAATAACAATAATAACAACAACA
ACAACCTGGGTATCGAAGAGAACCTGTACTTTCAGAGCAACGCGTGCTATAGCCCG
ACCAGCCCGGCGTACGAGCCGCGTAGCCCGGGTGGCTATACCCCGCAAAGCCCGAG
CTACAGCCCGACCAGCCCGTCTTACTCTCCTACCAGCCCGTCTTATTCTCCGACCAGC
CCGAATTACTCTCCCACCAGCCCGTCTTATAGTCCGACCAGCCCGTCTTATAGCCCT
ACCAGCCCGAGTTATTCTCCTACCAGCCCTTCTTACTCTCCAACCAGCCCGAGTTATA
GTCCTACCAGCCCTAGTTACTCTCCGACCAGCCCTTCTTATAGCCCCACCAGCCCGA
GTTACAGTCCTACCAGCCCGAGCTATTCTCCTACCAGCCCGAGCTATAGTCCTACCA
GCCCCTCTTATAGCCCAACCAGCCCTAGTTACAGTCCCACCAGCCCGAGTTATAGCC
CTACCAGCCCATCTTACAGTCCAACCAGCCCGAGCTATAGCCCTACCAGCCCTAATT
ACAGTCCGACCAGCCCTAACTACACCCCGACCAGCCCTAGTTATAGCCCGACCAGCC
CTAGCTATTCTCCCACCAGCCCGAACTATACCCCGACCAGCCCGAACTACTCTCCTA
CCAGCCCTAGCTATAGTCCCACCAGCCCTAGCTATAGCCCGACCAGCCCAAGCTATA
GCCCGAGCAGCCCGCGTTACACCCCGCAAAGCCCGACCTATACCCCGAGCAGCCCC
TCCTACAGCCCGAGCAGCCCCTCTTACAGCCCTACCAGCCCTAAATACACCCCGACC
AGCCCCTCCTATTCTCCGAGCAGCCCTGAGTACACCCCGACCAGCCCGAAATATTCC
CCTACCAGCCCGAAGTACTCTCCAACCAGCCCTAAATATTCCCCCACCAGCCCGACC
TATTCTCCGACCACCCCGAAGTATTCCCCAACCAGCCCGACATATTCTCCAACCAGC
CCTGTTTATACCCCGACCAGCCCTAAGTACTCTCCCACCAGCCCTACTTACAGCCCG
ACCAGCCCGAAGTATTCCCCCACCAGCCCTACATATTCCCCAACCAGCCCTAAAGGT
AGCACCTACTCTCCCACCAGCCCCGGCTATAGCCCGACCAGCCCGACCTACAGCCTG
ACCAGCCCGGCGATTAGCCCGGACGATAGCGATGAAGAAAATTAATGATAGTCGAC

Appendix II

Resonance assignments of Tat at different pHs at 298K.

Amino acid and position	pH 4		pH 5		pH 6		pH 6.5		pH 7	
	Chemical shift (ppm)									
	^{1}H	^{15}N	^{1}H	^{15}N	^{1}H	^{15}N	^{1}H	^{15}N	^{1}H	^{15}N
Met 1										
Gly 2										
Ser 3	8.766	115.848								
Ser 4	8.567	117.486	8.563	117.971	8.598	118.002				
His 5	8.624	120.093	8.587	120.232						
His 6	8.629	119.301	8.581	119.392						
His 7	8.798	119.951								
His 8	8.859	120.230								
His 9	8.869	120.591								
His 10	8.861	121.213								
Ser 11	8.642	118.392								
Ser 12	8.615	118.404	8.611	118.418	8.614	118.423				
Gly 13	8.483	110.645								
Leu 14	8.165	121.821	8.157	121.752	8.175	121.878	8.129	121.655	8.104	121.677
Val 15	8.231	123.192	8.161	122.861	8.245	123.116				
Pro 16										
Arg 17	8.548	122.180	8.548	122.175	8.584	122.267	8.557	122.085		
Gly 18	8.537	110.467	8.479	110.629	8.506	110.794	8.472	110.611		
Ser 19	8.272	115.505	8.275	115.618	8.329	115.792	8.333	115.656		

His 20	8.665	120.220	8.649	120.283						
Met 21	8.452	121.747								
Glu 22	8.523	123.448								
Pro 23										
Val 24	8.427	120.480	8.216	120.763	8.272	120.734		120.561		120.644
Asp 25	8.454	125.713	8.439	126.204	8.485	126.436	8.463	126.166	8.459	126.216
Pro 26										
Arg 27	8.427	118.866	8.442	118.896	8.475	119.027				
Lys 28	7.946	120.340	7.914	120.150	7.943	120.296	7.934	120.333	7.925	120.418
Glu 29	7.997	120.464	8.031	120.525	8.098	120.876	8.101	120.939	8.101	121.111
Pro 30										
Trp 31	7.527	117.743	7.481	117.373	7.553	117.701	7.599	117.905	7.627	118.137
Lys 32	7.697	122.454	7.629	122.182	7.635	122.415	7.612	122.559	7.601	122.756
His 33	8.185	119.226	8.104	119.060						
Pro 34										
Gly 35	8.659	109.898	8.676	109.847	8.775	110.044				
Ser 36	8.249	115.502	8.231	115.464	8.258	115.583	8.222	115.468		
Gln 37										
Pro 38										
Lys 39	8.581	121.919	8.575	121.889	8.619	122.039				
Thr 40	8.132	115.024	8.063	114.871	8.151	115.025				
Ala 41	8.416	126.435								
Cys 42	8.448	118.968	8.474	118.793	8.511	118.950	8.493	118.819	8.491	118.923
Thr 43	8.324	116.462	8.317	116.403	8.352	116.512	8.328	116.339		
Asn 44										
Cys 45	8.293	119.217	8.285	119.229	8.316	119.329				
Tyr 46	8.344	122.595								

Cys 47										
Lys 48	8.384	124.319	8.382	124.214	8.439	124.471	8.429	124.415	8.424	124.525
Lys 49	8.432	123.566								
Cys 50	8.441	120.591	8.373	120.584						
Cys 51	8.431	121.739								
Phe 52	8.364	122.729								
His 53	8.578	120.628	8.511	120.764	8.514	121.143				
Cys 54	8.425	120.601	8.421	120.463						
Gln 55	8.633	123.315	8.275	123.905	8.307	124.101				
Val 56	8.286	121.884	8.275	121.872	8.307	121.841				
Cys 57	8.415	122.936								
Phe 58	8.418	123.521								
Ile 59	8.167	122.950	8.050	123.110	8.198	122.959	8.192	122.716	8.167	122.647
Thr 60	8.218	118.963	8.208	118.887	8.244	118.874				
Lys 61	8.348	124.308	8.338	124.218	8.334	124.331				
Ala 62	8.339	125.312								
Lys 63	8.250	121.586	8.236	121.485	8.271	121.581				
Gly 64	8.382	109.222	8.374	109.213	8.405	109.290				
Ile 65	7.989	119.969	7.982	119.996	8.021	120.093	7.993	119.914	7.696	119.901
Ser 66	8.368	119.250	8.361	119.233	8.401	119.366				
Tyr 67	8.301	122.832								
Gly 68	8.403	110.154								
Arg 69	8.227	120.833								
Lys 70	8.388	122.203								
Lys 71	8.395	122.825								
Arg 72										
Arg 73	8.534	123.084								

Gln 74	8.544	122.471								
Arg 75	8.553	123.476								
Arg 76	8.554	123.195								
Arg 77	8.546	124.436	8.545	124.373	8.586	124.593	8.555	124.487		
Pro 78										
Pro 79										
Gln 80	8.607	120.997	8.607	120.959	8.646	121.141	8.623	120.959		
Gly 81	8.559	110.599	8.549	110.527	8.592	110.711	8.569	110.581		
Ser 82	8.350	115.620	8.345	115.624	8.376	115.763	8.356	115.660		
Gln 83	8.598	122.213	8.601	122.178	8.641	122.345	8.615	122.181		
Thr 84	8.195	114.858	8.116	114.964	8.231	114.964	8.227	114.851		
His 85	8.596	120.699	8.571	120.687						
Gln 86	8.544	122.471	8.544	122.478	8.598	122.615	8.565	122.428		
Val 87	8.431	122.432								
Ser 88	8.489	119.724	8.483	119.678	8.513	119.772	8.487	119.537		
Lys 89	8.468	125.183	8.461	125.129	8.488	125.205	8.449	124.996	8.432	125.049
Ser 90	8.334	116.476	8.316	116.401	8.352	116.512	8.326	116.336		
Lys 91	8.401	123.496								
Gln 92	8.449	122.832								
Pro 93										
Ala 94	8.523	124.474	8.509	124.468	8.565	124.629	8.541	124.536	8.541	124.584
Ser 95	8.305	114.989	8.298	114.841	8.332	114.921	8.322	114.864		
Gln 96	8.435	123.173								
Pro 97										
Arg 98	8.562	121.731	8.554	121.685	8.594	121.759	8.581	121.694		
Gly 99	8.432	110.019	8.407	110.150	8.452	110.224	8.446	110.083		
Asp 100										

Pro 101										
Thr 102	8.302	113.133	8.342	113.127	8.396	113.167	8.384	113.142	8.381	113.211
Gly 103	8.179	111.020	8.174	111.086	8.214	111.190	8.198	111.081	8.195	111.156
Pro 104										
Lys 105	8.521	121.429								
Glu 106	8.414	121.467	8.345	121.561						
Ser 107	8.439	117.486	8.439	117.501	8.478	117.579	8.474	117.503		
Lys 108	8.371	123.195								
Lys 109										
Lys 110	8.391	122.477								
Val 111	8.266	122.203	8.275	122.187	8.323	122.348	8.308	122.288	8.304	122.349
Glu 112	8.524	124.938	8.555	125.005	8.606	125.309	8.594	125.190	8.592	125.242
Arg 113	8.441	122.217								
Glu 114	8.483	123.329								
Thr 115	8.223	114.896	8.198	114.418	8.228	114.356	8.209	114.212	8.203	114.269
Glu 116	8.487	123.069								
Thr 117	8.421	114.774	8.189	114.709	8.211	114.482	8.194	114.490	8.187	114.524
Asp 118	8.451	123.551								
Pro 119										
Val 120	8.242	119.508	8.288	119.512	8.348	119.582	8.339	119.570	8.337	119.658
Asp 121	8.129	126.111	8.019	127.662	8.016	128.604	7.995	128.676	7.991	128.797

Overlay of ^{1}H-^{15}N HSQC Spectrum of His-tagged Tat protein at different pHs

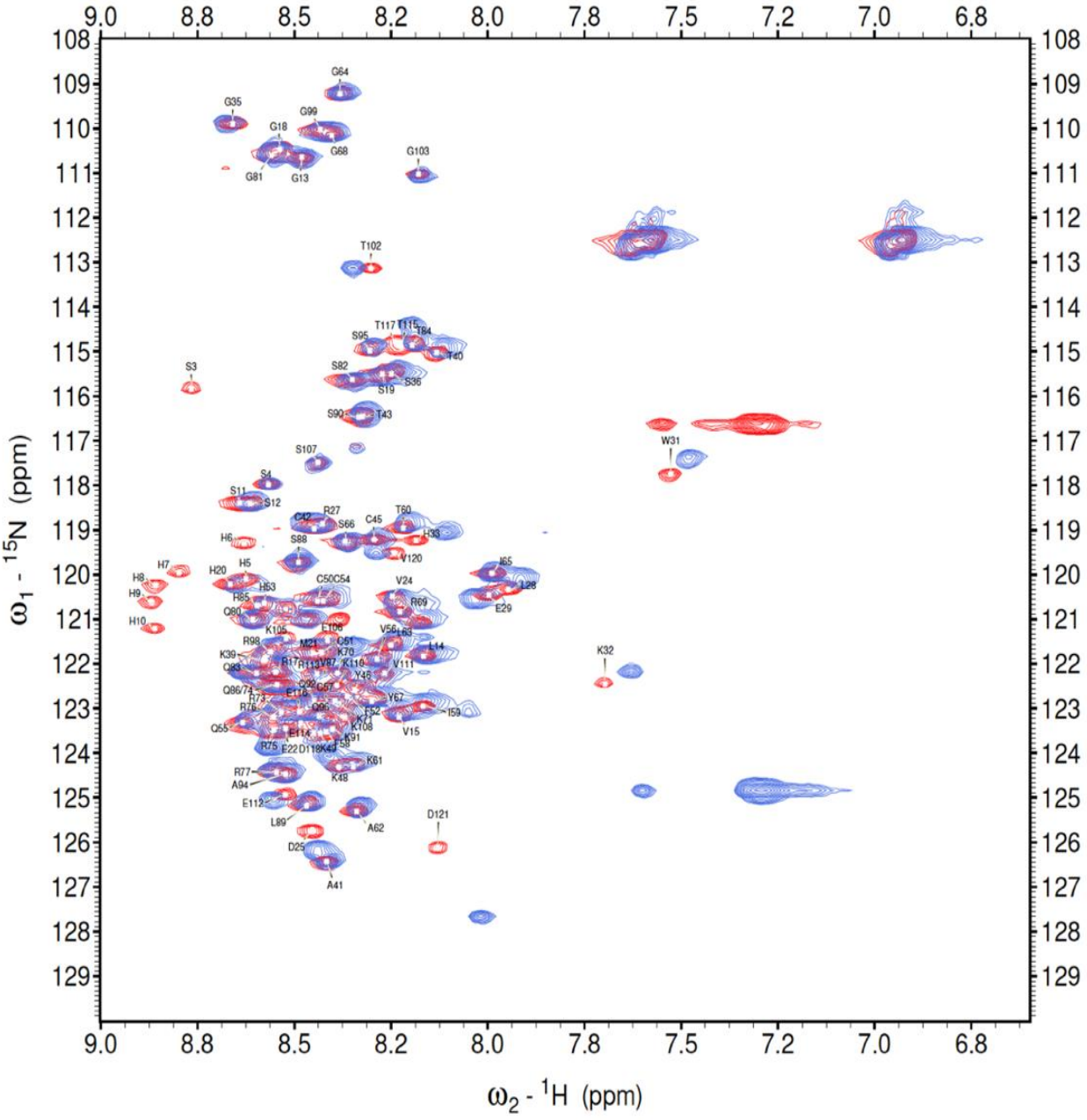


Figure 1. Overlay of ^{1}H-^{15}N HSQC spectrum of His-Tat at pH 4 (Red) and pH 5 (Royal blue) with 64 coadded transients.

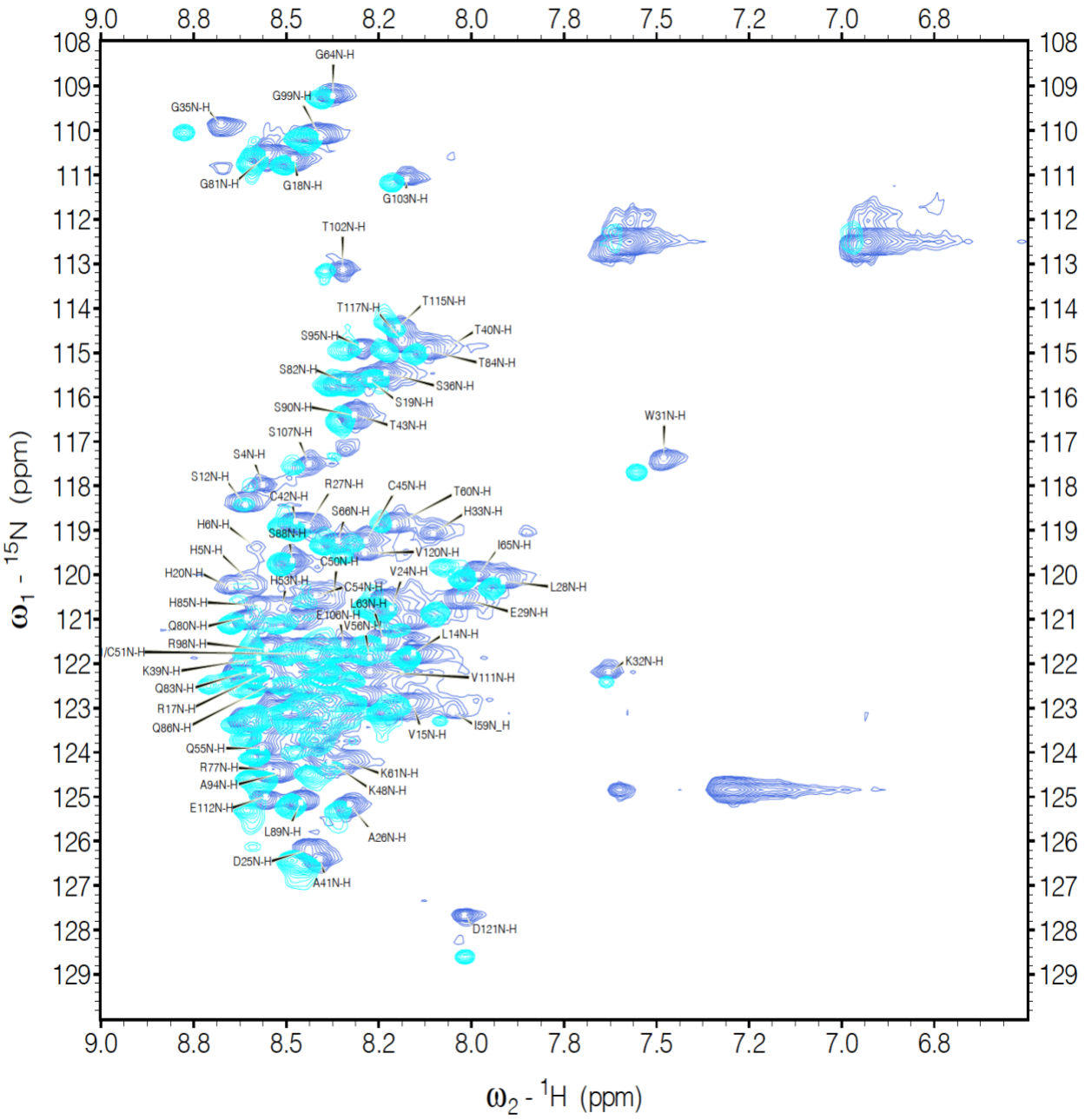


Figure 2. Overlay of ^{1}H-^{15}N HSQC spectrum of His-Tat at pH 5 (Royal blue) and 6 (Cyan) with 64 coadded transients.

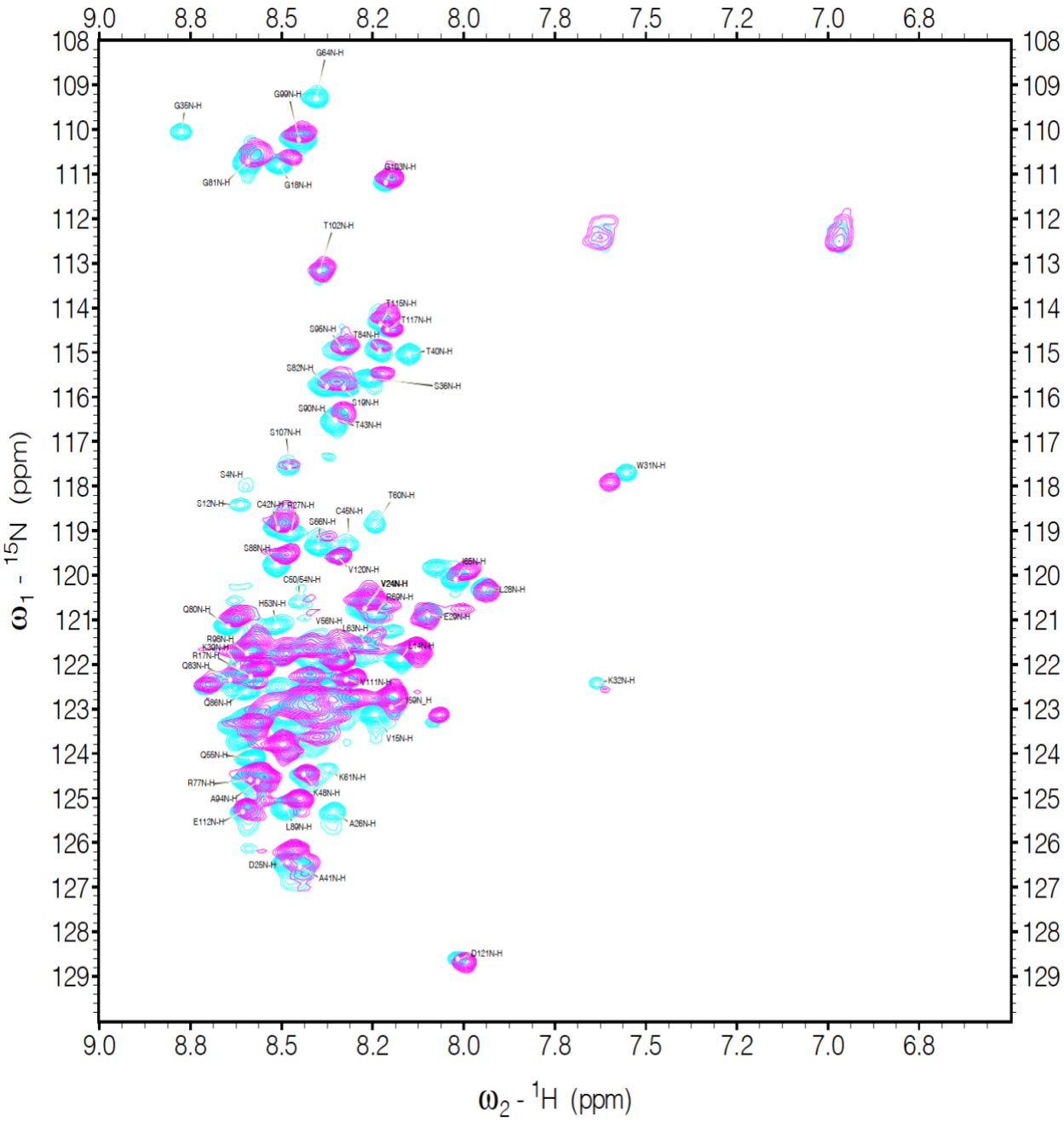


Figure 3. Overlay of ^{1}H-^{15}N HSQC spectrum of His-Tat at pH 6 (Cyan) and 6.5 (Magenta) with 64 coadded transients.

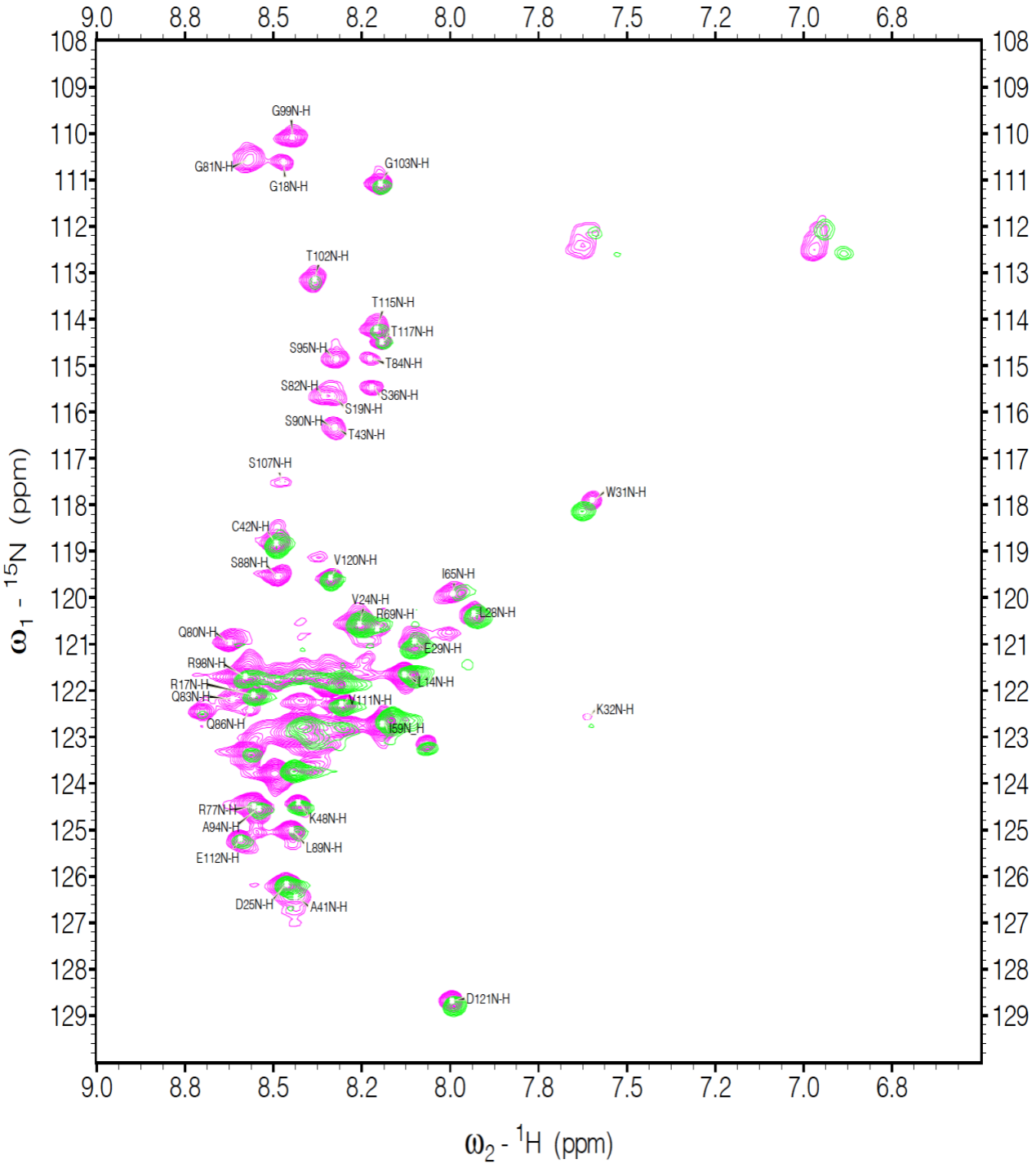


Figure 4. Overlay of ^{1}H-^{15}N HSQC spectrum of His-Tat at pH 6.5 (Magenta) and 6 (Cyan) with 64 coadded transients.

^{1}H-^{15}N HSQC of Supercharged Tat at different pHs

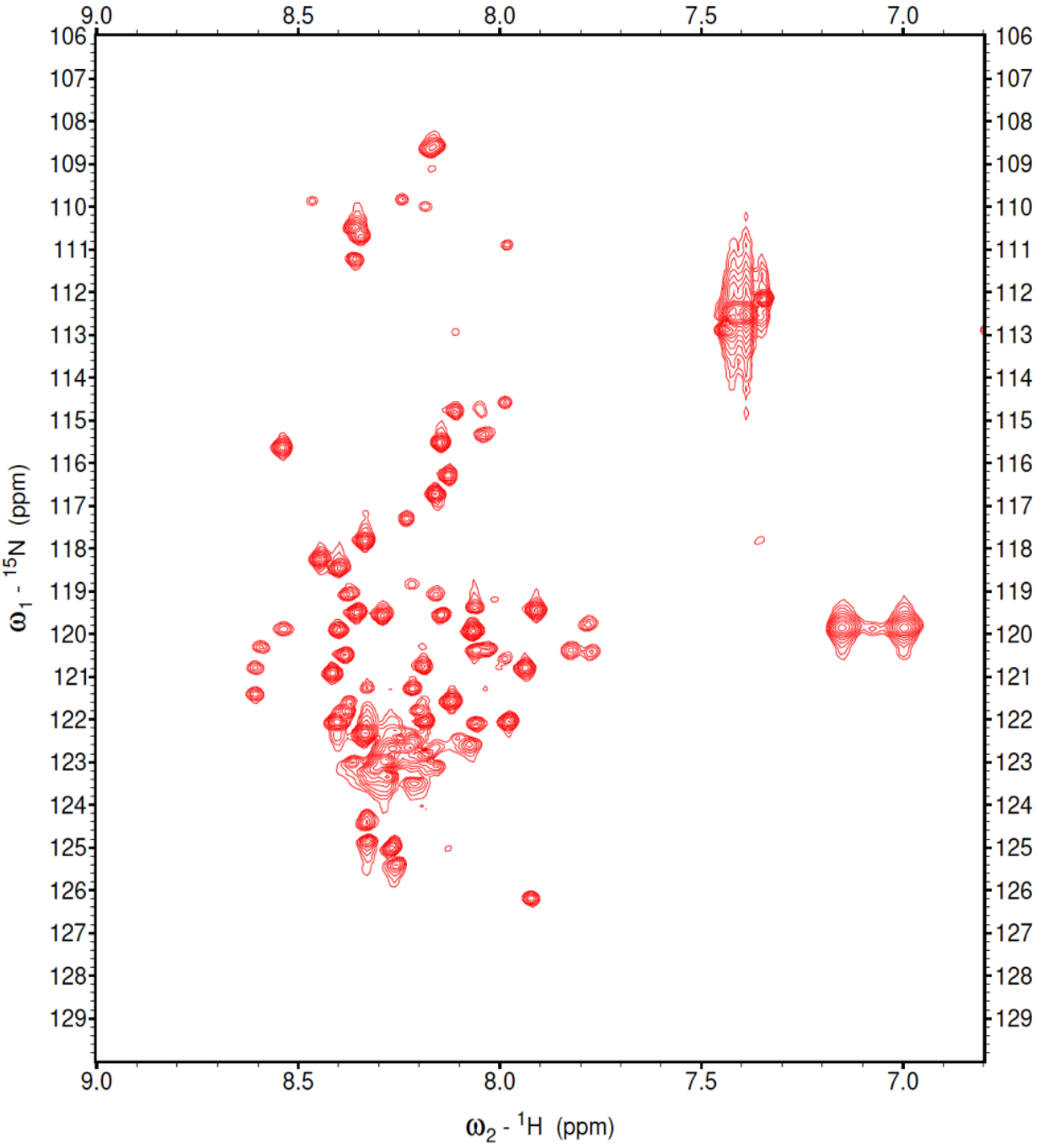


Figure 1. ^{1}H-^{15}N HSQC spectrum of ^{15}N-labelled 230 µM Supercharged-Tat-protein at pH 4 acquired with 90 coadded transients.

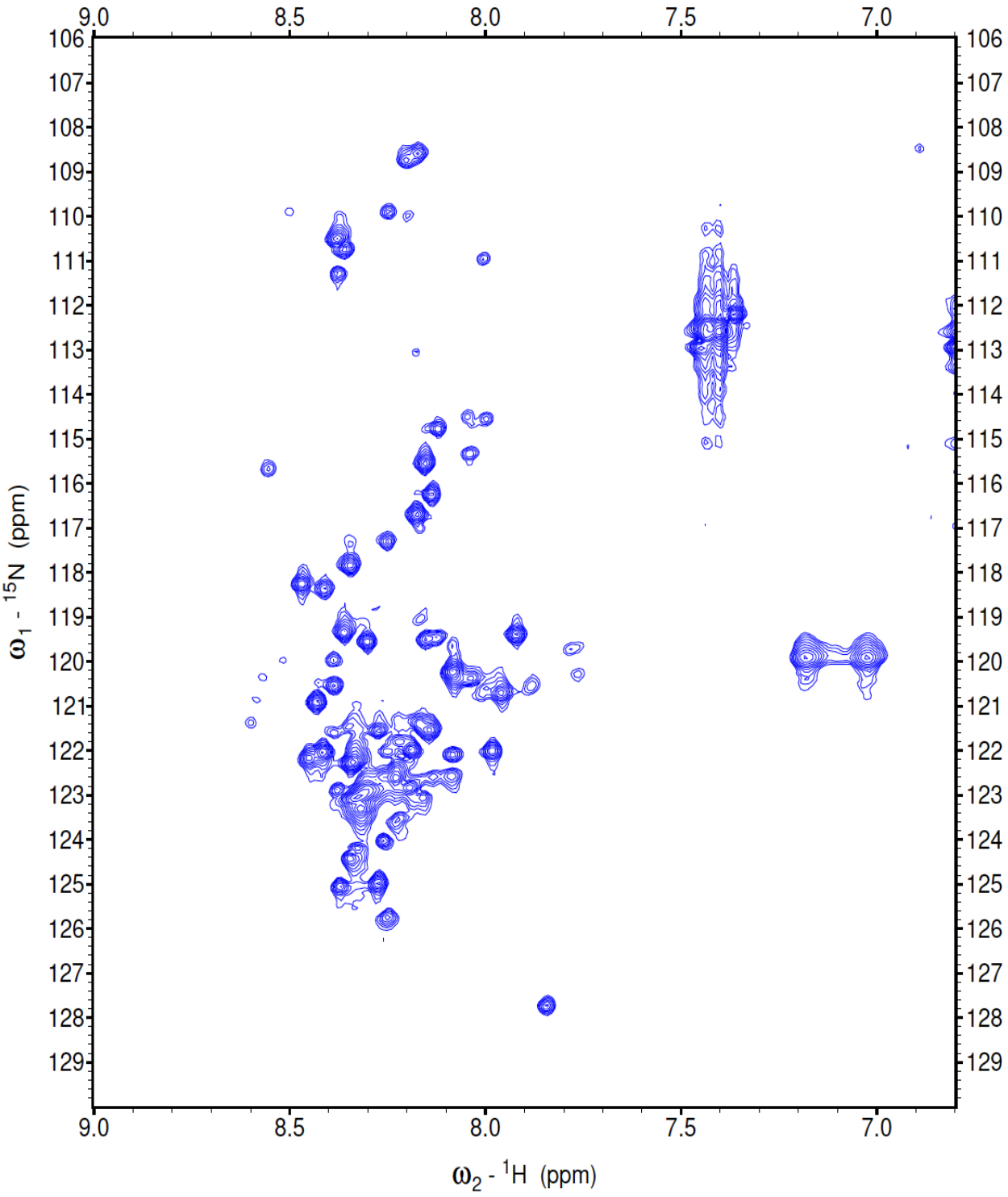


Figure 2. ^{1}H-^{15}N HSQC spectrum of ^{15}N-labelled 190 µM Supercharged Tat-protein at pH 5 acquired with 90 coadded transients.

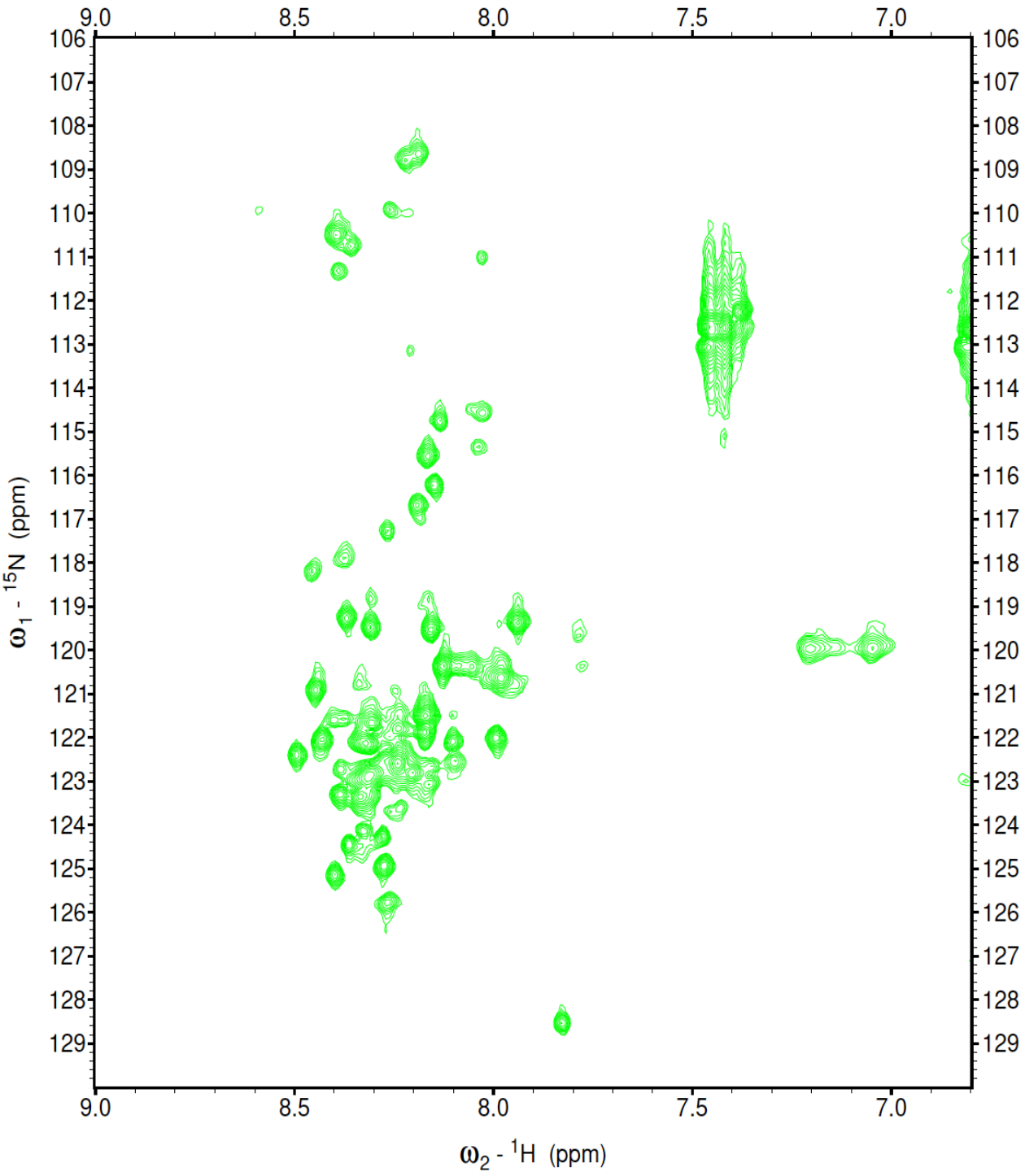


Figure 3. ^{1}H-^{15}N HSQC spectrum of ^{15}N-labelled 150 µM Supercharged Tat-protein at pH 6 acquired with 90 coadded transients.

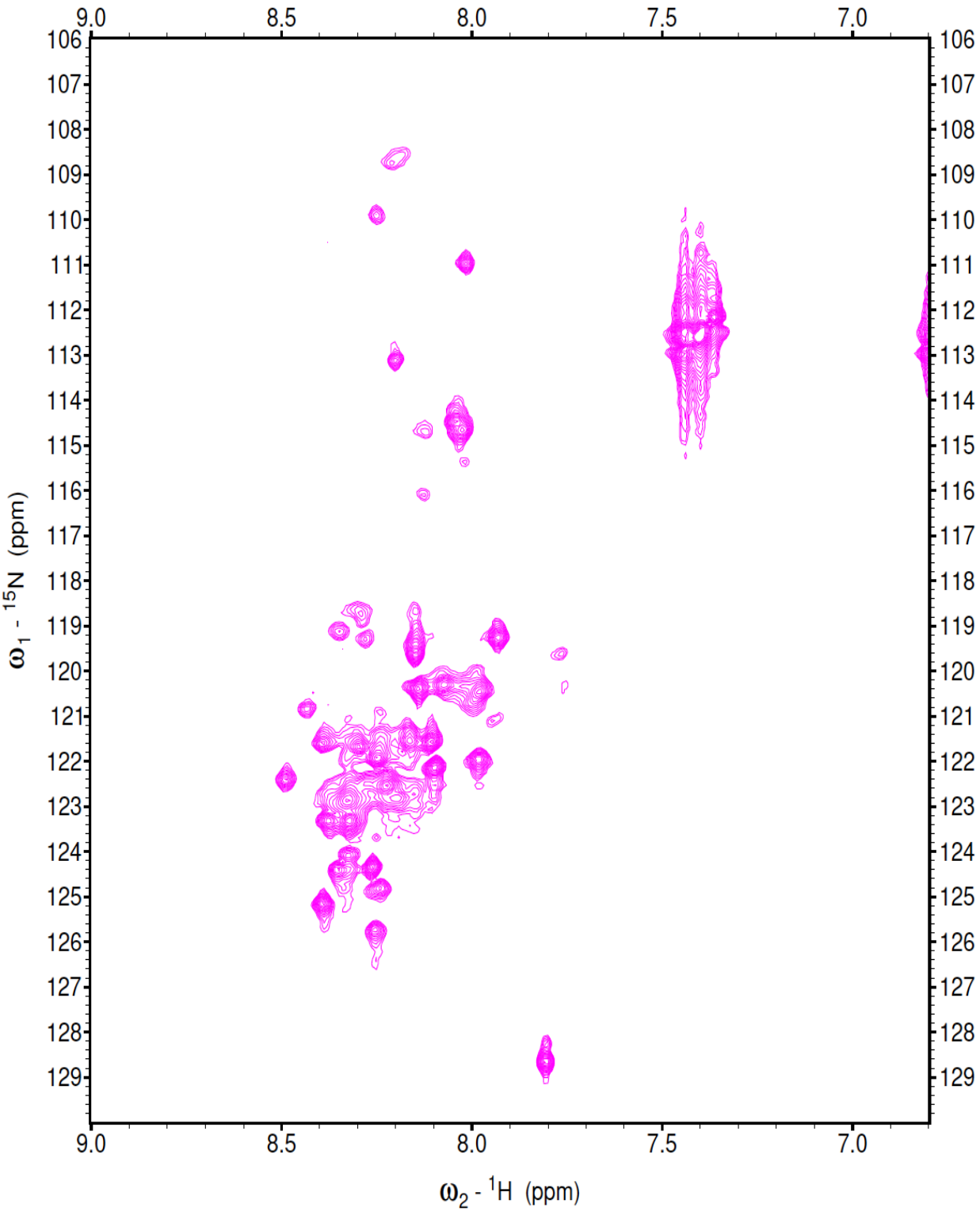


Figure 4. ^{1}H-^{15}N HSQC spectrum of ^{15}N-labelled 100 µM supercharged Tat-protein at pH 7 acquired with 90 coadded transients.

Amino acid composition of full-length Asp-Tat

Residue	3-letter code	1-letter code	Class	Number of residues	Count-%
Alanine	Ala	A	Hydrophobic	3	2.48
Arginine	Arg	R	Basic	10	8.26
Asparagine	Asn	N	Hydrophilic	1	0.83
Aspartic Acid	Asp	D	Acidic	11	9.09
Cysteine	Cys	C	Hydrophilic		
Glutamic Acid	Glu	E	Acidic	6	4.96
Glutamine	Gln	Q	Hydrophilic	8	6.61
Glycine	Gly	G	Hydrophobic	9	7.44
Histidine	His	H	Basic	10	8.26
Isoleucine	Ile	I	Hydrophobic	2	1.65
Leucine	Leu	L	Hydrophobic	4	3.31
Lysine	Lys	K	Basic	12	9.92
Methionine	Met	M	Hydrophobic	2	1.65
Phenylalanine	Phe	F	Hydrophobic	2	1.65
Proline	Pro	P	Hydrophobic	13	10.74
Serine	Ser	S	Hydrophilic	12	9.92
Threonine	Thr	T	Hydrophilic	7	5.79
Tryptophan	Trp	W	Hydrophobic	1	0.83
Tyrosine	Tyr	Y	Hydrophobic	2	1.65
Valine	Val	V	Hydrophobic	6	4.96