Backbone Dynamics of the Intrinsically Disordered HIV-1 Tat Protein

by

Vu To

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

DOCTOR OF PHILOSOPHY

Department of Chemistry
University of Manitoba
Winnipeg

Abstract

The type 1 Human Immunodeficiency Virus (HIV-1) transactivator of transcription (Tat) is a 101-residue protein that binds the viral mRNA, recruiting the cellular positive transcription elongation factor-b (P-TEFb), increasing the processivity of RNA polymerase II and leading to a significant increase of viral transcription. The full-length Tat protein (Tat_{101}) is encoded by two exons yielding the first 72 and the last 29 residues of the protein, respectively. The function and intrinsic disorder of the first 72 residues have been studied in great detail but relatively little is known about the structure and function of the second exon product despite its conserved expression in all lentiviruses.

My thesis, taking into account the biological importance of the second exon product of Tat, aims to study the impact of this region on the full-length protein in terms of disorder, dynamics and structural propensity. NMR spectroscopy has been used as the principle technique to study Tat_{101} protein in a fully reduced state. In order to study the Tat_{101} protein by NMR, multiple labeling strategies were applied and the labeled protein was expressed and purified in high yield. Backbone resonance assignment, chemical shift analysis, structure propensity analysis, fast (ps-ns) and slow (ms) timescale dynamics and hydrogen exchange studies confirm the intrinsically disordered nature of the second exon product and full-length Tat. The NMR results also revealed a propensity to alpha-helix in the acidic and cysteine-rich regions, and the propensity to beta-sheet/extended conformation in the core region and two other conserved motif regions. Reduced spectral density mapping and model-free analysis show that the fast internal motion on the ps-ns timescale dominates the relaxation, and that Tat_{101} has no slow motion / conformational exchange on the ms timescale. The hydrogen exchange measurements yield protection factors below 1, which can be explained by the high local concentration of hydroxyl ions in the vicinity of this

highly basic protein, leading to a higher local pH compared with the bulk solvent. Classical molecular dynamics simulations were used as a complementary technique to verify the NMR results, and more importantly, to sample the protein conformers that are invisible to NMR due to ensemble averaging. The two 100 ns trajectories from the simulations of Tat’s first exon product are dominated by non-structural elements such as coils, turns, and bends. The order parameters derived from the simulations are below 0.8 and in agreement with the NMR results, confirming the dynamic flexibility of the protein. The combination of NMR dynamics and simulation results indicate that some regions of the protein likely bind partners through a conformational selection mechanism while other parts of protein bind their targets through an induced – fit mechanism.

Contents

List of Figures

List of Tables

List of Abbreviations

AIDS	acquired immunodeficiency syndrome
ART	antiretroviral therapy
AXIN	axis inhibitor scaffold protein
AZT	azidothymidine
BBB	blood-brain-barrier
BMRB	Biological Magnetic Resonance Data Bank
CA	capsid protein
CBP	CREB binding protein
CCR5	chemokine receptor
CD	Circular dichroism
CD55	complement decay-accelerating factor
CdK	cyclin-dependent kinase
CDK9	cyclin-dependent kinase 9
cDNA	complementary DNA
cIAP1	cellular Inhibitor of Apoptosis 1 protein
CNS	central nervous system
CPMG	Carr-Purcell-Meiboom-Gill
CREB	cAMP response element binding protein
CSA	chemical shift anisotropy
CSI	chemical shift index
CypA	cyclophilin

DNA	deoxyribonucleic acid
DNase	deoxyribonuclease
drkN SH3	N-terminal SH3 domain from the adapter protein drk
DSIF	5,6-dichloro-1-β-D-ribofuranosylbenzimidazole sensitivity-inducing factor
Dsp	desiccation stress protein
DSS	2,2-dimethyl-2-silapentane-5-sulfonate
E. coli	*Escherichia coli*
E1A	early region 1A
Env	envelope protein
FlgM	flagellar anti-σ factor
Gag	polyprotein
gp120	glycoprotein 120
gp41	glycoprotein 41
GPU	graphical processing unit
HAdV	human adenovirus
HAART	highly active antiretroviral therapy
HAT	histone acetyl transferase
HDM2	human homolog of mouse double minute 2
His-tag	hexahistidine affinity tag
HIV	human immunodeficiency virus
HLA-DR1	Human leukocyte antigen DR1
HSQC	heteronuclear single quantum coherence
HX	hydrogen exchange

ICAM-1	Intercellular Adhesion Molecule 1
IDP	intrinsically disordered protein
IN	intergrase
INEPT	insensitive nuclei enhanced by polarization transfer
IPTG	isopropyl-β-D-thiogalactopyranoside
KID	kinase-inducible domain
LB	lysogeny broth
LEA	late embryogenesis abundant protein
MA	Matrix protein
MAP2	microtubule-associated protein 2
MD	molecular dynamics
MHC	major histocompatibility complex
MoREs	molecular recognition elements
MoRFs	molecular recognition features
MYPT-1	Myosin phosphatase targeting subunit 1
mRNA	messenger RNA
N-TEF	negative transcription elongation factor
NC	nucleocapsid protein
NCBI	National Center for Biotechnology Information
NDPK	human nucleoside diphosphate kinase
Nef	negative factor
NELF	negative elongation factor
NMR	Nuclear magnetic resonance

NOE	nuclear Overhauser effect
OD	optical density
ORD	optical rotary dispersion
OTU	Ovarian tumor protease
P-TEFb	positive transcription elongation factor-b
p130	tumor suppressor protein
PCAF	p300/CBP-associated factor
PDB	Protein Data Bank
PES	potential energy surface
PP1	protein phosphatase 1
ppm	parts-per-million
PR	protease
PRE	paramagnetic relaxation enhancement
PTM	post-translational modifications
RB	retinoblastoma protein
RDC	residual dipolar coupling
RF	radio frequency
RMSD	root-mean-square deviation
RNA	ribonucleic acid
RNAPII	RNA polymerase II
RNase	ribonuclease
RT	reverse transcriptase
Rev	Regulator of expression of virion proteins

SAXS	small angle X-ray scattering
SDS-PAGE	sodium dodecyl sulfate polyacrylamide gel electrophoresis
SEC	super-elongation complex
smFRET	single-molecule Förster Resonance Energy Transfer
smFS	single-molecule Force Spectroscopy
SSP	secondary structure propensity
TAD	transactivator domain
Tat	Transactivator of transcription
TAR	trans-activation response
TAZ	Transcription adaptor putative zinc finger domain
TCEP	tris (2-carboxyethyl) phosphine
TOCSY	Total Correlation Spectroscopy
UV	ultraviolet
UNAIDS	Joint United Nations Program on HIV and AIDS
Vif	viral infectivity factor
Vpr	lentivirus protein R

Acknowledgments

This work was supported by grants from the Natural Sciences and Engineering Research Council of Canada, the University of Manitoba and the Vietnam International Education Development organization.

The long Ph.D journey with lots of learning, working, failing, and succeeding is getting to the end, and a time for acknowledgments has come.

First, I would like to thank my supervisor, Prof. Joe O'Neil, who kindly accepted me and let me work in his laboratory, although I came with no background in the field of biochemistry. I really learnt a lot here and I am grateful for this chance. I also thank him very much for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my Ph.D. study.

Besides my supervisor, I would like to thank the other members of my committee, Dr. Jörg Stetefeld, Dr. Sean McKenna, and Dr. Peter Pelka, for the assistance they provided at all levels of the research project.

I would like to thank the following people for making this research and thesis possible:

- Dr. Kirk Marat, University of Manitoba, for assistance and training at the NMR facility at the University of Manitoba.
- Dr. Leo Spyracopoulos, University of Alberta, for his generous distribution of the *Mathematica*© notebooks for analysis of NMR relaxation data.

- Dr. Shaheen Shojania, University of British Columbia, for his assistance with the SOLEXSY experiment.
- Dr. Pascal Mercier, University of Alberta, for his assistance with the NMR data acquisition at NANUC.
- Mary Hernando, Sagar Saran and other students who have worked in Joe O'Neil research group for the great time in the lab.

I would like to thank my parents for all the encouragement and support that you gave me from half way around the world. I also thank my fiancé, Hoa Dang, for your endless love. I would also like to thank my 'Winnipeg family': Tam Thai, Anh Tuan Nguyen, Loan Le, Dung Pham, Ha Vu, Tuan Pham, Ngan Dang, Tuan Nguyen, Chi Doan, Khang Nguyen, "Alex" Huy Nguyen and others for your love, laughter and friendship which have kept me smiling and inspired. You are and always will be my family.

Chapter 1

Introduction

1.1. Intrinsically disordered proteins.

Proteins are the major component of the living cell, serving as vital parts in the cellular machinery. Protein dysfunction may lead to diseases. One of the best understood functions of proteins is catalysis (*i.e.*, enzymatic activity). In 1894, Emil Fisher observed the hydrolysis of β-glycosidic but not α-glycosidic bonds by enzymes and proposed that substrates need a correct shape to fit to the enzyme active site like a “lock and key” in order to exert a chemical effect on each other[1, 2]. The lock and key model attracted much attention in the early days of protein science for describing the mechanism through which proteins function. It still dominates our understanding of enzyme catalysis today.

The structure – function point of view was further backed up by later studies. In 1936, a survey on the structure of native, denatured and coagulated states of proteins by Mirsky and Pauling showed that upon denaturation, certain highly specific properties of native protein were lost, including enzymatic activity, immunological and physical properties[3]. In 1953, Anfinsen *et al.* successfully re-natured RNase A from the completely unfolded state and observed the restoration of the enzymatic activity[4]. This study clearly demonstrated the structure – function relationship. Early X-ray crystal studies of enzymes with their competitive inhibitors such as lysozyme in complex with N-acetyl-glucosamine and its dimer[5], or RNase-S with uridine 2’, (3’)-

phosphate in high concentration ammonium sulfate solution[6], shed light on structurally how the enzyme bound substrates. Thousands of deposited structures in the Protein Data Bank (PDB)[7] throughout the last few decades again strongly support the necessity of a 3D structure for explaining functionality.

In the 1950s and 1960s, optical rotary dispersion (ORD) was increasingly used to study protein structure. Even though proteins were believed to need a stable structure to function, there were studies that showed some proteins had “unusual structural properties”. Those included casein, a milk protein, which was determined to lack a 3D structure by single wave length ORD in 1952[8]; and phosvitin, an egg protein, for which the lack of a 3D structure was shown by multi-wave length ORD by Jirgensons in 1958[9]. Notably, Jirgensons also proposed a protein classification that included a category named “disordered”[10]. The observations on phosvitin were also confirmed by circular dichroism (CD) spectroscopy (Grizzuti and Perlmann, 1970[11]) and by Nuclear Magnetic Resonance (NMR) spectroscopy (Vogel, 1983[12]). Additionally, despite the explanatory power of the crystal structure which strengthened the structure – function paradigm in protein science, there remained a question about the missing electron density in most of the deposited structures in the PDB. In 1971, two missing electron density regions in staphylococcus nuclease were proposed to be “disordered”[13]. Such local regions of missing electron density in protein crystal structures were explained to arise either from static disorder, in which the missing region adopts multiple fixed positions, or from dynamic disorder, in which the missing region remains mobile even in the confines of the crystal lattice.

Interest in proteins that lack a stable structure began to increase in the late 1990s. Many more of this type of protein were discovered and they were named differently: floppy, pliable, rheomorphic[14], flexible[15], mobile[16], natively denatured[17]/unfolded[18], intrinsically unstructured[19]/

disordered[20]/denatured[17], vulnerable[21], chameleon[22], protein cloud[23], dancing protein[24], *etc.* From 2005, the term "intrinsically disordered" has become the predominant term thanks to a concerted effort in the field to use a consistent terminology.

An intrinsically disordered protein (IDP) is defined as having little or no ordered secondary structure or tertiary structure to emphasize its difference from other proteins when being studied by tools of structural biology. It has become increasingly common to define IDP as existing as a dynamic ensemble, within which atom positions and backbone Ramachandran angles exhibit extreme temporal fluctuations without specific equilibrium values[25].

1.1.1. Amino acid composition of IDPs

The amino acid composition of disordered proteins is different from that of ordered proteins. Figure 1.1 shows the result when I compared 2 data sets:

- I constructed the Disprot data set by including amino acid sequences of 1534 disordered protein regions reported by the Disprot data base[26] of IDP/IDR, updated on 26/09/2016;
- I constructed the Structured data set by including the amino acid sequences of 4091 proteins from the PDB, each of which has more than 20 amino acids where the structure was determined by X-Ray crystallography, and contains no ligand or modification.

All duplicated entries from multimeric proteins were removed. The amino acid composition is calculated for each IDP/IDR/protein, then is averaged for each data set. Comparison between the two data sets is expressed as (Disprot – Structure)/Structure. Negative peaks indicate depleted amino acids in IDPs compared to ordered proteins while positive peaks show the reverse.

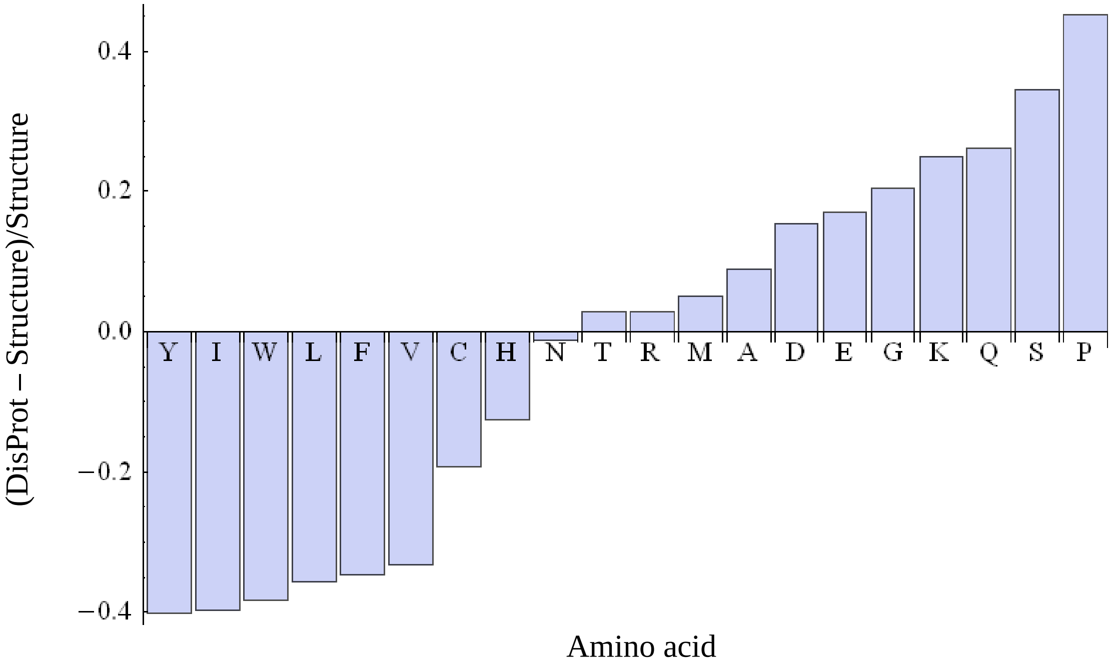


Figure 1.1. Relative amino-acid composition of IDPs compared with folded proteins.

IDPs/IDRs are shown to be enriched in P, S, Q, K, G, E, and D, which are charged, polar and structure-breaking residues; and they are depleted in hydrophobic and aromatic residues, including Y, I, W, L, F, V, and C residues which are termed "order-promoting amino acids"[27].

The low mean hydrophobicity and high net charge in the amino acid sequences of IDPs contribute to charge – charge repulsion within the proteins and limit the driving force for structural compaction. Uversky's plot[28] in Figure 1.2 reports the relationship between mean net charge and the mean hydrophobicity of two sets of proteins; IDPs are in red and ordered proteins are in blue. The IDPs are clustered on the left side of the separation line having high net charge and low mean hydrophobicity compared to ordered proteins located on the right side.

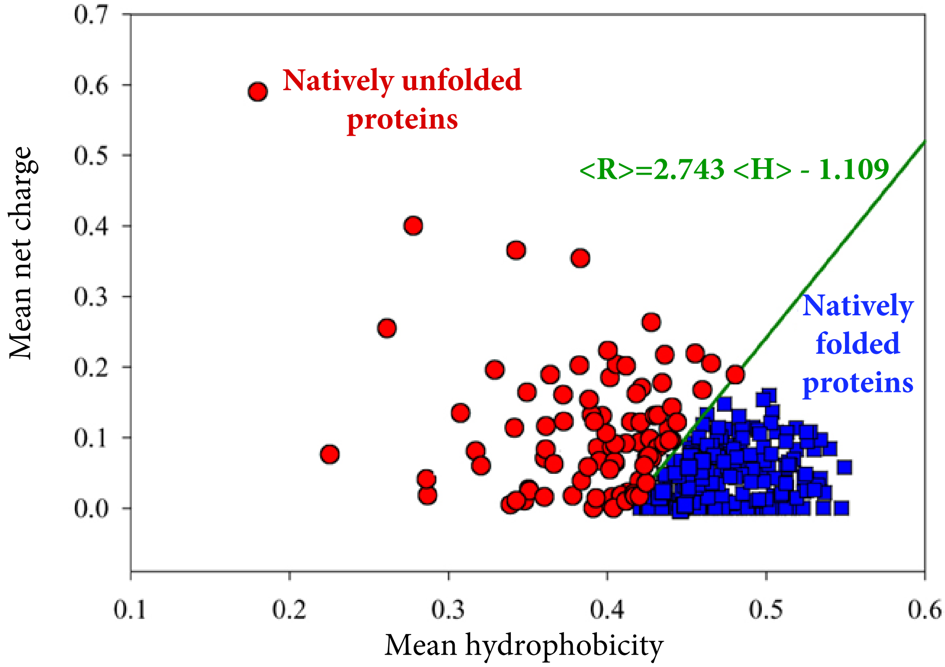


Figure 1.2. Charge-hydropathy plot of protein disorder. Mean net charge <R> vs mean hydrophobicity <H> is plotted for disordered (red circles) and ordered (blue squares) proteins. The two sets are separated by a straight line ⟨R⟩ = 2.743<H> – 1.109 shown as a green line. Adapted with permission from reference [25]*. Copyright © 2014 American Chemical Society.*

Low complexity in terms of protein sequence is also a distinctive feature of IDPs as they normally have low compositional variability and relatively high repetition of amino acids within a segment of protein, compared to folded proteins. And certain types of charged and proline – rich repetitive regions were also found in a statistical study by Lise *et al.*[29].

1.1.2. Disordered protein prediction.

Prediction of protein secondary structure has long been a goal for protein scientists. The above mentioned characteristics of IDP sequence are widely exploited in predicting protein disorder. FoldIndex[30], which uses Uversky's plot[28] was among the earliest methods. Other methods, *i.e.* PreLink[31] (no longer available) or GlobPlot[32], also used the same approach to predict IDPs/IDRs.

Along with the development of algorithms, more and more data became available and databases of experimentally characterized IDPs were established. Presently, besides the information about disorder which can be extracted from missing electron density regions in the PDB, there are 3 large databases of IDP/IDR available: Disprot[26], IDEAL[33], and MobiDB[34]. Making good use of the databases, a group of disorder prediction programs using machine – learning algorithms are trained on the data from either regions of missing electron density from the PDB, or from IDP/IDR databases. PONDR[35], DisEMBL[36], DISOPRED2[37] and AUCpreD[38] are among the ones using this approach.

Other prediction programs rely on an assumption that proteins are disordered because their protein sequences do not allow them to make enough inter-residue contacts to compensate for the reduction in entropy during folding. The FoldUnfold[39] program predicts protein disorder using the expected average number of contacts per residue from the amino acid sequence. First, the mean packing density for each of the 20 amino acids in a globular state is calculated from the database of 5829 folded protein structures as an average number of close residues (within the given distance). Next, an average of the mean packing densities of the residues within a window is calculated, and then assigned to the central residue of this window. As the window slides along the sequence, a profile of expected packing density is constructed for protein disorder prediction.

Another prediction program, the IUPred[40], instead of counting the number of contacts, evaluates the energy from the inter-residue interactions, taking the contact energy from globular proteins as a reference, to predict protein disorder. The reference contact energy can be calculated using a coarse-grained approach, as a sum of pairwise interactions between amino acid pairs within a distance cutoff.

The above-mentioned predictors have their own strengths and weaknesses for using different approaches. No standard definition of disorder or standard method of assignment for disordered regions has been set across all methods. Different data training sets, containing varying proportions of amino acid composition and/or different distributions of disorder lengths, are utilized. The methods for self-assessing the accuracy and reliability of the predictions also differ between prediction servers. Difference in the performance of prediction programs was demonstrated in the study by Atkins *et al.*[41], where they submitted the cardiac Muscle LIM protein (MLP) to various servers. All predictors returned different results, with some returning vastly different predictions.

Therefore, using a combination of methods (programs and servers), with different attributes for defining disorder, is the recommended approach, in order to obtain the most accurate prediction in the absence of direct experimental evidence. The Database of Disordered Protein Prediction (D^2P^2)[42] and several meta-predictors, such as metaPrDOS[43] or MFDp[44], allow the combination and comparison of multiple disorder predictor outputs. The combined result is more objective and reliable, thanks to a broader coverage of protein disorder properties from multiple predictors chosen for their high predictive efficacy instead of their availability or their ease of use.

1.1.3. Functions of disordered proteins.

IDPs have been found to have multiple functions. In 2002, when doing a literature search for the functions of proteins which contain disordered regions of at least 30 consecutive residues under native conditions, IDPs were suggested to have 28 separate functions[45]. These functions were then classified into 6 broader groups: entropic chain, effectors, scavengers, assemblers, display sites and chaperones. The physical characteristics of IDPs facilitate their function, including the presence of small recognition elements that fold on partner association (see 1.1.4); conserved sequence motifs to mediate binding interactions; a degree of flexibility, which enable them to interact with different targets; accessible sites for post-translational modification; and the ability to bind partners with high specificity and low affinity, leading to fast complex association/dissociation[46]. In the next section, I will briefly summarize the activities of each of these classes and give some examples of each.

1.1.3.1. Entropic chains

The function of proteins in this class of IDP comes directly from the flexibility and plasticity of the backbone. They serve as linkers, spacers, entropic clocks, entropic springs or entropic bristles.

- Linkers are found in a majority of protein crystal structures in the PDB often in missing electron density regions. They serve to regulate the distance between domains and enable conformational freedom in orientational searches.

- Microtubule – associated protein 2 (MAP2) is an example of an IDP that functions as a spacer. MAP2 consists of a C-terminal microtubule – binding domain, and an N-terminal projection domain which extends away from the surface of the microtubule[47]. The N-terminus of MAP2 is highly negatively charged and unstructured, as proven by circular dichroism, fluorescence spectroscopy, sedimentation equilibrium[48], ^{1}H[49] and ^{13}C-NMR spectroscopy[50]. When microtubules with MAPs are sedimented, they form a gel where the microtubules are widely spaced due to the long-range repulsion forces from the projection domains[47]. Spacing of microtubules by MAPs may contribute to cellular mechanics by resisting mechanical compression and/or maintaining cell shape. The unstructured projection domain would also occupy a very large volume, compared to a folded protein of the same length, and tend to exclude other macromolecules in a size-dependent manner (the larger the molecule the more it would be excluded), leading to a significant increase in the effective concentration of excluded molecules and thereby influencing their intracellular biochemical activity.

- An example of the entropic clock is the IDR in the Voltage-gated potassium channel (Figure 1.3). This tetrameric integral membrane protein consists of four domains: the two domains forming the potassium channel are linked to the positively charged inactivation domain (the ball domain) *via* a disordered linker (the chain domain). The whole system cycles among three main stages: closed (A), open (B and C), and inactivate (D). At the closed stage, prior to membrane depolarization, the pore is sealed, and a positive charge on the cytoplasmic side excludes binding of the inactivation domain. After membrane depolarization, the pore is open and the negative charge on the channel domain facilitates interaction with the ball domain that is positively charged. The potassium channel is inactivated when the ball domain occludes the pore. Transition from open-stage C to inactive-stage D is by a random walk of the ball domain toward its binding position, and the timescale of the process is determined by the length and flexibility of the disordered chain. Modification of the length of the chain domain shows that a shorter chain speeds up inactivation and longer chains slow the process down[51].

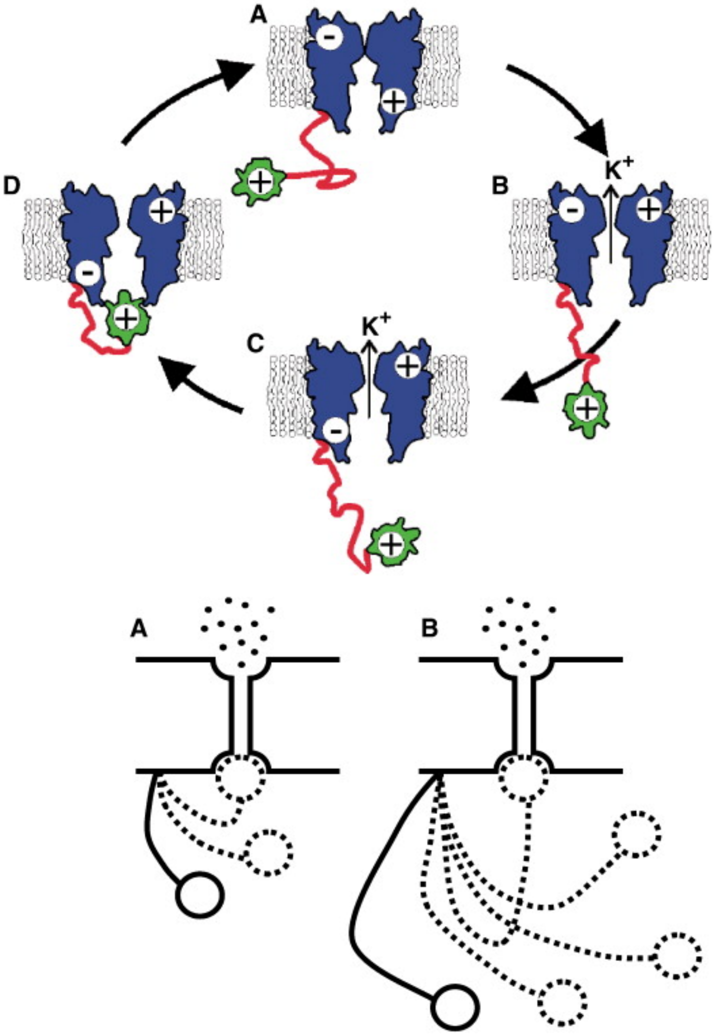


Figure 1.3. Example of an entropic clock. (a). Simplified model of the Voltage-gated K^+ ion channel, (b). “Ball and chain” model, showing the dependence of channel inactivation timing on the chain length. Adapted with permission from reference [52]. Copyright © 2010 Elsevier B.V.

1.1.3.2. Effectors

This group consists of IDPs that, upon binding, alter the action of binding partners. The earliest studied IDPs in this class are the two cyclin-dependent kinase (CdK) inhibitors p21$^{\text{Cip1}}$ [53] and p27$^{\text{Kip1}}$ [54], and the C-terminal half of the inhibitor of the sigma28 transcription factor in bacteria, FlgM[55]. p21$^{\text{Cip1}}$ and p27$^{\text{Kip1}}$ bind to the cyclin A – CdK2 complex *via* homologous sequences of approximately 60 amino acids within the N-termini of the two proteins. Binding inhibits the catalytic activity of the complex, inducing cell-cycle arrest in response to anti-proliferative signals. FlgM protein binds to and inhibits the sigma28 transcription factor which is necessary for the formation of flagella[56], propelling devices for bacteria to propel themselves through watery environments. Studies of these three proteins also uncovered the changes in their conformation, leading to important concepts in the field of IDP's such as folding induced upon biding, pre-existence of binding-competent secondary structure, and structural adaptability[25].

1.1.3.3. Scavengers

IDPs in this group bind to small ligands, *e.g.* ions or organic compounds, for disposal or storage and later release depending on the needs of the organism. The extended protein conformation is used mainly to maximize the ligand-binding capacity of the proteins. Proline–rich glycoproteins in saliva, which neutralize polyphenolic plant compounds such as tannins[57], and Desiccation stress protein (Dsp) 16, which retains water in plants[58], are examples of this group.

1.1.3.4. Assemblers

IDPs in this group assemble, stabilize and regulate large complexes, having a central role in regulation of signaling pathways and crucial cellular processes, including the regulation of transcription, translation and the cell cycle. An example for this group is the early region 1A (E1A) protein from the human adenovirus (HAdV)[59]. E1A protein forms a ternary complex with the retinoblastoma protein (pRb, also called RB) and the TAZ2 domain of transcriptional coactivator CREB binding protein (CBP) or its paralogue p300 (Figure 1.4). This complex brings the CBP histone acetyl transferase (HAT domain) and the cell cycle regulatory protein RB into proximity, promoting acetylation and degradation of RB, forcing S-phase entry and uncontrolled proliferation[60]. Another example is the axis inhibitor scaffold protein (AXIN). The long intrinsically disordered region of AXIN binds β-catenin, casein kinase Iα, and glycogen synthetase kinase 3β. The assembly of all four proteins accelerates interactions between them by raising their local concentrations and leads to the efficient phosphorylation and subsequent destruction of β-catenin[61].

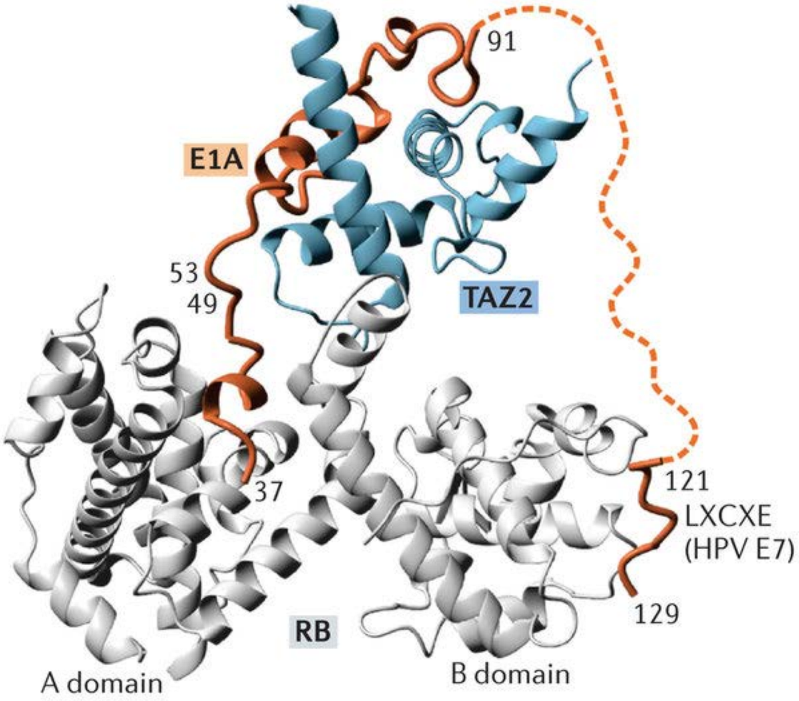


Figure 1.4. Structural model of the ternary RB-E1A-TAZ2 complex. The flexible linker between residues 83 and 120 of E1A is indicated schematically as a dotted line. Reprinted with permission from reference [46]*. Copyright © 2014 Macmillan Publishers Limited.*

1.1.3.5. Chaperones

There are chaperone proteins reported to be fully disordered including α-synuclein[62], β-casein[63] and late embryogenesis abundant (LEA) protein[64, 65]. Chaperone activity is also found in local disordered regions of other proteins, *e.g.* in the case of the small heat-shock protein α-crystallin[66, 67]. Disordered regions make up over one-half of the sequences of RNA chaperones and over one-third of the sequences of protein chaperones[68, 69]. Protein disorder seems well suited for chaperone function for the following reasons, despite a scarcity of mechanistic evidence. Firstly, disordered regions can structurally adapt to many different binding partners, allowing chaperones to bind a wide range of proteins. Secondly, disordered regions also enable fast macromolecular interactions. This is because of their highly dynamic nature enabling many different conformations to be rapidly sampled, resulting in a higher probability of sampling the specific conformation that exists in the stable interaction complex and thus increasing the association rate of the interaction[70]. Thirdly, binding of a misfolded protein may induce folding in the disordered chaperone protein, and the entropic penalty from the folding counterbalances the enthalpy gain from the binding. This allows weak and transient but still specific interactions, in contrast to folded proteins where interactions are expected to have high specificity along with high binding affinity. The uncoupling of specificity from affinity allows chaperone proteins to quickly cycle through the binding and releasing of partners. Finally, the highly hydrophilic character of IDPs provides a solubilizing effect, preventing the formation of toxic aggregates[70]. One theory of IDP chaperoning activity is the "entropy transfer" model whereby disorder in the chaperone is transferred to a misfolded protein helping it to unfold and search for the proper conformation[68].

1.1.3.6. Display Sites

IDPs in this group are subjected to post-translational modifications (PTM), which serve as regulatory signals for cellular processes. Computational and experimental evidence also supports the abundance of PTM such as phosphorylation[71] and ubiquitination[72] in IDPs. An advantage of IDPs is that their structural flexibility allows full accessibility of the entire protein to modifying enzymes[73].

The disordered N-terminal transactivation domain (TAD) of p53 is an example of an IDP in this group. It has nine phosphorylation sites and simultaneous phosphorylation of two to four sites within the p53 TAD has been detected in cell extracts[74]. In unstressed cells, p53 forms a ternary complex with CBP/p300 and HDM2 (the human homolog of mouse double minute 2). When DNA is damaged, p53 is phosphorylated, which modifies the ability of p53 to interact with its cellular partners[75]. Increasing the number of phosphorylation events on the p53 NTAD also increases its affinity for the Transcription Adaptor putative Zinc finger domains (TAZ1 and TAZ2), and KIX domains of CBP/p300[76]. In the case of p53, PTMs play an important role in fine-tuning the binding affinity for its various cellular partners.

1.1.4. Coupling of folding and binding of disordered proteins, fuzzy complexes and the fly-casting mechanism.

Most IDP functions, except for the entropic chain, are related to molecular recognition and disorder-to-order transitions, or induced folding upon binding to some extent. Unbound IDPs in cells do not adopt stable structures; instead, they are best characterized as ensembles of dynamic structures that interconvert freely[73, 77-79]. When interacting with partners, most coupled folding and binding events involve only relatively short motifs contained within longer disordered sequences[80]. These motifs, termed molecular recognition elements/features (MoREs/MoRFs), usually have fewer than 30 residues, and are possibly identified by bioinformatics analysis of protein sequence[81]. They can fold into helix, β-strand or irregular structure on binding to a target (Figure 1.5).

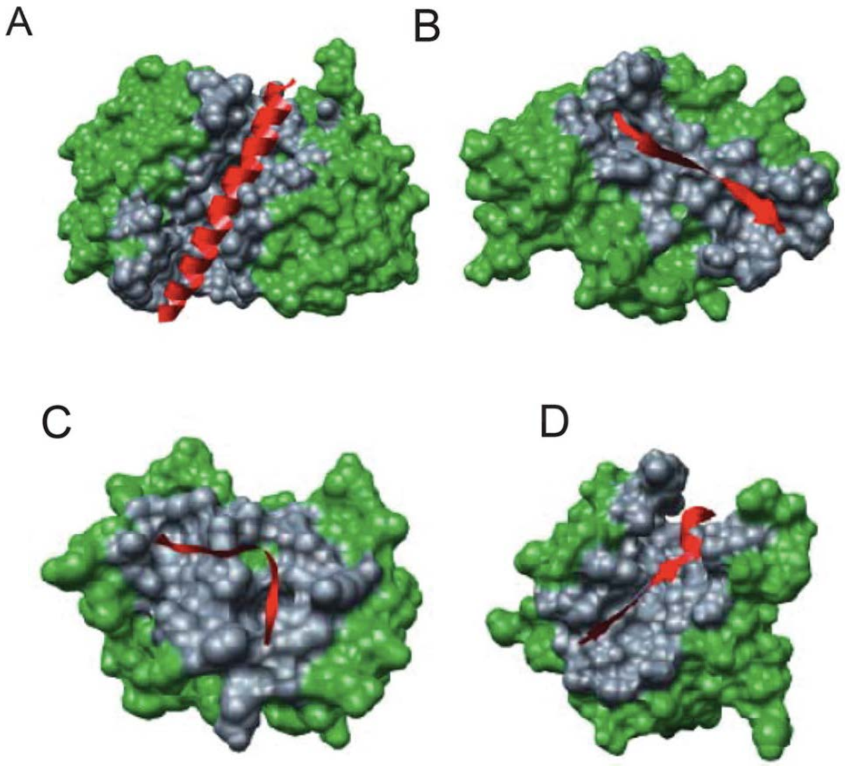


Figure 1.5. Examples of structurally divergent MoREs. MoREs (red ribbons) and partners (green surfaces) are shown. (A) An α-MoRF, proteinase inhibitor IA3, bound to proteinase A (PDB entry 1DP5). (B) A β-MoRF, viral protein pVIc, bound to human adenovirus 2 proteinase (PDB entry 1AVP). (C) An i-MoRF, amphiphysin, bound to α-adaptin C (PDB entry 1KY7). (D) A complex-MoRF, β-amyloid precursor protein (βAPP), bound to the PTB domain of the neuron specific protein X11 (PDB entry 1X11). Partner interfaces (gray surfaces) are also indicated. Reprinted with permission from reference [25]*. Copyright 2014 American Chemical Society.*

Some IDPs also have longer binding regions, rather more of a domain than a motif, with their lengths exceeding 20 – 30 residues[82]. These regions have typical properties of domains for being independent, in term of structure and function, of the rest of the protein molecule, for being recognized by homology due to sequences that are conserved through evolution, and for possessing at least one specific function[83, 84].

By having the binding regions one after another, an IDP can bind to many partners[52]. One IDP binding region may also change its shape and thereby bind to different partners[85-87]. Figure 1.6 shows a fraction of the proteins that are known to interact with the disordered region of the p53 protein. The disordered N-terminus of p53 may bind to more than 40 different partners, and the C-terminus, which is disordered as well, binds to an even larger number of partners. As multiple partners can bind to the same region of an IDP, the binding competition among them certainly has biological consequences[88].

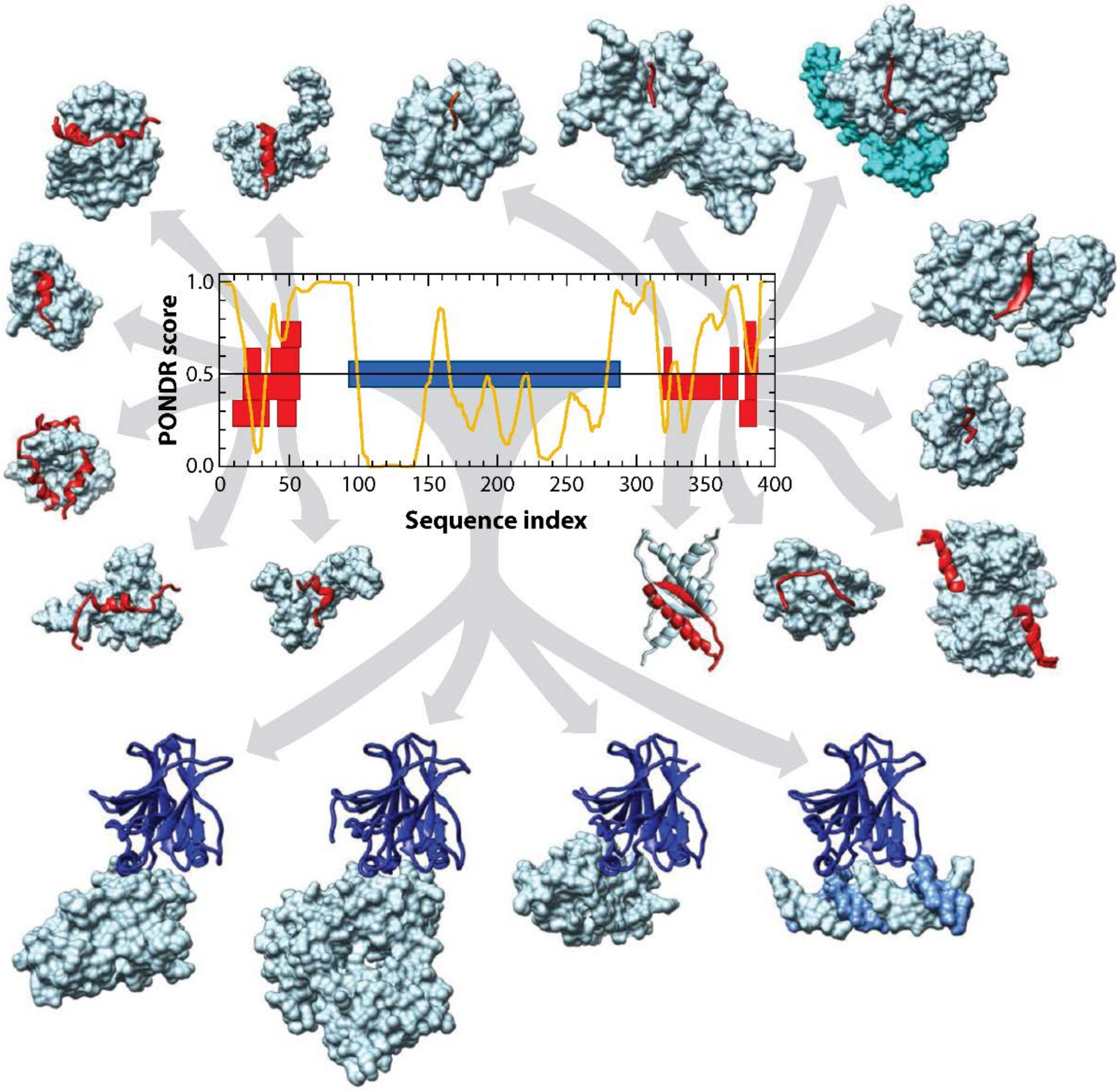


Figure 1.6. Intrinsic disorder and molecular interactions of the tumor suppressor p53. The graph in the center shows the PONDR VL-XT prediction of intrinsic disorder for p53. Values above or below the 0.5 threshold indicate predictions of disorder or order, respectively. The prediction indicates that the N and C termini are largely disordered, whereas the central DNA binding domain is ordered. Also shown are the structures of several discrete regions of p53 (ribbons) that have been determined in complexes with partners (surfaces); the corresponding horizontal bars indicate the region of p53 that participates in each structure. Five partners of the

N terminus (clockwise from the lower right) are high-mobility group protein B1 (PDB entry 2LY4), Taz2 domain of p300 (PDB entry 2K8F), nuclear coactivator–binding domain of p300 (PDB entry 2L14), MDM2 (PDB entry 1YCR), the N-terminus of replication protein A (PDB entry 2B3G), and the PH domain of RNA polymerase II transcription factor B subunit 1 (PDB entry 2GS0). Eight partners of the C-terminus (clockwise from the upper left) are the histone acetyltransferase domain of Tetrahymena general control nonderepressor 5 (PDB entry 1Q2D), SET9 (PDB entry 1XQH), CDK2/cyclin A (PDB entry 1H26), Sir2 (PDB entry 1MA3), the bromodomain of CBP (PDB entry 1JSP), S100B(ββ) (PDB entry 1DT7), the Tudor2 domain of PHF20 (PDB entry 2LDM), and p53 oligomerization domain (PDB entry 3SAK). Four partners of the central DNA binding domain (left to right) are 53BP2 (PDB entry 1YCS), the large T antigen of simian virus 40 (PDB 2H1L), the BRCT domain of 53BP1 (PDB entry 1GZH), and DNA (PDB entry 1TSR). Reprinted with permission from reference [88]*. Copyright © 2014 by Annual Reviews.*

The mechanism through which IDPs fold upon binding is of significant interest. Two extreme cases have been proposed and evidence has been obtained in support of each:

- ***Induced folding mechanism***: Folding of an IDP takes place after the association with its target, and the contact with the binding partner provides a major driving force for the folding event. An example for this mechanism is the binding between the disordered C-terminal domain of the measles virus nucleoprotein (N_{TAIL}) and the X domain (XD) of the viral phosphoprotein (Figure 1.7). Initially, N_{TAIL} forms a weak encounter complex in a disordered conformation with XD. Subsequently, N_{TAIL} goes through a folding step, gaining α-helicity within a short region of the domain, to form the final bound complex[89].

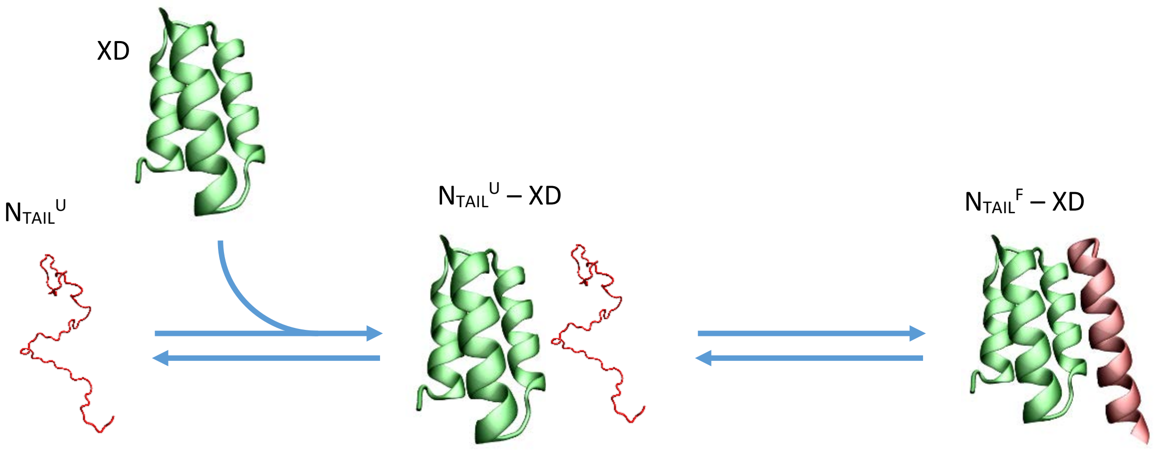


Figure 1.7. Induced fit mechanism for the binding between the disordered C-terminal domain of the measles virus nucleoprotein (N_{TAIL}^{U}) and the X domain (XD) of the viral phosphoprotein. The binding event goes through an intermediate step of forming an encounter complex N_{TAIL}^{u} – XD where N_{TAIL} remains disordered. The final bound form of the complex is N_{TAIL}^{F} – XD, with a region within N_{TAIL} adopting an α-helical conformation. Structures of XD and the complex were generated using VMD[90], using coordinates from the PDB entry 1T6O[89].

- ***Conformational selection mechanism***: The bound conformation is populated in the conformational ensemble of the unbound IDP, and this preexistent conformation dominates the binding process. An example for this case is the binding of the transactivation domain of the transcription factor c-Myb, a disordered protein, to the KIX domain of the general transcriptional coactivator CREB-binding protein[91] (Figure 1.8)

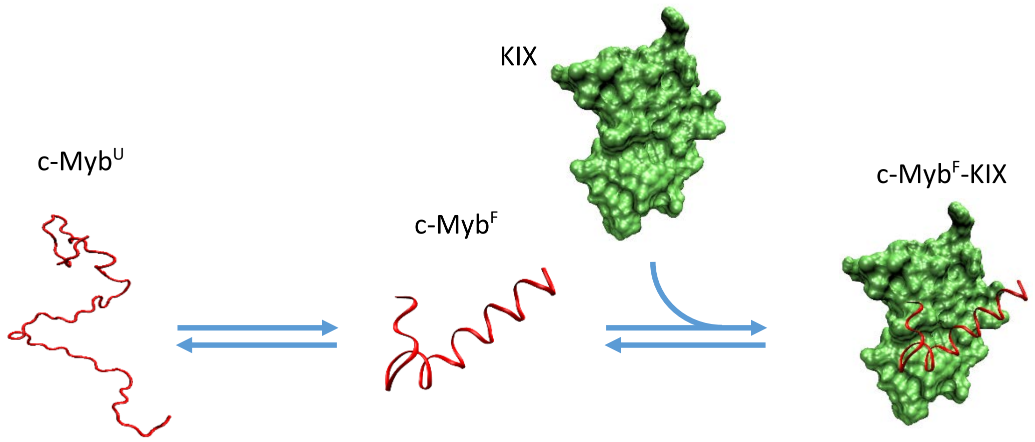


Figure 1.8. Conformational selection mechanism for the binding between c-Myb and KIX. Structures of KIX and c-Myb^F-KIX complex were generated using VMD[90], using coordinates from PDB entry 1SB0[92].

It is more likely that folding upon binding of an IDP to its target follows a mechanism that is a combination of the two extreme cases just described[93, 94]. Both mechanisms work synergistically in the binding event; and their contribution to the overall binding mechanism depends on the required rate of binding, the concentration and the local flexibility of the IDP, the degree of binding degeneracy (the ability to bind to multiple targets) and the type of disorder-to-order transition[94]. The first step of this synergistic mechanism is a non-specific 'reeling' of the IDP by the target molecule *via* a mechanism called 'fly-casting'[95]. The 'fly-casting' mechanism proposed that the disorder of the IDP enhances its capture radius, improving its ability to search for a partner[95, 96]. Once the IDP is close enough to its target, a specific encounter is facilitated by MoRFs/MoREs. Conformational selection here comes into play by the selection of a specific conformational state of the MoRF/MoRE among the ones existing in the free IDP. The different capture radii in the fly-casting mechanism may assist the conformational selection of

MoRFs/MoREs by filtering out irrelevant conformations. In the last step, formation of the final complex is then largely dominated by the induced-folding mechanism. The driving force for this last step is provided by both a favorable intermolecular interaction energy and a considerable gain in solvent entropy from the release of solvation water molecules due to the folding and binding of MoRFs/MoREs[25, 94]. An example of this synergetic mechanism is the binding between the KIX domain of CREB binding protein and the disordered phosphorylated kinase inducible domain (pKID) of CREB[97].

Even though most IDPs will fold upon binding, there are cases where the IDP remains disordered even after binding to its target[80], forming an ensemble of "fuzzy complexes"[98]. The fuzzy regions establish alternative contacts between specific partners. Most fuzzy complexes that have been experimentally characterized are involved in gene-expression regulation, signal transduction and cell-cycle regulation. They also exist in viral protein complexes, in cytoskeleton structures and a few metabolic enzymes[98].

1.1.5. Characterization of disordered proteins.

Early characterization of IDPs was done by low-resolution techniques such as ORD and CD. These techniques cannot provide a description of individual disordered regions, but only the structural properties of the whole protein. When used in combination with intrinsic viscosity[9] or size exclusion chromatography[99], ORD and CD are effectively methods to confirm the non-globular, extended structure[100] of IDPs. When using CD to study protein secondary structure, measurements in the far – UV region are used. α-helices in proteins are characterized by large positive bands at 193 nm and negative bands at 208 and 222 nm; β-sheets yield a positive band at

193 nm and a negative band at 218 nm; and random coils exhibit a negative band at 195 nm, and very low positive ellipticity above 210 nm (Figure 1.9).

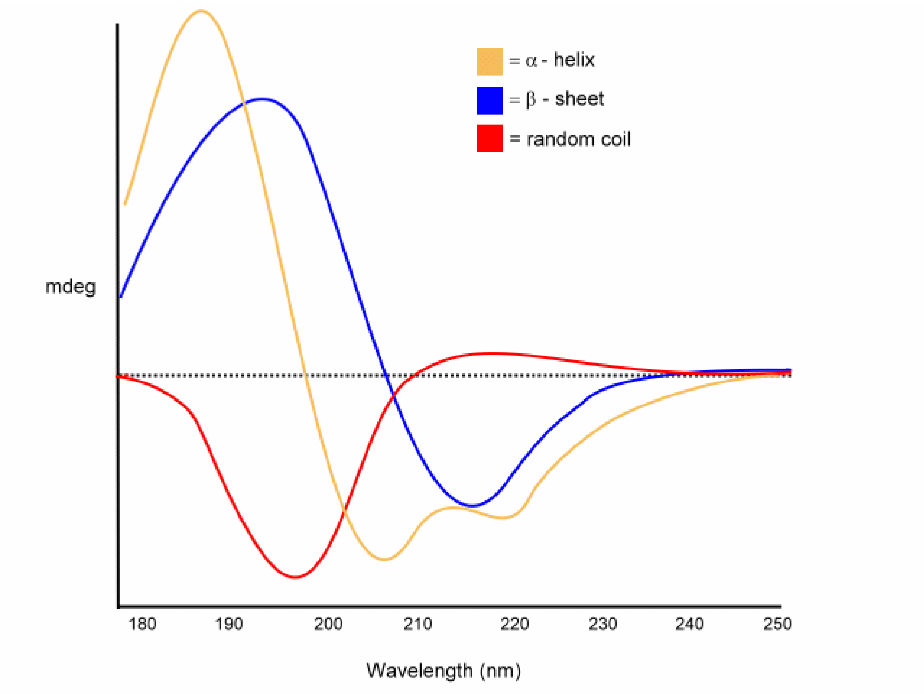


Figure 1.9. Circular dichroism spectra of pure secondary structures. Reprinted with permission from reference [101]*. Copyright © Frontiers in Bioscience, 1995.*

Disordered regions are also detected by the absence of electron density in *single-crystal X-ray diffraction* patterns. Because of their flexibility, such regions of the protein fail to scatter X-rays coherently due to differences in the atomic orientation among those sections of proteins in the crystal lattice. However, the conclusion of disorder needs to be verified because, in certain cases, a well-folded domain of a protein may be highly dynamic. The mobility causes inhomogeneity in the crystal lattice in a similar way that a disordered region does, leading to the absence of electron density for the whole structured domain. Finally, the missing electron density could be the result of a crystal defect or proteolysis during purification[102].

Small angle X-ray scattering (SAXS) is another method to characterize disordered proteins and disordered regions of structured proteins in solution. A solution of particles (IDP) is placed in a quartz capillary and illuminated by a monochromatic X-ray beam. The intensity of the scattered X-rays is recorded at small angles by a detector. The scattering pattern of the pure solvent is also collected and subtracted from the scattering of the sample, leaving only the signal from the particles of interest. The resultant scattering pattern is related to the sizes and shapes of proteins and complexes, ranging from a few kDa to GDa. A Kratky plot, defined in the legend of Figure 1.10 below, derived from SAXS data is used to qualitatively identify disordered states and distinguish them from globular proteins. The Kratky plot of a typical globular protein is bell-shaped, whereas the plot of a disordered protein will have no clear maximum[103, 104]. The radius of gyration is an important quantitative parameter that can be derived from the SAXS data, providing information on the average size of the particles. It can be compared to theoretical or experimental values published for globular and unfolded proteins of the same number of residues to verify the disorder of the sample[104].

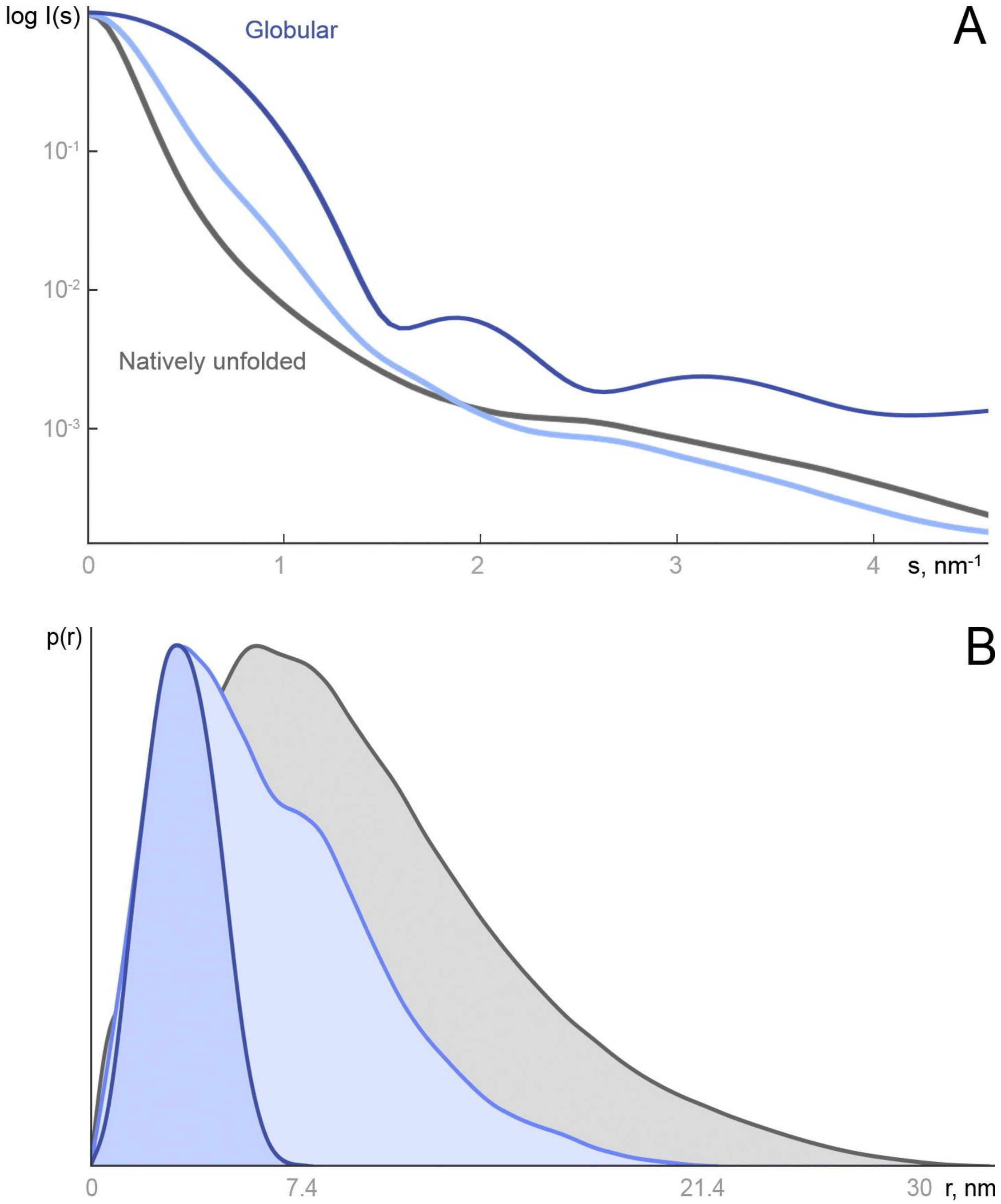


Figure 1.10. SAXS data simulated for three 60 kDa proteins: globular (dark blue), 50% unfolded (light blue) and fully disordered (gray). (A) Logarithmic plot of the scattering intensity I(s) (arbitrary units) vs. s (nm^{-1}), the momentum transfer defined as $s = [4\pi \sin(\theta)]/\lambda$*, where* λ *is the X-ray wavelength and* 2θ *is the scattering angle. (B) Distance distribution functions p(r) (arbitrary units) vs. r (nm).*

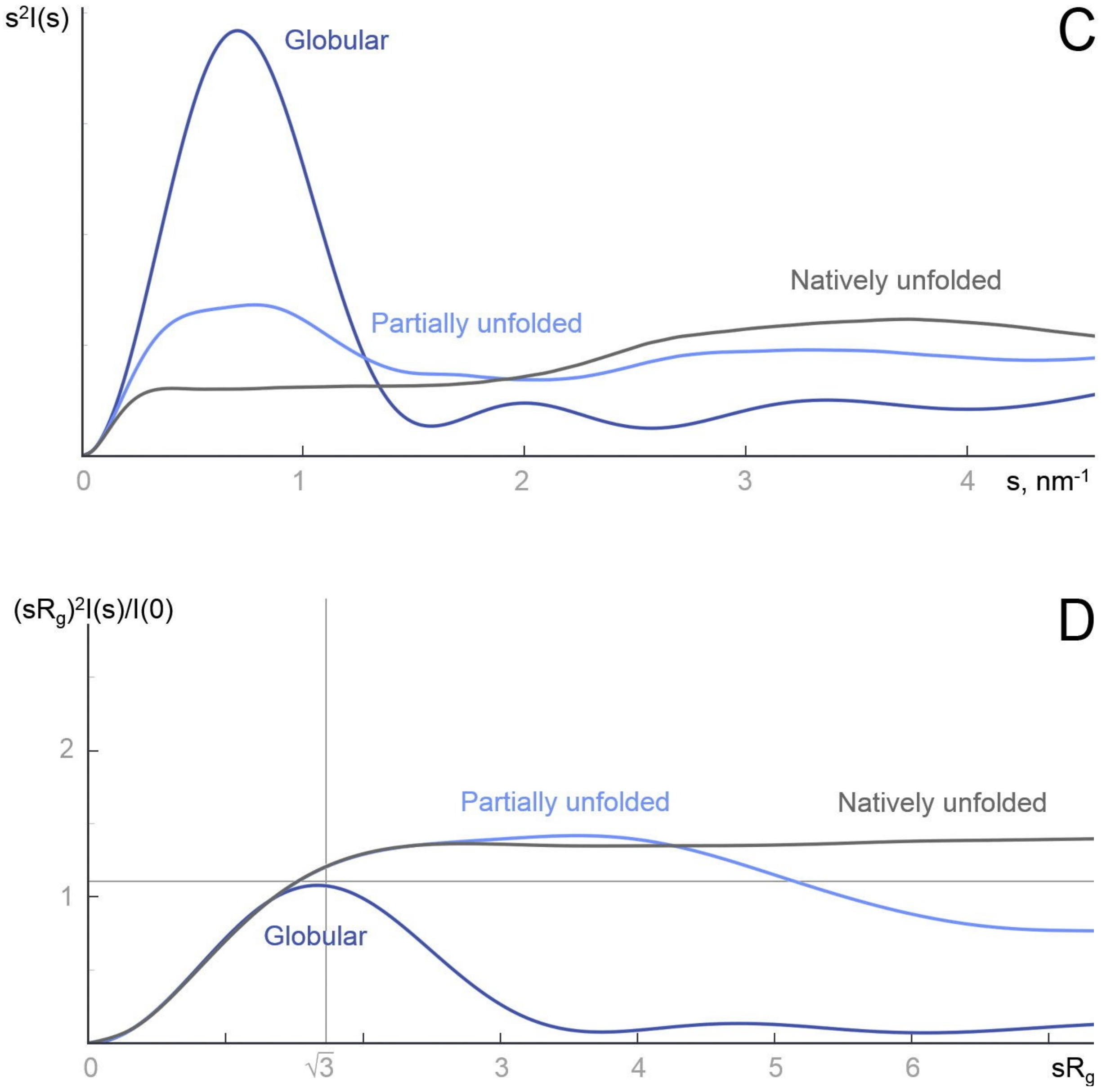


Figure 1.10. (Continued) SAXS data simulated for three 60 kDa proteins: globular (dark blue), 50% unfolded (light blue) and fully disordered (gray). (C) Kratky plot. (D) Normalized Kratky plot. Reprinted with permission from reference [105].

Single-molecule Förster Resonance Energy Transfer (smFRET) has recently emerged as an additional powerful method to study IDPs. smFRET measures the nonradiative energy transfer between fluorescence donor and acceptor chromophores[106], providing information about intramolecular distance, the distance distribution between the chromophores and the underlying dynamics on length scales of 2 – 10 nm and timescales from nanoseconds to days. An advantage of smFRET is that because the signal is recorded on single molecules, structural and dynamic heterogeneity can be resolved, which is often impossible to resolve by ensemble-averaged methods. From smFRET data, information about folding intermediates and pathways can be extracted, as for example in the study by Ferreon *et al.*[107], on the coupling between folding and ligand binding of α-synuclein. In this study, the folding landscape of α-synuclein was mapped as a function of sodium dodecyl sulfate (SDS) concentration (Figure 1.11). At very low SDS concentrations, α-synuclein molecules bound to a single SDS molecule, and transitioned from a disordered (U-state) to a folded conformation (I-state). Increasing the SDS concentration resulted in the binding of additional SDS monomers, and α-synuclein adopted a more extended, folded conformation (F-state).

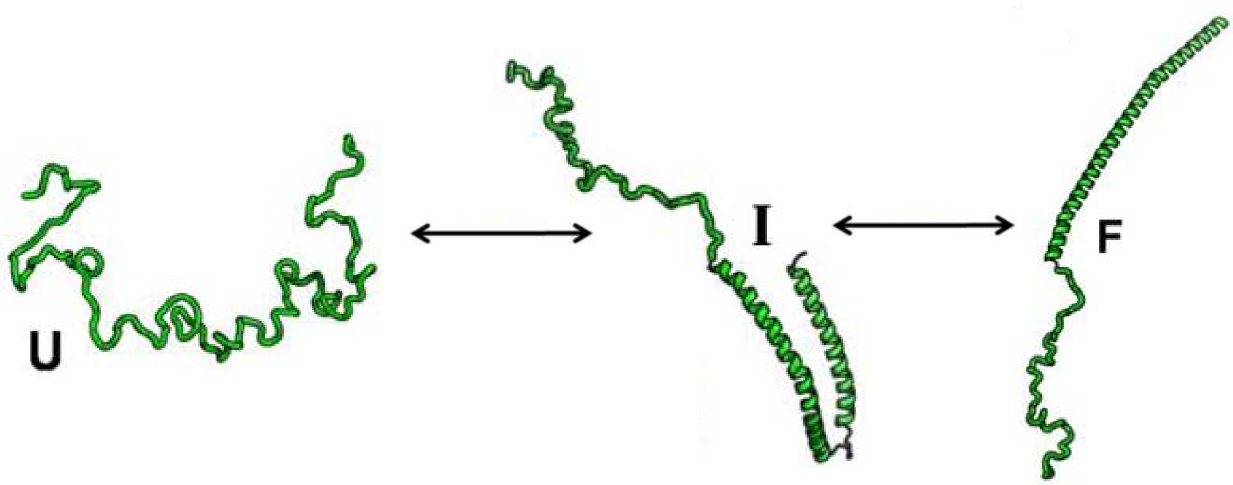


Figure 1.11. The conformational states of α-synuclein. The unfolded state (U) is largely disordered. The U-state transitions to an intermediate state (I) upon binding of SDS monomers. The I-state is also known as the hairpin or horseshoe conformational state. The folded state (F) is populated upon additional SDS monomer binding. Adapted with permission from reference [108].

In some cases where IDPs are amyloidogenic, high protein concentration samples often lead to aggregation, so the low concentration required for smFRET allows the feasibility of the experiment. Besides smFRET, other single-molecule methods, such as *single-molecule Force Spectroscopy (smFS)* (atomic force microscopy and laser optical tweezers) and *single-nanopore sensing*, are under active development and contribute to the exploration of IDPs[109].

Nuclear magnetic resonance (NMR) spectroscopy and *Molecular dynamics simulation* are two powerful tools to study IDPs that have been used in this research. Chapters 2 and 3 will present further discussion of these methods.

1.2. Transactivator of Transcription of Human Immunodeficiency Virus type 1

1.2.1. Human Immunodeficiency Virus type 1.

In 1984, scientists discovered the virus responsible for acquired immunodeficiency syndrome (AIDS)[110]. Soon after its identification, the virus, later named Human Immunodeficiency Virus (HIV), was predicted to be a “terrible disease” by U.S. Health & Human Services. Indeed, 35 years into the epidemic, the disease is neither preventable nor contained. During peak mortality in 2005, AIDS killed about 2 million people per year. The Joint United Nations Program on HIV and AIDS (UNAIDS) estimated there were 36.7 million people living with HIV at the end of 2015 and 1.1 million people died from illnesses caused by AIDS that year. South Africa is particularly hard hit, with about 19% of the population between the ages of 15 to 49 living with HIV[111].

HIV belongs to a group of retroviruses called lentiviruses. There are 2 HIV strains: HIV-1 and HIV-2. Worldwide, the predominant strain is HIV-1, and generally when people talk about HIV without specifying the type of virus they are referring to HIV-1. HIV-2 virus is relatively uncommon and is concentrated in West Africa. It is thought to be less infectious and progress slower than HIV-1. While commonly used anti – HIV-1 drugs are active against HIV-2, optimum treatment is poorly understood[112, 113].

The genome of HIV is made of ribonucleic acid (RNA), and each virus has two identical copies of the RNA[114] that interact, forming a dimer[115]. The HIV-1 RNA is 9719 nucleotides long[116, 117] with multiple open reading frames (ORF). There are 3 major genes, encoding major structural proteins and essential enzymes (Figure 1.12). Translation of these genes yields polyproteins which are cleaved into individual protein units by a viral protease (Table 1.1)

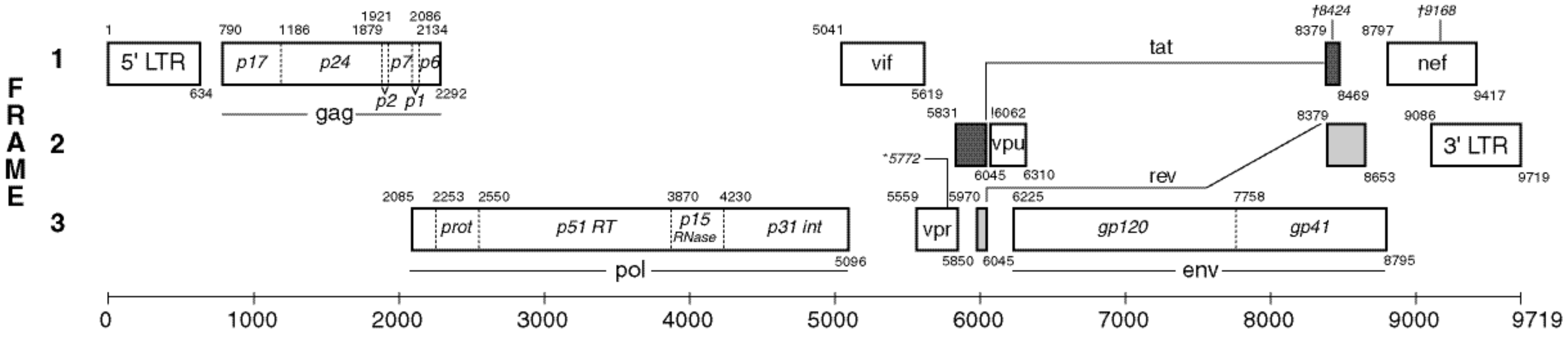


Figure 1.12. Genome organization of HIV – 1. Open reading frames are shown as rectangles. The gene start, indicated by the small number in the upper left corner of each rectangle, normally records the position of the a in the atg start codon for that gene; while the number in the lower right records the last position of the stop codon. For pol, the start is taken to be the first t in the sequence ttttttag, which forms part of the stem loop that potentiates ribosomal slippage on the RNA and a resulting −1 frameshift and the translation of the Gag-Pol polyprotein. The tat and rev spliced exons are shown as shaded rectangles. Reprinted with permission from reference[118]

Table 1.1. Proteins encoded by the HIV genome.

Class	Gene name	Primary products	Processed products
Viral structural proteins	Gag	Gag polyprotein	Matrix protein, MA Capsid protein, CA Nucleocapsid, NC Late assembly protein, p6 Spacer peptide, SP1, SP2
	pol	Pol polyprotein	Reverse transcriptase, RT Ribonuclease H, RNAse H Intergrase, IN Protease, PR
	Env	gp160	gp120, gp41
Regulatory proteins	Tat	Transactivator of transcription, Tat	
	Rev	Regulator of expression of virion proteins, Rev	
Accessory regulatory proteins	nef	Negative factor, Nef	
	vpr	Lentivirus protein R, Vpr	
	vif	Viral infectivity factor, Vif	
	vpu	Viral protein unique, Vpu.	

For replication, viral RNA is reverse-transcribed into DNA and then integrated into the host DNA of the infected cells. For HIV, its main target is the activated CD4 T-lymphocytes, which play a vital role in the human immune system. Figure 1.13 shows a structural model of an HIV-1 particle. The viral genome (two single strands of RNA) and other enzymes necessary for early replication steps, such as reverse transcriptase, proteases and integrase, are enclosed by a conical capsid composed of approximately 2000 capsid proteins (CA or p24)[119]. Reverse transcriptase (RT or p51) is the enzyme that copies the viral RNA into viral DNA inside the host cell; integrase (IN or p31) inserts the viral DNA into the host DNA in the nucleus; and protease (PR) cleaves the viral polyproteins into their functional units. The inner core also contains the negative factor (Nef), primarily involved in down regulation of CD4 surface expression; nucleocapsid (NC or p9) protein, which functions to deliver unspliced RNA for assembly of new virions and other accessory regulatory proteins such as lentivirus protein R (Vpr) or viral infectivity factor (Vif). The capsid proteins also shield the viral genome from cytoplasmic sensors that are capable of inhibiting infection and activating innate immune signaling pathways[120-122]. The viral core is surrounded by a matrix composed of the viral protein p17[123]. A small protein, p6, is located between the matrix and the capsid. It is important for virion budding and for the incorporation of Vpr protein into the particle[124]. The viral particle does not have a cell wall or a nucleus, instead, it is, in turn, surrounded by the viral envelope that consists of a lipid bilayer, glycoproteins (SU or gp120), and a transmembrane protein (TM or gp41). The envelope proteins form trimers, and an average virion contains approximately 20 surface glycoprotein trimers[125, 126]. A mature HIV virion is spherical and has a diameter of about 120 nm[127].

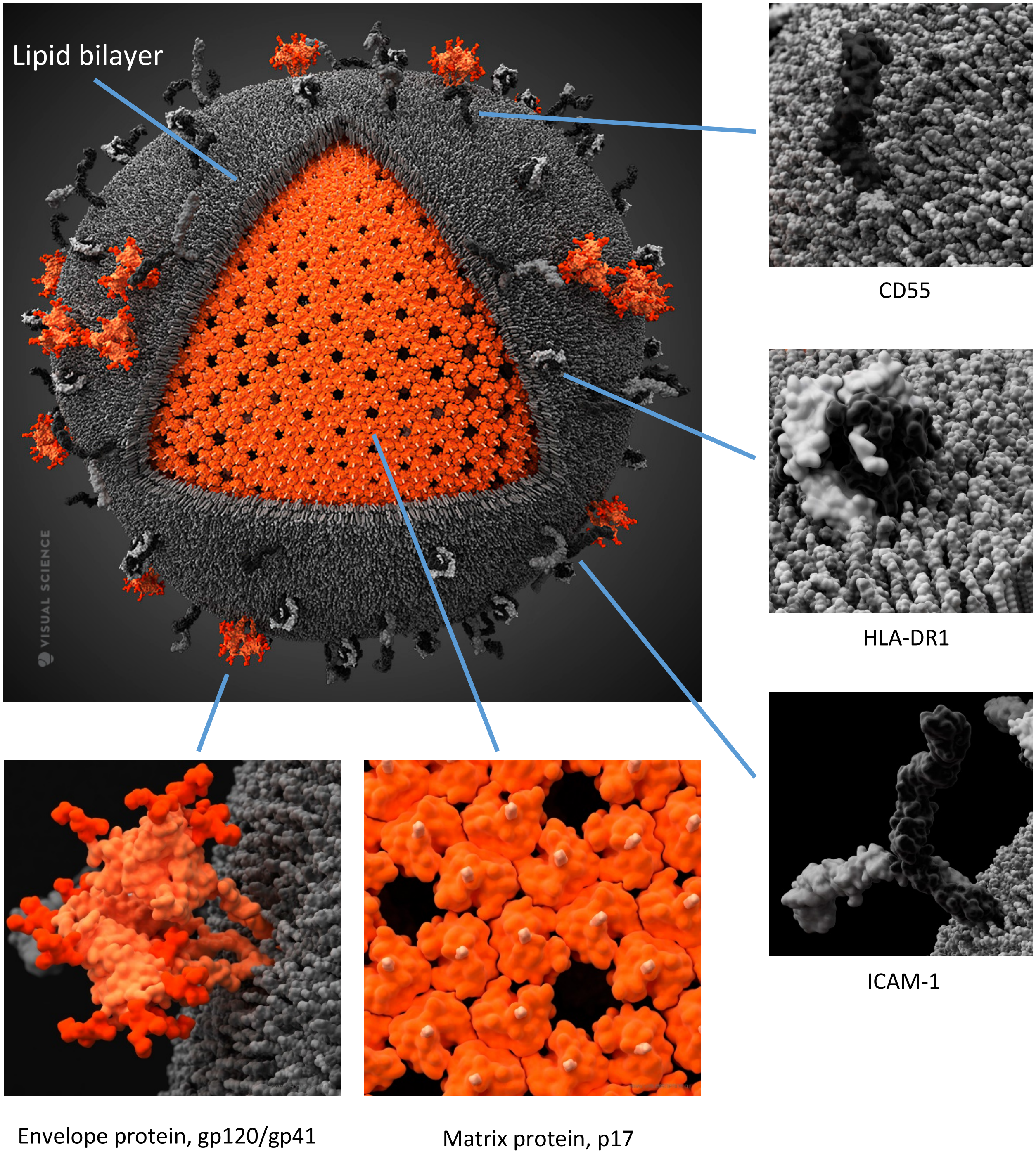


Figure 1.13.a. Structural model of an HIV-1 particle. Lipid bilayer is in gray, viral proteins are in orange, and membrane proteins captured from host cell are in black and white. Adapted with permission from reference [128] .

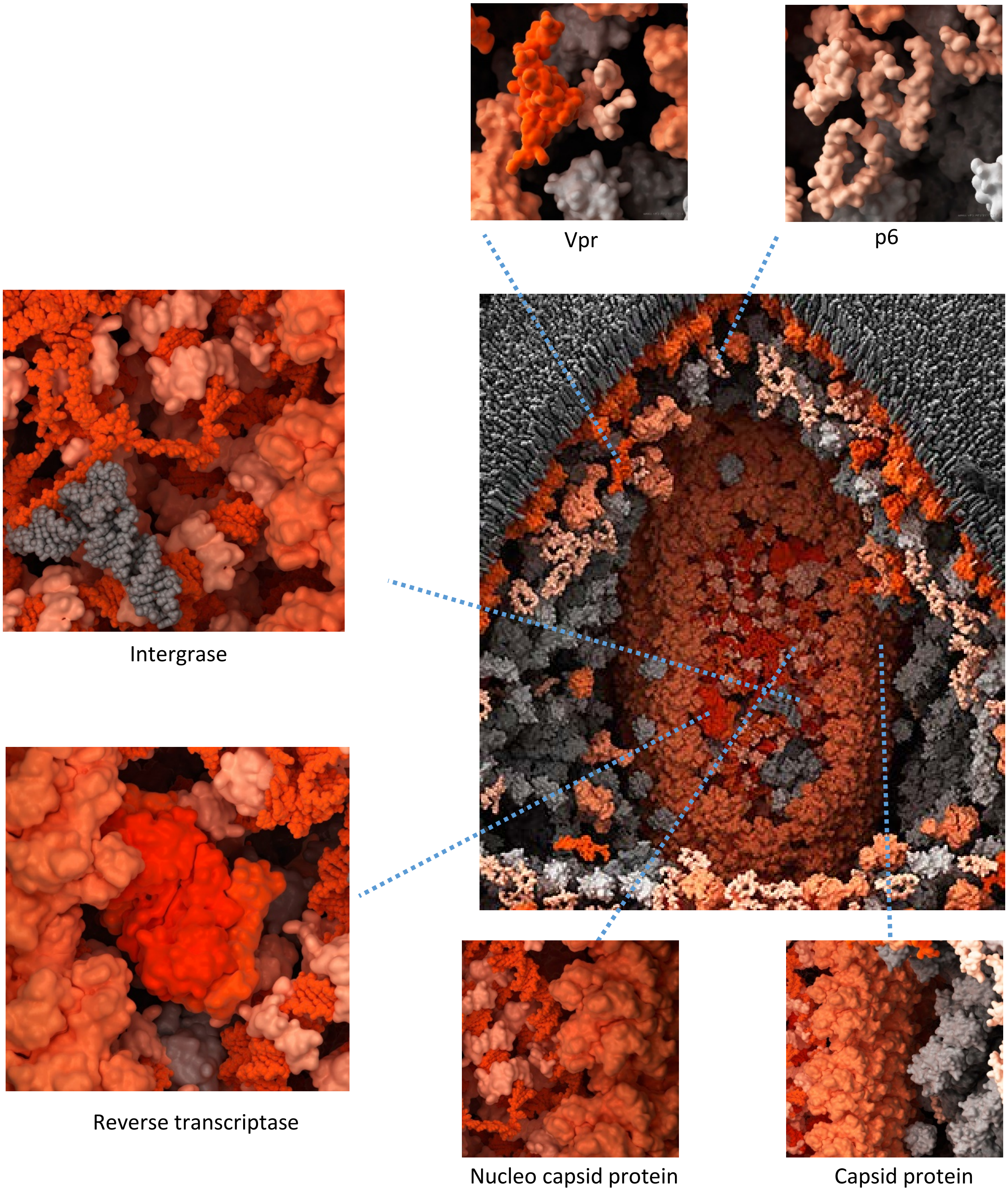


Figure 1.13.b. Structural model of the inside of an HIV-1 particle shows the capsid layer and the inner core of an HIV-1 particle. Adapted with permission from reference [128]

The HIV replication cycle (Figure 1.14) begins with virus entry into the host cell. HIV enters macrophages and CD4 T-cells *via* the binding of gp120 protein to CD4, a glycoprotein found on the surfaces of immune cells such as T-cells, monocytes, macrophages, dendritic cells and microglia. Once gp120 is bound with CD4, the envelope complex undergoes a conformational change, exposing the chemokine binding domains of gp120, allowing it to interact with the target chemokine receptor (CCR5 or CXCR4)[129]. The HIV viral envelope then fuses with the host cell membrane, and HIV RNA and enzymes are injected into the cell and the entry process is completed. In order to uncoat the capsid correctly, HIV requires host cyclophilin (or CypA), which is a peptidylprolyl isomerase that facilitates protein folding in the host cell. Cyclophilin is among many host cell proteins captured from the budding process of an HIV particle. It is found attached to viral capsid hexamers.

Following entry into a target cell, the relatively highly error-prone reverse transcription takes place, copying viral RNA into complementary DNA (cDNA). The resulting mutations from transcription errors may cause drug resistance or allow the virus to evade the immune system. cDNA and its complement form double stranded viral DNA that is then transported into the nucleus. The subsequent integration into the host cell chromosome takes place under the direction of another viral enzyme called integrase[130]. The intergrated DNA, called the provirus, may remain inactive for several years, producing few or no new copies of HIV.

When the host cell becomes active, the provirus uses the host enzyme RNA polymerase to create copies of HIV genomic material, as well as shorter strands of RNA called messenger RNA (mRNA). Some of the full-length RNAs function as new copies of the virus genome, while others function as mRNAs and will be translated to produce structural protein: the polyprotein (Gag), and the envelope protein (Env). The shorter mRNAs are translated into different viral regulatory proteins, including the transactivator of transcription protein (Tat) and the regular expression of virus protein (Rev). Rev protein binds to full-length viral RNA, allowing it to leave the nucleus for further translation or for viral assembly. Tat protein is essential for virus expression and its function is described in the next section.

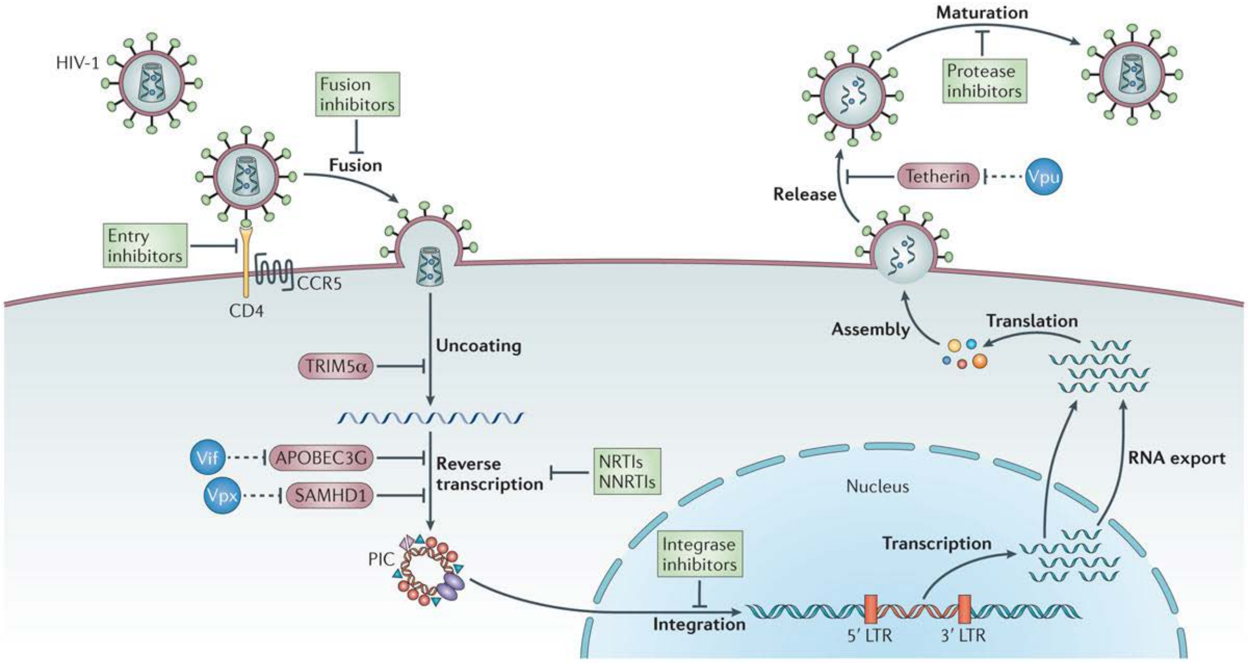


Figure 1.14. Illustration of the main steps of the HIV-1 replication cycle. Major families of antiretroviral drugs (green), and the step of the life cycle that they block, are indicated. Also shown are the key HIV restriction factors (tripartite motif-containing 5α (TRIM5α), APOBEC3G, SAMHD1 and tetherin; red) and their corresponding viral antagonist (Vif, Vpx and Vpu; blue). CCR5, CC-chemokine receptor 5; LTR, long terminal repeat; NRTIs, nucleoside reverse transcriptase inhibitors; NNRTIs, non-nucleoside reverse transcriptase inhibitors. Reprinted with permission from reference [131]. Copyright © 2013 Macmillan Publishers Limited.

After transcription, new viral RNA and viral proteins translocate to the cell surface to assemble into new immature virus forms, which then bud off and are released from the host cell. During this process, HIV protease cleaves the polyprotein to form the mature Gag proteins, resulting in new infectious virions. The virion also acquires lipid membrane from the host cell. Therefore, some cellular membrane proteins, including cell adhesion molecules and major histocompatibility complex (MHC) receptors, are incorporated into the HIV particle. For example, the Human Leukocyte Antigen DR1 (HLA-DR1)[132] and the Intercellular Adhesion Molecule 1 (ICAM-1)[133] are proteins that increase HIV infectivity, whereas the Complement Decay-accelerating factor (CD55)[134] (also referred to as DAF, decay-accelerating factor) participates in down-regulating the complement system, which blocks the formation of the membrane attack complex.

Medical treatment for HIV infection has made tremendous progress in terms of therapeutic options available, transforming the disease from a fatal to a manageable chronic one. In the early days of the epidemic, only the associated opportunistic infections could be treated with limited success. Not until HIV was identified to be responsible, and its life cycle was characterized, were the medical and scientific communities able to start investigating anti-HIV drugs. Azidothymidine (AZT) was the first compound reported to decrease mortality and opportunistic infections in patients with AIDS[135]. AZT, originally synthesized for anticancer purposes, was found to block the reverse transcription step of the HIV-1 replication cycle[136]. Due to the high mutation rate, viral resistance quickly developed and new drugs were needed. A decade later, the antiretroviral therapy (ART), which used a combination of a protease inhibitor (PI) and 2 other nucleoside reverse transcriptase inhibitors (NRTIs), was introduced and helped to reduce the morbidity and mortality remarkably[137, 138]. ART does not kill or cure, instead, it prevents the growth of the virus, thus, slowing down the progression towards AIDS. ART nowadays is recommended for all people with HIV, regardless of how long they have had the virus or how healthy they are.

Besides targeting the reverse transcription process, drugs that target other steps in the HIV replication cycle have also been developed. Drugs commercially approved by the Food and Drug Administration (FDA) are categorized into 6 classes, including non-nucleoside reverse transcriptase inhibitors (NNRTIs), NRTIs, PIs, fusion inhibitors, CCR5 antagonists (CCR5s) also called entry inhibitors, and HIV integrase strand transfer inhibitors (INSTIs). Standard treatment of HIV nowadays consists of a combination of at least 3 drugs (often called "highly active antiretroviral therapy" or HAART).

1.2.2. Transactivator of Transcription.

The HIV-1 transactivator of transcription (Tat), a small RNA binding protein essential for viral gene expression and replication[139-141], is expressed during the early stage of viral infection[142]. During the early stages of infection, viral transcription is halted due to the binding between RNA polymerase II and negative transcription elongation factor (N-TEF). After intergration into the human genome, HIV-1 establishes a low basal expression of Tat, Rev, and Nef proteins from multiply spliced short transcripts. Expressed Tat is transported into the nucleus, and acetylated at Lys28 by the p300/CBP-associated factor[143, 144]. It then recruits the positive transcription elongation factor b (P-TEFb) and an AF4 family member (AFF1/2/3/4) to RNA polymerase II (RNAPII) to form a super-elongation complex (SEC). The SEC binds to the transactivation response element (TAR), a 59- nucleotide stem−loop RNA structure located at the 5′ end of nascent viral transcripts[145, 146]. P-TEFb is composed of a regulatory cyclin subunit (cyclin T1), which interacts extensively with Tat, and a cyclin-dependent kinase 9 (CDK9). CDK9 hyper-phosphorylates the RNA polymerase II (RNAPII) carboxy-terminal domain and other elongation factors[141] such as negative elongation factor (NELF)[147] and the 5,6-dichloro-1-β-D-ribofuranosylbenzimidazole sensitivity-inducing factor (DSIF)[148], resulting in an increased polymerase processivity. Tat also relieves nucleosome repression by recruiting histone acetyl transferase (HAT)[149] to the HIV-1 promoter region, leading to nucleosome acetylation so that transcription of the viral genes by RNA polymerase II is not aborted. Tat is released from TAR and P-TEFb after being acetylated at Lys50 by p300/CBP and hGCN5[150]. Binding of Tat leads to a significant increase in the transcription elongation rate and the production of long transcripts in contrast to the short transcripts that result from transcription pausing in the absence of Tat.

Significant levels of Tat are detected in the serum of HIV-1 infected patients, even during antiretroviral therapy[151, 152]. Uptake of Tat into neighboring non-infected cells[153] alters the functions of components inside these cells[154]. For example, Tat has been proven to have an effect on the development of HIV-1 associated neoplasms[155]. Extracellular Tat also contributes to the growth and tumorigenesis of human Kaposi's sarcoma cells[156]. Tat interacts with tumor suppressor proteins (Rb2/p130), inhibiting the growth control activity of these protein[155]. Interaction of Tat with the component of the DNA repair systems, which protects the genome from deleterious damage, may also result in cancer[157]. Tat can cross the Blood-Brain-Barier (BBB) and is detected in the central nervous system (CNS)[158]. It has been shown to adversely impact a variety of cell types in the CNS including neurons, astrocytes, brain microvascular endothelial cells, microglia and macrophages[159] , leading to a variety of harmful consequences[159].

Tat is a 101-amino acid protein encoded by two exons[160-163]. The amino acid sequence of Tat_{101} contains a low overall hydrophobicity and high net positive charge. Early reports suggested that the first exon product (residues 1− 72) is fully active in transactivation[162], and indeed, many of the important functions of the protein are present within this segment[164]. The first Tat exon encodes a proline-rich and acidic segment (residues 1−21), a cysteine-rich segment (residues 22−37) through which Tat binds zinc and interacts with CycT1[165], a hydrophobic core (residues 38−47); a basic segment (residues 48−57) necessary for TAR binding, Tat cellular uptake, and nuclear translocation[145] and, finally, a glutamine-rich segment (residues 58−72) involved in T-cell apoptosis[166]. Despite the high mutation rate of HIV-1, the first Tat exon product is relatively highly conserved, especially its first 56 residues, among which are the tryptophan residue in the proline-rich segment[167] and the essential cysteine residues (except for Cys-31)[165] in the cysteine-rich segment. This is likely an indication of the functional importance of these conserved residues in Tat activity and/or structure.

The second tat exon product (residues 73−101) has greater sequence variation than Tat_{72}. Figure 1.15 shows no conserved residues within the Tat second exon. The high level of sequence diversity is mostly due to the error-prone nature and low fidelity of reverse transcriptase, the poor proofreading by the polymerase, and perhaps other pressures exerted by the host immune response or by the antiretroviral chemotherapy[161]. Even though the transactivation function can be solely accomplished by the first segment, the biological importance of Tat’s second exon product is suggested by its conserved expression in all lentiviruses[168]. $\text{Tat}_{73\text{–}101}$ has two identified sequence motifs: the RGD motif, which is involved in cell adhesion[169], and the ESKKKVE motif, which is related to optimal HIV-1 replication *in vivo*[161, 170]. Previous studies showed that the second exon is essential for Tat-mediated cell genome dysregulation[171]. It may also control non-viral gene transcription (TAR-independent activation) *via* binding to canonical enhancer sequences of cellular transcription factors such as NF-κB or Sp1 and indirectly changing the expression of several genes[168, 171, 172]; this binding in turn may alter cellular functions such as T-cell activation and apoptosis. The biological requirement for the second Tat exon in HIV-1 replication and pathogenesis *in vivo* is further supported by a study that showed that macaques infected with SIVtat1ex virus, which expresses a one-exon tat gene, survived longer than animals infected with wild-type virus SIVmac239[173]. The authors also compared this result with the accidental infection of three laboratory workers with HIV IIIB (two cases in 1985 and another case in 1990), an HIV-1 isolate that has a premature stop codon in its *tat* gene. The virus from one of the infected patients had its one-exon Tat reverted to two-exon Tat, causing a steep decline in CD4+ T-cell count and a rapid progression to AIDS within 1 year[173, 174].

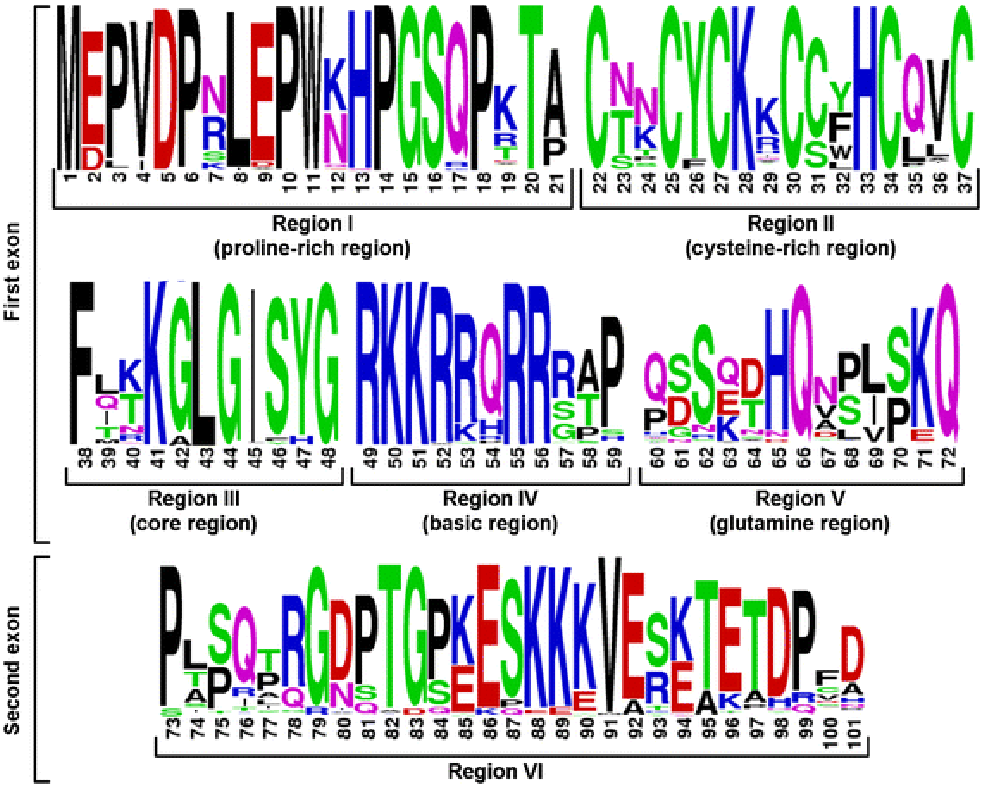


Figure 1.15. HIV-1 Tat$_{101}$ protein sequence. The WebLogo was constructed using 3378 HIV-1 Tat sequences with 101 amino acids from the NCBI database. Identical sequences were

removed. The letter size in the Weblogo is proportional to residue conservation. Reprinted with permission from reference [151]. Copyright © Springer Basel 2015.

Tat is multifunctional[151, 175] with a wide range of interactions with many different proteins[161, 176]. This is likely a direct effect of its intrinsic disorder[93, 160]. Because of its structural flexibility, Tat may adopt different local conformations to interact with different binding partners. One example of this can be observed in the crystal structure of Tat in complex with P-TEFb, AFF4 and TAR (Figure 1.16). The crystal structure reveals an extended loop conformation in the N-terminal acidic/proline-rich segment containing two type II β-turns, one type II′ β-turn, and a 3_{10} helix[165]. On the other hand, a peptide corresponding to residues 1−16 of Tat binds to an antibody fragment (Fab′) in a standard type I β-turn conformation without forming the 3_{10} helix or any of the other turns[177].

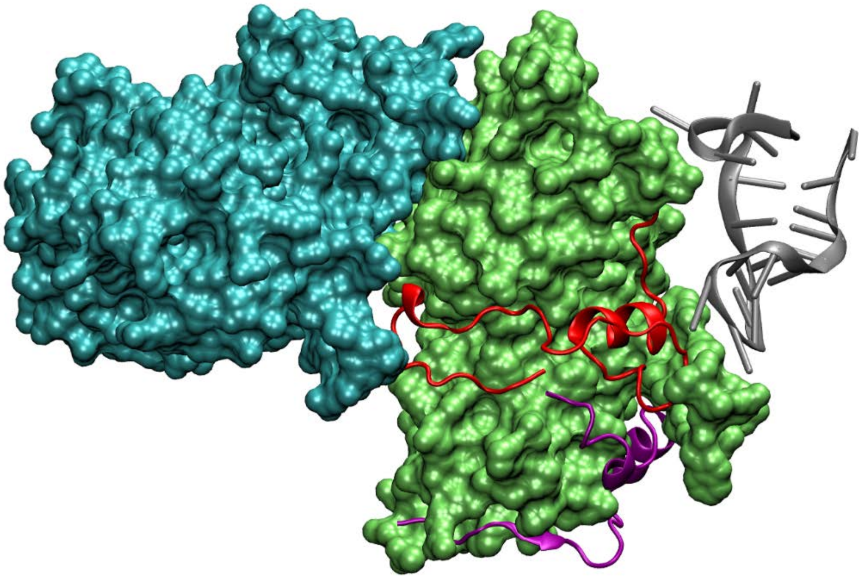


Figure 1.16. Crystal structure of complex of 58-residue Tat (red), P-TEFb complex of Cdk9 (blue) and CyclinT1 (green), AFF4 (purple), and TAR (gray). Tat and AFF4 are disordered

proteins but they form helical regions in the complex. Structure representation was created from entry 5L1Z[178] from the PDB using VMD.

1.3. Protein dynamics – the energy landscape.

Proteins in solution are not static objects, instead they experience fluctuations at various timescales (fs to s or even longer), amplitudes (from tenths of Å to nm) and directions. Protein motions result from thermal energy and cause protein structural deviations from the native state. These structural deviations vary across proteins; some with lower, and some with higher structural variability, measured by root-mean-square deviation (RMSD) and other parameters[179]. These alternative structures exist in equilibrium and are so-called protein conformations. Proteins, thus, are seen as dynamic entities that sample a large ensemble of conformations around an average structure[180].

Functions of proteins are determined not only by their structures, but also by their dynamics properties. Classic examples are myoglobin (Figure 1.17) and hemoglobin, where structural fluctuations are needed to open transient pathways for substrate to get into the binding sites[181]. For many enzymes, solvent and substrate molecules gain access to the binding sites, which are completely buried in the interior of the protein, in a similar fashion. Another example is the case of hexokinase[181]. This enzyme transfers a phosphate to glucose much faster than to water because the binding of glucose induces a large conformational change in the protein that brings the two protein domains together and creates a protected catalytic site. The closure of the domains is slow, but it would not be possible without the numerous rapid atomic motions that allow groups of atoms to slide past each other, despite their repulsive interactions. Protein folding is another aspect of protein dynamics where understanding how the newly synthesized polypeptide chain is able to fold to its native structure is fundamental to the description of life at a molecular level. This issue is especially critical since misfolding of proteins typically leads to disease[182]. Protein dynamics are also crucial for signal transduction. In order to relay signals, proteins shift among different energy states to respond to inputs[183].

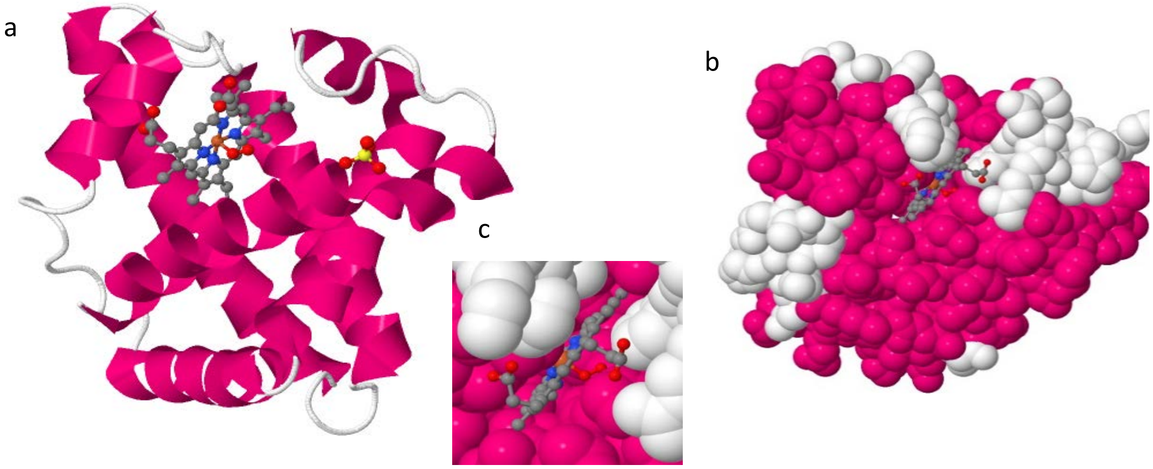


Figure 1.17. Crystal structure of Myoglobin in (a) ribbon and (b) surface representation. The heme group is represented in ball and stick model. Structural fluctuations are needed to open transient pathways[181] for substrates to access the heme group that is buried inside the protein. (c) Close up view of the active site. The figure was created by VMD, using the PDB entry 1mbn[184].

Dynamic properties of proteins are characterized by their thermodynamics (the populations/probabilities/lifetimes of conformational states of the ensemble) and the kinetics (the transition times and energy barriers between conformations), both of which are governed by free energy landscape theory. Figure 1.18 represents an energy landscape describing the energetic relationships between possible protein conformations in 2 particular sets of conditions (dark and light blue) such as temperature, pressure, solvent composition and mutations. A change in any of these conditions leads to the conversion of the landscape from one to another (dark to light blue), and the redistribution of the population of conformational states. Within one energy landscape, a state/conformation is a minimum in the energy surface, whereas a transition state/conformation is the maximum between the wells. Each dynamic tier corresponds to one type of motion, happening within a certain timescale.

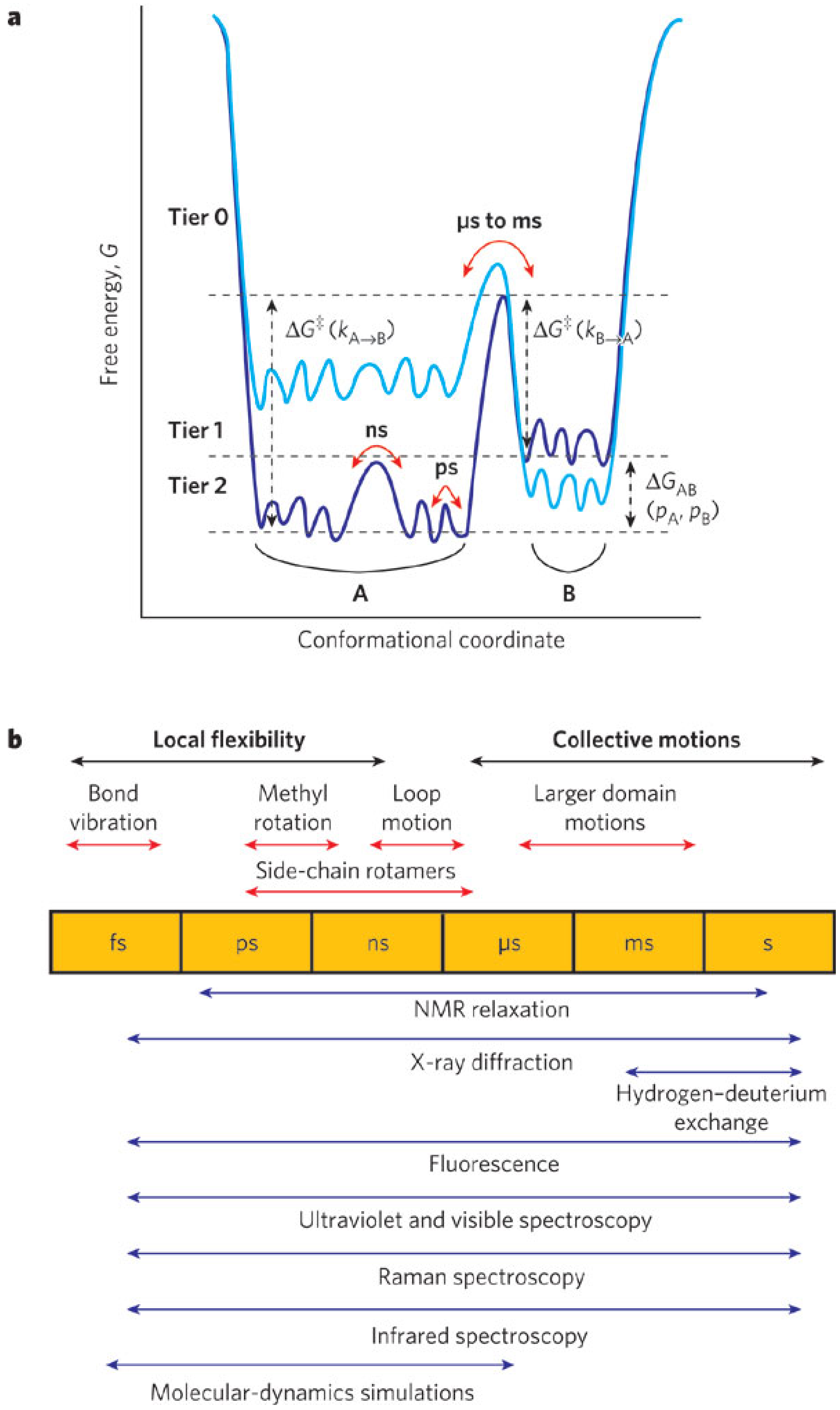


Figure 1.18. a, Energy landscape of a protein showing the hierarchy of protein dynamics and the energy barriers. b, Timescale of dynamic processes in proteins and the experimental

methods that can detect fluctuations on each timescale. Reprinted with permission form reference [180]. Copyright © Nature Publishing Group.

For the tier-0 dynamics, the conformations A and B are separated by energy barriers of several kT, where k is the Boltzmann constant and T is the absolute temperature[180]. The population of these conformations follows the Boltzmann distribution, depending on the difference in free energy (ΔG_{AB}). The number of conformations in this dynamic tier is typically small (2 conformers, A and B, for the energy landscape in Figure 1.18). The high energy barrier between the states restricts the conversion rate, and the fluctuation events happen on the µs to ms timescale or longer[185]. Excited/transition states have important functional roles in biochemical processes, including molecular recognition and ligand binding[186-192], enzyme catalysis[193-195], and protein folding[196-198]. Conformational transitions in this dynamic tier are found in domain/multiple domain protein motion.

Lower dynamic tiers (tier-1 and tier-2) describe fluctuations between large numbers of closely related sub-states within each tier-0 state. Energy barriers that separate these states are low (less than 1 kT), and therefore they can be easily overcome by thermal energy, leading to ps to ns timescale dynamics. Motions of tier-1 dynamics are of a small group of atoms that are collectively fluctuating (loop motion) whereas, tier-2 dynamics describe motions belonging to atomic fluctuations such as bond vibrations or side-chain rotations. These motions are local and their information is often obtained from spin relaxation NMR measurements, from other spectroscopic methods such as fluorescence, UV-VIS, Raman, IR spectroscopy, or from room temperature X-ray crystallography[199]. Local fluctuations working concertedly lead to collective motions, happening on the µs and longer timescales.

The energy landscape for IDPs is significantly different from folded proteins. The folded protein energy landscape has a "funnel-shape" with a global energy minimum corresponding to the native structure[200, 201]. The native state entropy determines the width of the "funnel". In the case of IDPs, the energy landscape has multiple local energy minima separated by small barriers, forming "shallow wells". These barriers can be easily and frequently overcome, leading to the property of existing as an ensemble of a large numbers of states having approximately equal energies. Figure 1.19 shows an example of a "funnel-like" energy landscape of a folded protein, human nucleoside diphosphate kinase (NDPK, PDB ID: 1nsk)[202], and an energy landscape of an intrinsically disordered peptide (CcdA C-terminal, PDB ID: 3tcj)[203], where the peptide may exist in different conformations, occupying different shallow wells on the energy surface. The disordered regions in NDPK also adopt different conformations, resulting in a rugged global minimum as shown on the enlarged Figure 1.19.c.

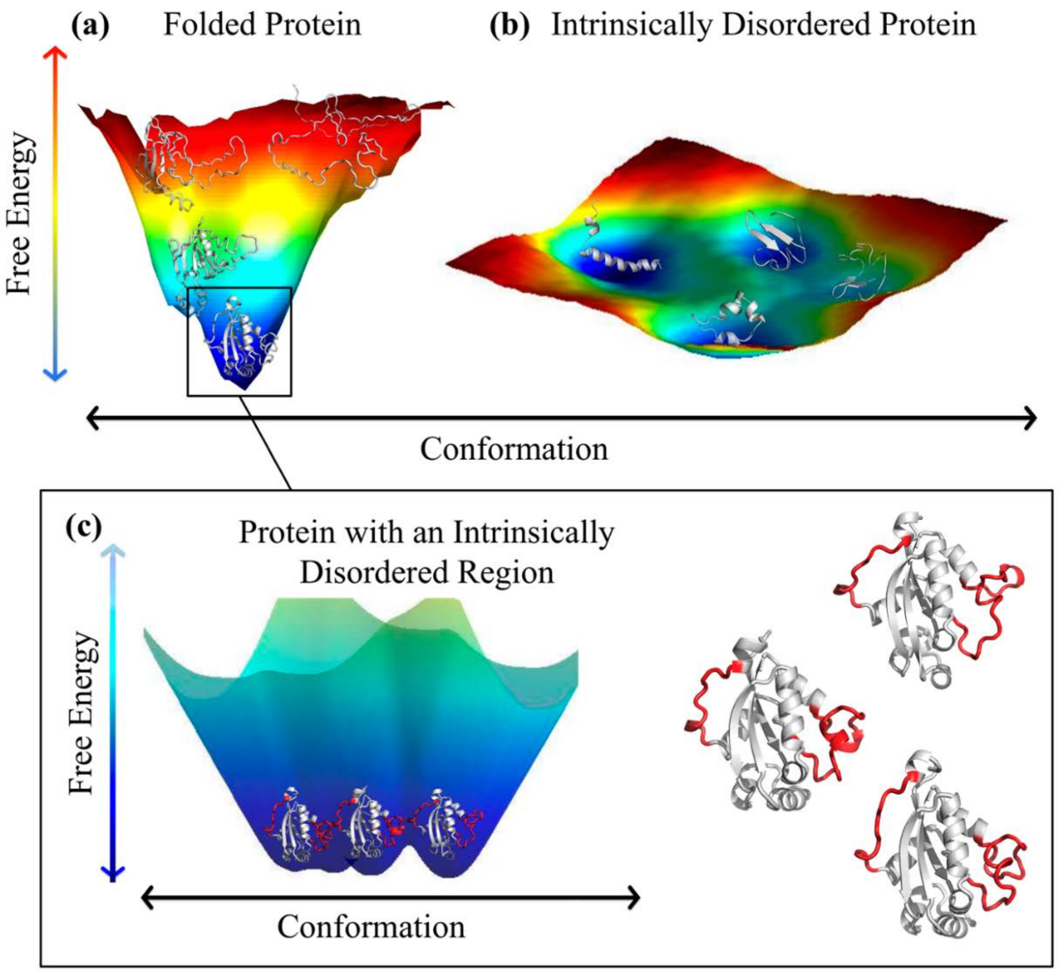

(a) Folded Protein
(b) Intrinsically Disordered Protein
Free Energy
Conformation
(c) Protein with an Intrinsically Disordered Region
Free Energy
Conformation

Figure 1.19. Schematic of protein free energy landscapes for (a) human nucleoside diphosphate kinase (NDPK), (PDB ID: 1nsk) and (b) an intrinsically disordered peptide (CcdA C-terminal, PDB ID: 3tcj); (c) close-up of the minimal free energy well in (a), where IDRs are shown in red and ordered regions are shown in white. The example NDPK conformations are shown again enlarged to the right for better visualization. In (a–c) lower free energy (dark blue) represents more probable conformations whereas less probable conformations are represented in red. Reprinted with permission form reference [204].

1.4. Purpose of the thesis.

Conformational ensemble and dynamics of proteins provide important information about proteins functions and mechanism (see section 1.3). So far, there is only one dynamics study conducted on the first exon product of Tat consisting of 72 residues in fully reduced condition using nuclear magnetic resonance (NMR) spectroscopy[160]. This study showed a predominantly disordered and extended (random coil) conformation with some extent of folding propensity in the cysteine-rich region and the hydrophobic core region of this protein segment. The fast dynamics on the timescale of ps-ns of this segment were also explored, revealing the uniformly distributed internal motion throughout the sequence, except for the two termini with a higher degree of motion. Several earlier studies on shorter segments of Tat suggested conformational transitions of the peptide upon binding. For example, NMR spectroscopy suggested a conformational change in Tat_{32-72} in the region of Gly-42 and Gly-44 upon binding to TAR[205], and the extended conformation of the bound form of Lys-50 acetylated $\mathrm{Tat}_{46\text{-}55}$ with the bromodomain of p300/CBP-associated factor (PCAF)[206]. No study has been conducted on full-length Tat_{101} protein to explore the conformational ensemble and dynamics, especially on the biologically important timescales of μs – ms and longer.

My thesis, taking into account the biological importance of the second exon product of Tat, aims to study the structure, dynamics, disorder and structure propensity of this region and of its impact on the structure, dynamics, disorder and structure propensity of the rest of the full-length protein. NMR spectroscopy will be used as the principal technique to study Tat_{101} protein in a fully reduced condition. The chemical shifts and the data from relaxation and hydrogen exchange measurements will be extensively analyzed to extract information about the protein conformational ensemble and dynamics. Classical Molecular Dynamics simulations will be used as a complementary technique to verify the NMR results, as well as to sample the protein conformers that are invisible to NMR due to ensemble averaging. The combined results on the full-length protein may also provide a deeper understanding of the mechanisms through which it interacts with partners as an IDP; and this is an important step toward rational drug design. Considerable effort has been expended in the design of Tat antagonists to inhibit HIV-1 transcriptional elongation. This continues to be an important goal, but more attention is now focused on promoting proviral transcription to reactivate latent virus and so eradicate latent viral reservoirs. Whether the therapeutic goal is activation or inhibition, the insight into the mechanisms derived from protein dynamics information will be helpful in pursuing both.

In Chapters 2 and 3, I will describe some of the background and theory behind the principal methods used in this research namely, NMR spectroscopy and molecular dynamics simulations, paying particular attention to their application to intrinsically disordered proteins. Some readers may prefer to skip or skim these chapters.

Chapter 2

NMR spectroscopy in studies of Intrinsically Disordered Proteins.

The phenomenon of nuclear magnetic resonance (NMR) was discovered and developed by Purcell and his team at the Massachusetts Institute of Technology[207], and Bloch and coworkers at Stanford University[208]. The two scientists were granted the Nobel Prize in 1952 for "*their development of new methods for nuclear magnetic precision measurements and discoveries in connection therewith*", marking the world's recognition for their scientific contribution to the invention of one of the most powerful tools available for structural biology. A few years after the initial discovery, NMR was introduced into the field of chemistry following the observation of the chemical shift from two different signals from nitrogen atoms in NH_4NO_3, and three lines in the spectrum of ethanol[209]. In 1953, the first evidence of spin-spin interactions, known later as the "nuclear Overhauser effect" (NOE), was reported as an observed increase of nuclear polarization upon the saturation of electrons in metals[210]. The emergence of superconducting magnets in the early 1960s helped solve issues related to magnet inhomogeneity and instability[211]. Application of the Fourier transformation to NMR in 1966 by Ernst and Anderson was another major breakthrough, opening the door to the development of NMR pulse sequence[212]. Ten years later, two-dimensional (2D) NMR, an experiment to show the correlation between spins *via* J-coupling or the NOE, was introduced, paving the way for resonance assignment[213]. 2D NMR was quickly applied in the field of biomolecules and the first protein structure was reported in 1985[214] using the NMR technique.

Rapid progress in the development of NMR hardware and software has been made. Improvements in magnet technology and wire technology allowed the construction of NMR machines with ultra – high field strengths up to 23.48 Tesla (Lyon's European Nuclear Magnetic Resonance Center, and Japan's National Institute for Materials Science), compared with the first commercial 0.7 Tesla machine available in the 1950s[215, 216]. Introduction of cryogenically cooled probes has significantly increased the experimental sensitivity. Modern console electronics allow radio frequency (RF) digitization with timing resolution of a few tens of nanoseconds, enabling the fast switching of frequency, amplitude and phase as required for modern NMR experiments. Introduction of new pulse sequence programs[217] granted access to the measurement of samples with increasing molecular weight. For example, in comparison with the limit of 20 residues in some peptide studies in the 1970s, molecular machines with aggregate molecular masses up to 1 MDa can be studied at atomic detail[218] today.

Besides X-ray crystallography and cryoelectron microscopy, NMR is the only method that provides information at atomic resolution on the structures of biological macromolecules. In contrast to X-ray crystallography where protein must be crystallized, NMR allows studies of proteins in the liquid state. This makes structural studies of highly dynamic systems such as intrinsically disordered proteins (IDPs), where the proteins do not crystalize, possible. Various NMR experiments are capable of providing dynamics information over a wide range of timescales, from slow exchange, where interconverting species are visible, to fast motions using relaxation measurements, and under equilibrium conditions (Figure 2.1). In many cases, the exchanging species with slightly higher free energy, called excited – state conformers, are sparsely populated, just a few percent of the whole population, and exist transiently with extremely short lifetimes, making them invisible to many tools of modern biophysics. Understanding these conformers is important for their functional roles in biochemical processes, and NMR spectroscopy has paved the way for studying such 'invisible' excited states[185].

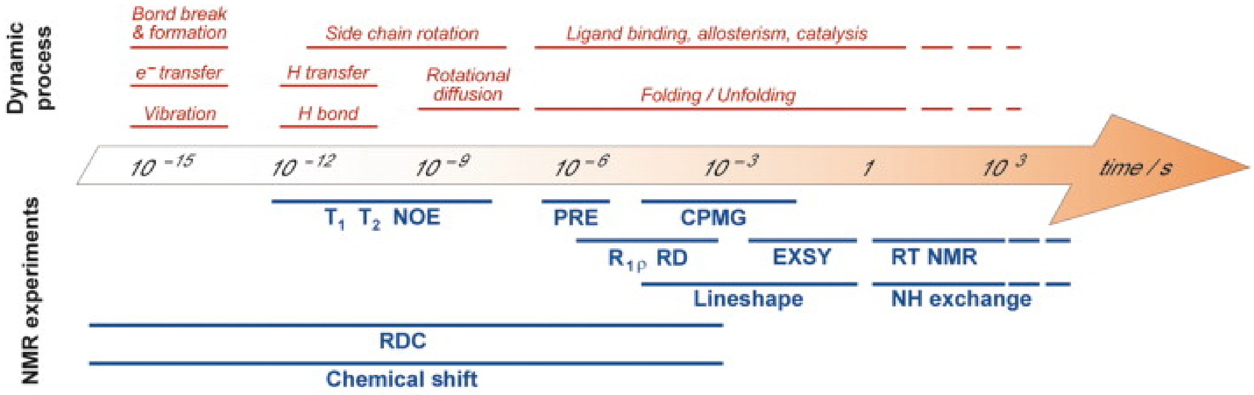


Figure 2.1. Protein dynamics timescale (top) and NMR experiments suitable for dynamic quantification (bottom). Reprinted with permission from reference [219]. Copyright © 2013 Elsevier Inc.

NMR is used in an integrated fashion with other experimental techniques in protein studies. Dynamics information from NMR relaxation experiments is normally complemented with 3D structural information from X-ray crystallography, SAXS, or cryoelectron microscopy. For example, dynamic data can be interpreted in terms of interconversion between states that a protein adopts when crystalized with and without ligands or among conformers with amino acid mutations. Many molecular dynamics (MD) simulation force fields were developed and validated based on NMR data[220]. MD simulations and NMR are also used in tandem to study protein dynamics. For example, in a study on the enzyme cyclophilin A, the authors compared the generalized order parameters in the Lipari-Szabo approach[221, 222], derived from nuclear spin relaxation data and calculated from simulations, to come to a conclusion that this enzyme catalyzed protein isomerization *via* an electrostatic handle mechanism[223]. SAXS and NMR in combination can provide complementary information on highly dynamic systems, especially on IDPs. An ensemble of structures representing the dynamic behavior of a protein can be calculated using paramagnetic relaxation enhancements (PREs), residual dipolar couplings (RDCs) and NOEs along with SAXS data[104]. Phillips *et al.* successfully used time-resolved SAXS and NMR to explore the relationship between the internal motion and the activation of the protein cellular Inhibitor of Apoptosis 1 (cIAP1).

Albeit NMR is a powerful tool, it also faces some challenges when applied to proteins. The foremost issue is the requirement for a concentrated isotopically-labeled protein sample. Solving the problem means having to answer the questions of how to make enough labeled protein for the study, how to keep the protein stable at high concentration in a certain amount of time (up to several days), and how to reduce the amount of isotopic material to a minimum since they are costly expensive in some cases (certain labeled amino acids for site specific labeling, or labeled pyruvate for side-chain labeling). For IDPs, this problem becomes more challenging since IDPs are more prone to aggregate at high concentration[224]. The conformational dynamics and transient structure of IDPs are highly sensitive to experimental conditions such as pH, buffer composition, salt concentration and temperature. Under certain conditions, some parts of an IDP may undergo conformational dynamics on a timescale that leads to extensive line broadening, and thus diminished or missing resonance peaks. Another challenge for NMR is the limit on the size of the molecule. Signals from proteins are typically weak, broadened and overlapped, hindering the assignment of resonances and acquisition and interpretation of high quality data.

An IDP is different from a folded protein in terms of structure and dynamics and therefore NMR signals, chemical shifts and relaxation parameters, the quantities that describe protein dynamics, are also different between the two classes. For an ordered protein, the local electronic environment for each nuclear spin is unique due to the stable compact three-dimensional structure; therefore the induced shielding fields are different, leading to a distribution of chemical shifts over a broad range. In contrast, a less compact structure and conformational averaging in IDPs significantly reduces the contributions of the local environment to chemical shifts, causing severe overlap of resonances (Figure 2.2). The low chemical shift dispersion of IDPs makes resonance assignment very challenging. To tackle this issue, two approaches have been used, namely residue-specific NMR, and multi-dimensional NMR experiments. The former approach can be achieved either by using different labeling strategies such as site specific labeling[225] or unlabeling[226], or by NMR experiments that can give rise to signals from only a specific amino acid type[227-231]. The later approach uses non-uniform sampling[232, 233], projection reconstruction[234, 235] and parallel acquisition[236] techniques to measure carbon-detected NMR signals at higher dimensionality (≥ 4D)[237-239]. Carbon detection gives higher peak dispersion, superior to the conventional HN signal detection. It is also particularly useful in the case of IDPs where amide proton signals are broadened, sometimes even beyond detection, because of chemical or conformational exchange. Notably, carbon-detected NMR allows the assignment of proline residues, which frequently are abundant in IDPs[240]. Proline residues do not give a signal in HN-detected spectra, causing breaks in the sequential assignment of the protein backbone resonances.

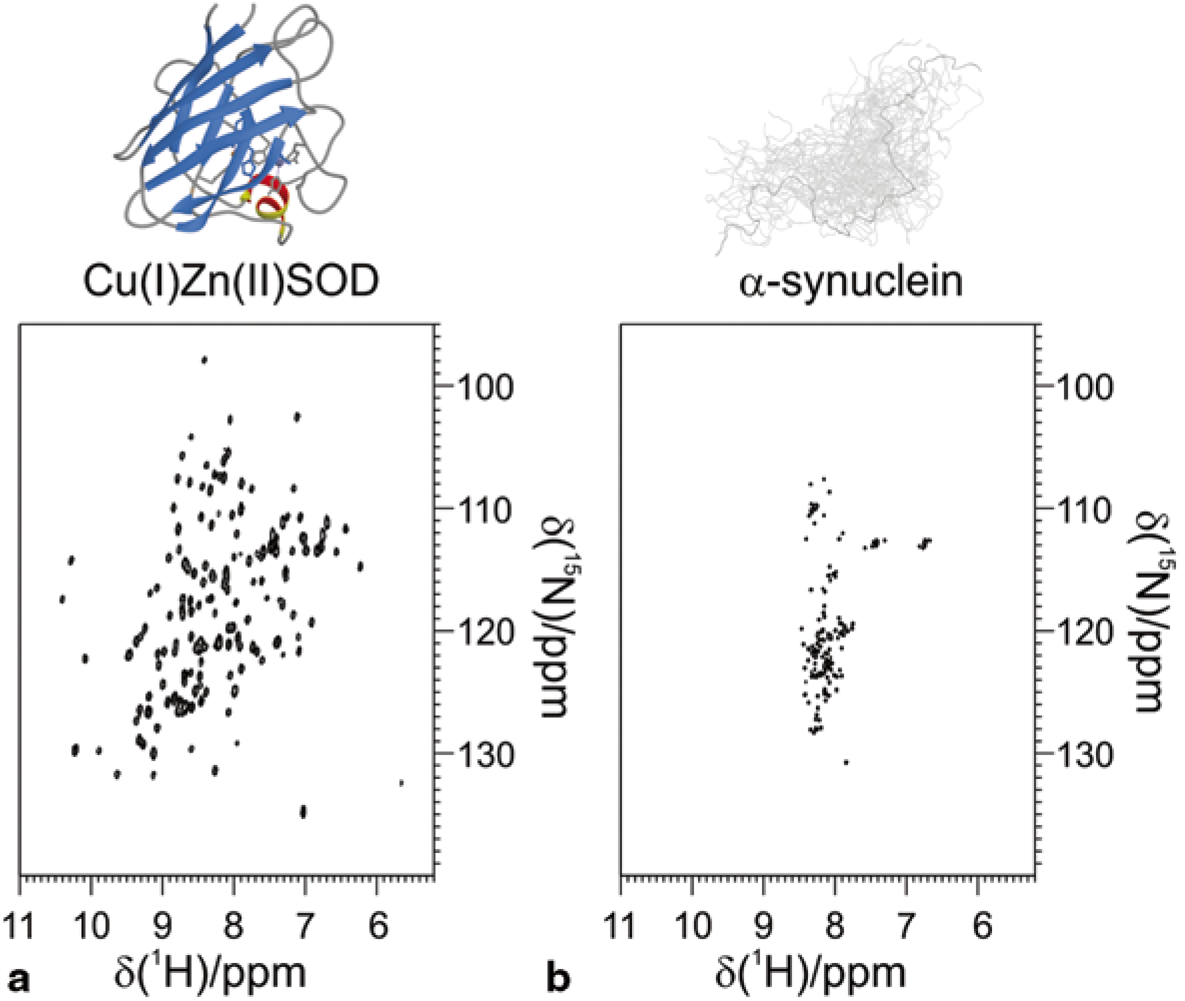


Figure 2.2. A comparison of chemical shift dispersion for a structured and an intrinsically disordered protein in 2D ^{1}H-^{15}N HSQC spectra acquired on two proteins of similar size, but characterized by different structural properties. (a) The HSQC spectrum of structured monomeric Cu(I)Zn(II) superoxide dismutase (1.5 mM sample in 20 mM phosphate buffer, pH 5.0, at 298 K; PDB code: 1BA9). (b) The HSQC spectrum of intrinsically disordered α-synuclein (1.0 mM sample in 20 mM phosphate, pH 6.4, 0.5 mM EDTA, 200 mM NaCl, at 285.5 K). The experiments were acquired on a 700 MHz Bruker AVANCE spectrometer equipped with a CPTXI probe. Reprinted with permission form reference [241]*. Copyright Springer International Publishing Switzerland.*

In terms of dynamics, the compact stable structure of a folded protein allows the description of its Brownian motion with a single overall rotational correlation time. For an IDP, because of the flexibility resulting from the inter-conversion between conformations due to their small energy difference, a single overall rotational correlation time cannot be defined[241]. The high flexibility of IDPs also strongly affects the relaxation rates (discussed in 2.2.1) and NMR line widths. Compared to folded proteins, IDPs have lower spin relaxation rates. This dynamic feature allows the use of longer magnetization transfer pathways in higher dimensional experiments. Fast motion in IDPs gives rise to narrow line widths, making it amenable for NMR characterization. Finally, the non – compact structure of IDPs allows exposure of the entire protein backbone to solvent, resulting in high hydrogen exchange rates. Measurements of solvent exchange rates can be used to distinguish between highly structured and disordered regions of proteins, and to predict the folding propensity by identifying the amide sites that are exposed to and/or protected from solvent[242].

In the following sections, I introduce the NMR methods used in this project to characterize the structure, dynamics on the ps-ns timescale (nuclear spin relaxation), dynamics on the µs-ms timescale (relaxation dispersion), and dynamics on the ms timescale and slower (hydrogen exchange NMR) of an IDP, the HIV-1 Tat_{101} protein.

2.1. Chemical shift, secondary structure propensity and coupling constant.

Chemical shifts have long been known to carry sufficient information to determine protein structure[243-245]. The effect of protein secondary structure on a protein's chemical shifts is theoretically explained by the dependence of the nuclear shielding effect on the dihedral angles ϕ

and ψ[246]. Structural information can be extracted from the so – called secondary chemical shifts of each nucleus, which are derived as the difference between the measured chemical shift and the predicted random coil values. For an IDP, because of the lack of a unique three-dimensional structure, the secondary chemical shift of a spin indicates the relative tendency of a measured residue to adopt a specific conformation at that point in the primary sequence.

For a ^{13}C and ^{15}N enriched protein, a large number of chemical shifts ($^{1}H_{N}$, $^{1}H_{\alpha}$, ^{15}N, $^{13}C'$, $^{13}C_{\alpha}$, $^{13}C_{\beta}$) can readily and precisely be measured, therefore, the quality of the secondary chemical shifts mostly depends on the predicted random coil chemical shifts used. Many data sets of random coil chemical shifts have been reported either from experimental measurement of disordered peptide[247-253] or from statistical analyses of deposited chemical shift assignments[254-257] in the Biological Magnetic Resonance Data Bank[258] (BMRB). These data sets have slightly different values due to the differences in the methods used and the external conditions including pH, co-solvents and temperature. The contribution of neighboring residues to random coil chemical shifts is also necessary to be taken into account for many reasons. Firstly, the side chains of neighboring residues may change the polarization of chemical bonds, leading to a change in nuclear shielding that determines the random coil chemical shifts. Secondly, chemical shifts may be altered by the through-space ring-current effect induced by the π-electrons in aromatic rings of some residues, and the magnetic anisotropy shielding effect from all carbonyl bonds. Furthermore, the hydrogen bonding potential of the neighbors may also affect the chemical shift through side-chain-to-backbone hydrogen bonds. Finally, steric clashing between the side chains of neighboring residues affects the free energy of the backbone conformations, thus affecting the populations of the major wells in the Ramachandran map[259]. This effect is notably strong for the residues before a proline because of the clash between the proline C_{δ} and the preceding-residue's C_{β}[260], explaining the abnormal random coil chemical shift of such residues[253]. The effect of neighboring residues on a chemical shift is considered in some data sets such as the one reported by Tamiola *et al*.[256] and the one by Kjaergaard *et al*.[253].

Chemical shifts of different nuclei are reporters of secondary structure to a different extent. For example, several studies have shown that $^{1}H_{\alpha}$ experiences upfield shifts of about -0.3 ppm in α-helical conformations, and downfield shifts of approximately 0.5 ppm for β-sheets. Amide protons only show downfield shifts for β-sheets, while the influence on the shifts for α-helices is negligible. The distribution of secondary chemical shifts of amide proton are larger than those observed for alpha protons[261]. This evidence suggests that the correlation between amide protons and backbone conformation is not as strong as for the alpha proton. For ^{13}C, secondary chemical shifts of $^{13}C_{\alpha}$ and $^{13}C_{\beta}$ are the most sensitive to backbone torsion angles. In helices, the $^{13}C_{\alpha}$ chemical shift is downfield by 2.2 ppm (for Ala) to 4.5 ppm (for Thr), whereas it is upfield by -0.4 ppm (for Thr) to -1.8 ppm (for Arg) in β-sheets[261]. $^{13}C_{\beta}$ secondary chemical shifts are smaller, and of opposite sign. The secondary chemical shift difference, defined by the difference between the secondary chemical shift of alpha and beta carbons, has been used as an even better structure predictor[262]. Using a joint probability calculation, Wang and Jardetzky[261] found that the reliability to discern α-helix from random coil using chemical shifts is in the order $^{13}C_{\alpha} > {}^{13}C' > {}^{1}H_{\alpha} > {}^{13}C_{\beta} > {}^{15}N > {}^{1}H_{N}$, whereas the order for β-sheet structure is $^{1}H_{\alpha} > {}^{13}C_{\beta} > {}^{1}H_{N} \sim {}^{15}N \sim {}^{13}C_{\alpha} \sim {}^{13}C'$. Algorithms have been developed to predict protein secondary structure using chemical shift data, such as the Chemical Shift Index program (CSI 3.0)[263] and the Secondary Structure Propensity program (SSP)[264] which is specifically designed for IDPs.

Mutual interactions between nuclear spins give rise to coupling, including scalar coupling and dipolar coupling. Dipolar couplings happen through space and are averaged out in isotropic solution. Scalar couplings, mediated by the electrons in covalent bonds, contain structural information because they depend on the intervening dihedral angle[265]. $^3J_{HNH\alpha}$ depends primarily on the φ backbone angle and has been used extensively to distinguish α from β secondary structure[266]. $^3J_{HNH\alpha}$ of α-helices are below 6 Hz, above 8 Hz for β-strand. IDPs have $^3J_{HNH\alpha}$ distributed in the range of 6 - 8 Hz. One-bond coupling constants $^1J_{C\alpha H\alpha}$ have also been used to validate IDP ensembles[267].

2.2. Nuclear spin relaxation, reduced spectral density mapping, and model-free analysis.

Protein dynamics on the ps-ns timescale has long been explored by nuclear spin relaxation measurements. This section will give a brief introduction on why and how fast protein motions can be interpreted by NMR observables (mostly by signal intensities) in relaxation experiments.

2.2.1. Nuclear spin relaxation and the spectral density.

In NMR, populations of nuclear quantum states and coherence among the populations are altered by applying a time – dependent magnetic field with a frequency ω_{RF} selected to satisfy the energy requirement of a transition. The magnetic field fluctuations are an excitation when the spins are forced into higher energy levels, and in this case, the system adopts an excited state. The excitation process in an NMR experiment is achieved by using RF pulses. To return to equilibrium, the excited system goes through a relaxation process that also requires a time-dependent magnetic field. Such a field is usually not intentionally created in NMR experiments, instead, it is spontaneously created by the random rotational motions of the molecule or by internal motions within the molecule. In proteins, there are two main mechanisms for the generation of oscillating fields from molecular motions: the chemical shift anisotropy (CSA) and dipolar couplings.

- CSA arises whenever the electronic environment around a nucleus is not the same in every direction, *i.e.* is anisotropic, such that a nucleus experiences different electronic shielding effects and, thus different effective magnetic fields, depending on the orientation of the molecule with respect to the external field. CSA contributes to relaxation, for example, when an amide nitrogen experiences a different local magnetic field as the protein rotates relative to the main magnet field B_0 due to Brownian tumbling in solution, and/or the bond vector rotates relative to the protein backbone.
- Dipolar coupling is a through-space interaction between pairs of nuclear spins that affects the magnetic field experienced by each spin. The distance and orientation of the pair of nuclei relative to B_0, which may change over time due to ps-ns motions, affects the strength of the dipolar field, leading to local magnetic field oscillations.

It needs to be noted that only temporal changes in the magnetic field strength that contain frequencies that match the nuclear resonance frequencies may cause relaxation. The contribution of protein motions, or equivalently the intensities of the oscillating magnetic fields, to relaxation as a function of frequency is represented by a spectral density function, $J(\omega)$.

Nuclear spin relaxation includes the restoration of an excited spin system to thermal equilibrium with its surroundings (spin – lattice or longitudinal relaxation), and the loss of coherence of transverse magnetization (spin – spin or transverse relaxation). Longitudinal relaxation occurs due to the net loss of energy from the excited state; while transverse relaxation occurs due to both the energy loss and the energy exchange between spins (Figure 2.3)

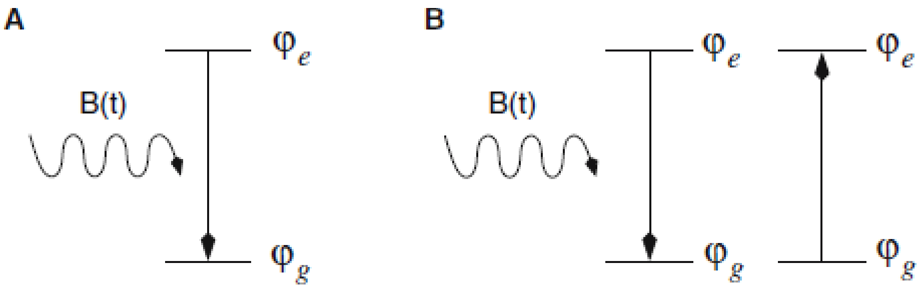


Figure 2.3. Relaxation of the excited state. Nuclear spin relaxation is a consequence of time-dependent fluctuations in the magnetic field. These fluctuations can cause the net loss of energy from the excited state (A) or enhance the rate of spin-spin exchange (B). The former contributes to both T_1 and T_2 relaxation while the latter only contributes to T_2 relaxation. Reprinted with permission from reference [242]*. Copyright © 2006 Springer.*

Longitudinal and transverse relaxation are characterized by time constants (T_1 and T_2, respectively), which are the times, on the average, required for a certain spin to relax to about 63% of its equilibrium state. Relaxation rates (R_1 and R_2) are reciprocals of the relaxation time constants. More details on longitudinal and transverse relaxation will be discussed below. NMR relaxation also investigates the relationship between dipolar coupling and the NOE effect which measures how perturbation of the ground and excited state populations of one spin affects the populations of another dipole-coupled spin, given that dipolar coupling is one of two main mechanisms whereby protein motions contribute to relaxation. T_1, T_2 (or equivalently R_1, R_2) and

NOE are the three most frequently measured parameters of protein NMR nuclear spin relaxation studies.

The relationships between the spectral density functions (J(ω)), which describes the oscillating magnetic fields, and the relaxation parameters (R_1, R_2 and NOE) for ^{15}N are as follow[268]:

$$R_1 = d\,[J(\omega_H - \omega_N) + 3\,J(\omega_N) + 6\,J(\omega_H + \omega_N)] + c\,J(\omega_N)$$

$$R_2 = R_2^0 + R_{ex}$$

$$R_2^0 = \frac{d}{2}[4\,J(0) + J(\omega_H - \omega_N) + 3\,J(\omega_N) + 6\,J(\omega_H) + 6\,J(\omega_H + \omega_N)] + \frac{c}{6}[4\,J(0) + 3\,J(\omega_N)]$$

$$NOE = 1 + \frac{\gamma_H}{\gamma_N}\frac{d}{R_1}[6\,J(\omega_H + \omega_N) - J(\omega_H - \omega_N)]$$

where:

- d, dipolar constant, is defined as:

$$d = \frac{1}{4}\left(\frac{\mu_0}{4\pi}\right)^2 \frac{(\hbar\boldsymbol{\gamma_N\gamma_H})^2}{r_{N-H}^6}$$

- c, chemical shift anisotropy constant, is defined as:

$$c = \frac{(\omega_N \Delta\sigma)^2}{3}$$

- μ_0, permittivity of free space,
- ħ, Planck’s constant divided by 2π,
- γ_N and γ_H, gyromagnetic ratios of ^{15}N and ^{1}H, respectively,
- $r_{N\text{-}H}$ = 1.02 Å vibrationally averaged effective N-H bond length,

- $\Delta\sigma$ = -172 ppm, chemical shift anisotropy,
- ω_N and ω_H, Larmor frequency of ^{15}N and ^{1}H in rad^{-1}
- R_2^0, pure transverse relaxation rate, not influenced by exchange processes.
- R_{ex}, exchange rate. For a system exchanging between 2 states A and B:

$$A \underset{k_B}{\overset{k_A}{\rightleftharpoons}} B$$

where k_A and k_B are exchange rates between the states, and R_{ex} is given by:

$$R_{ex} \approx \frac{p_A(1-p_A)k_{ex}}{1+(k_{ex}/\Delta\omega)^2}$$

p_A and p_B are populations of states, and $\Delta\omega$ is the chemical shift difference between resonances in the two states.

The relaxation time constants are measured by NMR experiments in which the spin system at thermal equilibrium is forced into an excited state by RF pulses; after that, the bulk magnetization of the system is allowed to evolve in a defined amount of time, called a relaxation delay, before detection. Figure 2.4 shows the behavior of the magnetization in 1D T_1 and T_2 experiments. At equilibrium before any RF pulse is applied, the spin system of the sample is subjected to a static magnetic field B_0. The magnetic moments of spins are aligned in such a way that when their contributions are all added up, there is a net magnetic moment, called bulk magnetization, along the direction of the applied field B_0. The bulk magnetization arises from the slight difference in the population of spins that are aligned with and against the field.

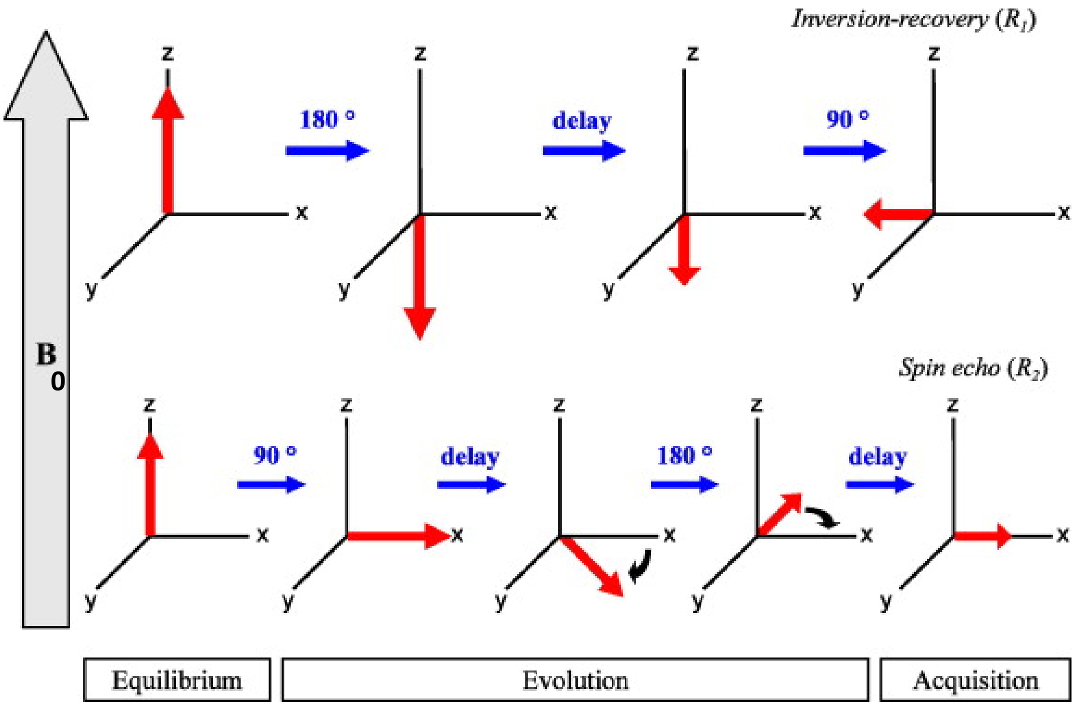


Figure 2.4. Experimental approaches to the recording of R_1 (inversion-recovery) and R_2 (spin echo). Bulk magnetization is shown as an arrow on the Cartesian coordinate system. Varying the delays allows the recording of signal amplitudes for relaxation rate measurements. In the case of the spin-echo experiment, both delays are of equal duration. This figure does not take into account details other than longitudinal and transverse relaxation. Reprinted with permission from reference [268]*. Copyright © 2010 Elsevier.*

T_1 measurement generally uses the inversion-recovery approach. The experiment starts with a 180^0 pulse, placing the magnetization along the z axis. This serves as a perturbation pulse to excite the spin system. After a delay during which the spins relax, a 90^0 pulse brings the magnetization into the transverse plane (xy plane) for acquisition. Varying the delay between the two pulses allows the system to relax to different extents, leading to different magnetization amplitudes, or, equivalently, different signal intensities for each delay. The T_1 value is extracted from fits of the decaying signals as a function of the delay times. For protein dynamics, the pulse sequence for T_1 measurements is derived from a sequence originally proposed for the recording of ^{1}H-^{15}N correlation spectra[269]. This approach has higher resolution to overcome the problem of overlapping peaks. In this case, the perturbation period increases the polarization of the pure nitrogen longitudinal magnetization *via* 2 INEPT (Insensitive Nuclei Enhanced by Polarization Transfer) periods. INEPT is a signal resolution enhancement method used in NMR spectroscopy. It transfers nuclear spin polarization from spins with a larger Boltzmann population difference to the spins of interest having a lower Boltzmann population difference using J-coupling[270]. During the delay time, the nitrogen magnetization relaxes toward thermal equilibrium and it will then be transferred back to a proton for detection. The intensity of the peaks at different delay times will follow an exponential decay function:

$$I_t = I_0\, e^{\frac{-t}{T_1}}$$

where I_t is the peak intensity with delay time t, and I_0 is the peak intensity at time 0.

An NMR experiment to measure T_2 generally uses the spin-echo approach implemented in a pulse sequence that is also a variation of the sequence originally proposed for the recording of ^{1}H-^{15}N correlation spectra[269]. The effect on the bulk – magnetization of the spin echo pulse sequence in a 1D experiment is depicted in Figure 2.4. The experiment starts with a 90^0 pulse, placing the magnetization in the transverse plane (xy plane) and creating coherence among the spins, or in another words, bringing the spins into phase. After a delay, a 180^0 pulse “flips” the magnetization, putting it on the other side of the x axis. The magnetization is then allowed to evolve for a further delay of equal length compared with the first delay. An echo is created where signals from different chemical shifts (*i.e.* different angular frequencies) are refocused to the same position and coherence is recovered. Recording proceeds immediately after the second delay.

Transverse relaxation processes occur during the two delays. Varying the length of the relaxation delays will affect the signal intensity. Similar to T_1, intensity of peaks at different delay times also follow an exponential decay function:

$$I_t = I_0\, e^{\frac{-t}{T_2}}$$

where I_t is the peak intensity with delay time t, and I_0 is the peak intensity at time 0. In practice, for both T_1 and T_2 measurements, a series of spectra with varied relaxation delays are acquired, and peak intensities are fit to the exponential models to extract the relaxation time constants.

The steady – state heteronuclear NOE, which arises from cross-relaxation between two dipolar – coupled spins (^{15}N and ^{1}H), is generally measured by calculating the ratio of the peak intensities in two 2D spectra acquired with and without proton pre-saturation.

$$NOE = \frac{I_{sat}}{I_{eq}} - 1$$

In order to extract dynamics information, the measured ^{15}N relaxation data (T_1, T_2 and NOE) can be analyzed using several approaches. A first quick analysis of the relaxation data, yields valuable qualitative information on the global and local conformational properties of the protein. Figure 2.5 shows an example of a quick analysis of the relaxation data of Myosin Phosphatase Targeting subunit 1 (MYPT-1) in solution. The protein has a disordered region on the N-terminus which is more flexible in comparison with the folded ankyrin-repeat regions on the C-terminus. The raw relaxation parameters clearly show the differences in flexibility of MYPT-1 protein domains where the N-terminus has clearly lower R_2 relaxation rates (Panel B) and NOE values (Panel C). Within the disordered region, relaxation data also reflect the populations of conformations such that the short segment that forms a transient helix also has higher relaxation rates and NOE values than the rest of the region.

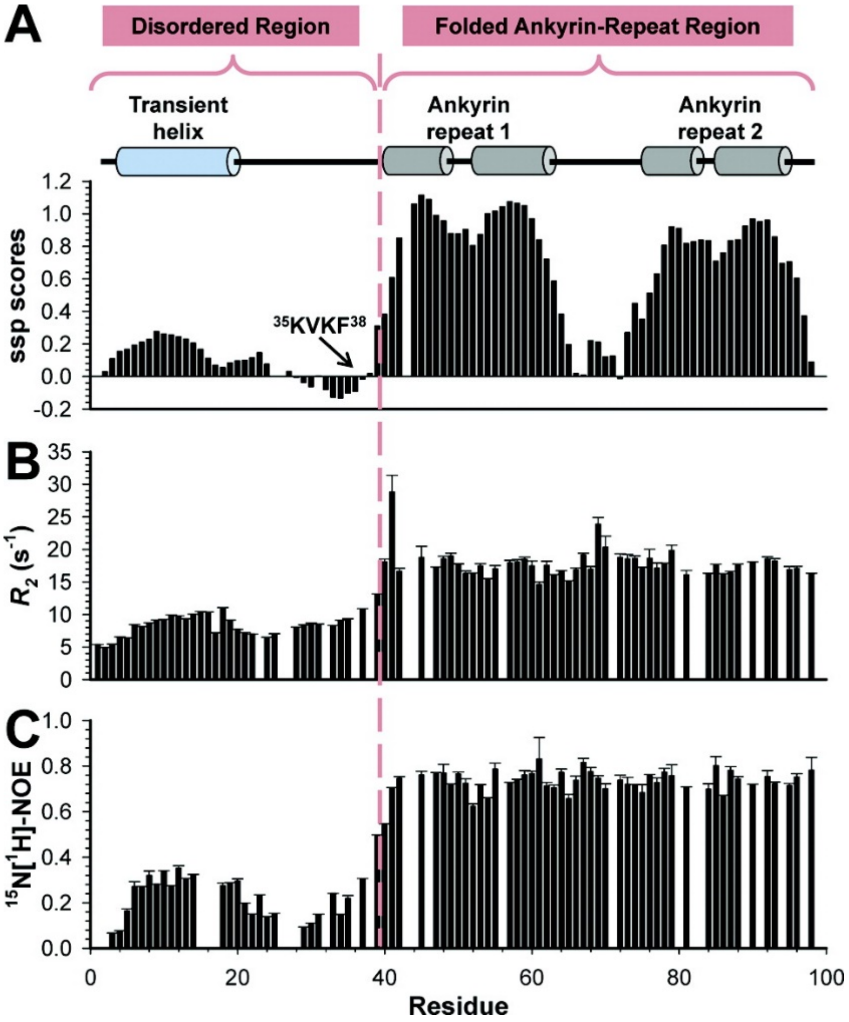


Figure 2.5. (A) Secondary structure propensity scores, (B) experimental R_2 relaxation rates, and (C) heteronuclear ^{15}N[^{1}H]-NOE measurements demonstrate the two-domain behavior of MYPT1$_{1-98}$ in solution. The cartoon representations above the SSP scores indicate the presence of secondary structure based on the MYPT1−PP1 complex structure. A dashed line separates the two different regions of MYPT1$_{1-98}$, the N-terminal flexible region and the C-terminal folded ankyrin-repeats, which constitute the MYPT1 PP1-binding domain. An element of transient

structure and reduced backbone motion within residues 5−17 in the N-terminal disordered region can be readily identified by SSP scores greater than ~0.2, elevated R_2 rates and $^{15}N[^1H]$-NOE values. This transient helix is colored in blue, which differentiates it from the fully populated α-helices of the ankyrin-repeat domain, colored in gray. Reprinted with permission from reference [271]*. Copyright © 2011, American Chemical Society.*

Additionally, the T_1/T_2 ratio provides information on the local protein rigidity[272]. For folded proteins, it is used estimate a protein's overall rotational correlation time. For IDPs, T_1/T_2 allows to distinguish regions having significant structural propensities, characterized by longer effective correlation times, from segments lacking any residual structure, characterized by shorter effective correlation times.

There are other approaches to extensively analyze relaxation data including the model-free formalism[221, 222] and reduced spectral density mapping[273, 274]. The following sections will provide a discussion of these methods.

2.2.2. Model-free analysis.

Model-free formalism was introduced by Lipari and Szabo[221, 222], further extended by Clore *et al.*[275] and optimized in terms of model selection by d'Auvergne and Gooley[276, 277]. It has been the preferred approach for protein spin relaxation data analysis for allowing the extraction of parameters with more physical meaning than the relaxation data alone. The approach relies on an assumption that protein internal motions can be decoupled from the global tumbling of the molecule, given that these two stochastic processes are separated by at least an order of magnitude (with global tumbling generally lying between a few ns to tens of ns, whereas local motions are on the order of tens or hundreds of ps). Autocorrelation functions are used to define and decouple global and internal motions:

$$C(\tau) = C_0(\tau)\, C_I(\tau) \qquad (1)$$

where C(τ) is the overall autocorrelation function, $C_0(\tau)$ is the autocorrelation function for overall tumbling, and $C_I(\tau)$ is the autocorrelation function for internal motions. NMR relaxation of the protein backbone is often determined by the fluctuations of the ^{15}N-^{1}H vectors with respect to the external magnetic field, and the correlation function describes the dynamic tumbling of these dipoles. It represents the autocorrelation between the orientation of the dipole vector at time 0 and its orientation at time t. Fourier transformation of an autocorrelation function is another function that represents the contributions of the motions, as a function of frequency, to the reorientation of relaxation vectors and this resulting function is the spectral density function as mentioned above.

Overall rotational motion of a protein in solution is stochastic, thus it is reasonable to assume that the autocorrelation function has the form of an exponential decay during protein relaxation. According to the work by d'Auvergne[278], the function can be written as:

$$C_0(\tau) = \frac{1}{5}\sum_{i=-k}^{k} W_i e^{-\frac{\tau}{\tau_i}} \qquad (2)$$

where

- W defines a weighting factor,
- the index i ranges over the number of exponential terms,
- and k, depending on the complexity of the diffusion tensor, may take value of 0, 1 or 2 if the global model of the protein is spherical, spheroidal, or ellipsoidal, respectively.

In the case of isotropic tumbling, or if the protein has a spherical shape in solution, the equation is reduced to $C_0(\tau) = \frac{1}{5}e^{-\frac{\tau}{\tau_m}}$, where τ_m is the global rotational correlation time constant.

A simple correlation function for a protein with internal motion, introduced in the original model-free theory by Lipari and Szabo[221, 222], and the extended version for 2 internal motions, proposed by Clore *et al.*[275], are respectively:

$$C_I(\tau) = S^2 + (1 - S^2)e^{-\frac{\tau}{\tau_e}} \qquad (3)$$

$$C_I(\tau) = S^2 + \left(1 - S_f^2\right)e^{-\frac{\tau}{\tau_f}} + \left(S_f^2 - S^2\right)e^{-\frac{\tau}{\tau_s}} \qquad (4)$$

where

- S^2 is the square of generalized order parameter, and $S^2 = S_f^2 . S_s^2$,
- τ_e is the effective correlation time of the internal motion,
- and the subscripts s and f refer to the slower and the faster of the two internal motions, respectively.

The internal motion of the relaxation vector (the NH bond vector) is restricted, so the correlation function of the internal motion is not decaying to zero, but to S^2, when $\tau \rightarrow \infty$. The order parameter describes the restriction of fast motions. S^2 is limited to values between 0 (completely unrestricted motion) and 1 (completely restricted, absence of motion). The higher the order parameter, the more restrained the N–H bond motion is by its environment, *i.e.*, the more tightly the proteins is packed. Empirically, S^2 values for backbone amide ^{15}N sites are found to be >0.8 in the secondary structures and between 0.5 and 0.8 for loops, turns and termini[279]. The correlation time constants account for the effective timescale of the motions. The overall rotational correlation time represents the time required for the molecule to tumble through one radian in an arbitrary direction. The overall rotational correlation time provides information about the size and shape of the molecule, indicating the presence of quaternary structure and providing useful comparisons to other physical measurements such as viscosity, sedimentation data, fluorescence anisotropy, and gel filtration retention times. The internal correlation time measures the timescale for internal motions of bond vectors sweeping through an amplitude quantified by the order parameter and thus reflects both the frequency and amplitude of the internal motion. This parameter is extracted from model-free analysis with low precision for restricted residues (S^2>0.8), and usually is not analyzed in detail[280].

The general spectral density function is derived by Fourier transforming the autocorrelation function (1), where the autocorrelation function of rotational motion takes the general form in equation (2), and the autocorrelation function of internal motion takes the extended form in equation (4):

$$J(\omega) = \frac{2}{5}\sum_{i=-k}^{k} W_i\tau_i\left(\frac{S^2}{1+(\omega\tau_i)^2} + \frac{(1-S_f^2)(\tau_f+\tau_i)\tau_f}{\left(\tau_f+\tau_i\right)^2+\left(\omega\tau_f\tau_i\right)^2} + \frac{\left(S_f^2-S^2\right)(\tau_s+\tau_i)\tau_s}{(\tau_s+\tau_i)^2+(\omega\tau_s\tau_i)^2}\right) \qquad (5)$$

The general spectral density function (5) can be reduced to the original spectral density function proposed in the original model-free theory by Lipari and Szabo[221, 222] when the protein is tumbling isotropically and has only one internal motion:

$$J(\omega) = \frac{2}{5}\left(\frac{S^2\tau_m}{1+(\tau_m\omega)^2} + \frac{\left(1-S^2\right)\tau}{1+(\tau\omega)^2}\right), \text{ where } \tau^{-1} = \tau_m^{-1} + \tau_e^{-1}, \qquad (6)$$

or to the extended version by Clore *et al.*[275] when the protein is tumbling isotropically and has two internal motions:

$$J(\omega) = \frac{2}{5}\left(\frac{S^2\tau_m}{1+(\tau_m\omega)^2} + \frac{\left(1-S_f^2\right)\tau_f'}{1+\left(\tau_f'\omega\right)^2} + \frac{(S_f^2-S^2)\tau_s'}{1+\left(\tau_s'\omega\right)^2}\right) \qquad (7)$$

where $\tau_s'^{-1} = \tau_m^{-1} + \tau_s^{-1}$ and $\tau_f'^{-1} = \tau_m^{-1} + \tau_f^{-1}$.

Equation (6) and (7) are derived in the absence of a physical model delineating all trajectories and microstates of the internal motions of the system. These motions are solely characterized by the model-independent order parameters and correlation time constants. By assuming certain order parameters or correlation times to be statistically negligible, either being one or zero respectively, a number of model-free models can be constructed.

m0: ( ),

m1: (S^2),

m2: (S^2, τ_e),

m3: (S^2, R_{ex}),

m4: (S^2, τ_e, R_{ex}),

m5: (S^2, S_f^2, τ_s),

m6: (S^2, τ_f, S_f^2, τ_s),

m7: (S^2, S_f^2, τ_s, R_{ex}),

m8: (S^2, τ_f, S_f^2, τ_s, R_{ex}),

m9: (R_{ex}),

The models of model-free motions above are constructed by combining the parametric restrictions, in which statistically insignificant parameters are dropped, together with the addition of a parameter accounting for chemical exchange relaxation. An order parameter is neglected and dropped when the motion is statistically insignificant. In the spectral density function, it takes the value of 1, meaning the motion is restricted. In the case of the correlation time constant, when the motion is too fast for this parameter to be reliably extracted, the time constant will take the value of zero in the spectral density function. The local models m0 – m4 are used with the model described in equation (6), and the local models m5-m9 are used with the model described in equation (7). The special case of local model m0 corresponds to the situation whereby no internal motions are statistically significant.

To extract information about protein motion, model-free parameters of each residue, including the order parameters and correlation time constants, are extracted by fitting relaxation data to a system consisting of a spectral density function (equation (6) or (7)), and the functions that describe the relationship between T_1, T_2, NOE and spectral density function in the previous section (2.2.1). An appropriate combination of rotational motion model (sphere, spheroid, or ellipsoid) and internal motion model (models m0 to m9) that correctly describes the system need to be chosen. A protocol for model selection and minimization was proposed by d'Auvergne and Gooley, and it was applied in a program, called Relax[281], to analyze relaxation data.

For IDPs, the important assumption of model-free analysis regarding the separation of overall and internal motions may not hold true due to the structural flexibility of IDPs on the ps-ns timescales, and this affects the accuracy of model-free analysis. Local reorientations of bond vectors, due to conformational changes, may occur on timescales similar to that of the overall rotational motions, making the statistical mutual independence of the two motion-types less plausible. Additionally, one single overall diffusion frame is not sufficient to describe the rotational motion of IDPs. At best, this overall diffusion tensor would only correspond to an average over a set of time-dependent diffusion tensors[282]. A modified model-free approach can still be applied for IDPs by using effective residue - specific or segmental correlation times that vary over the polypeptide chain instead of a global correlation time, thus an overall diffusion tensor can be avoided. However, defining a single local correlation time per residue or per segment is facing two issues. Firstly, the ensemble of conformational states is large for an IDP and it may consist of both compact and extended conformers with different overall rotational times. Secondly, an IDP will have many modes of motion[283], each of which has its own correlation time. For example, in one peptide segment all residues will move in a correlated fashion, but even segments that are distant in the sequence might move correlated because they (temporarily) interact mutually. Further improvement has been made by introducing a distribution function for the correlation times that is applied to the spectral density function. The spectral density function in Lipari and Szabo's work will become[284]:

$$J(\omega) = C(0) \int_0^\infty f(\tau_C) \frac{\tau_C}{1 + (\omega \tau_C)^2} d\tau_C$$

This method was successfully applied in the studies of disordered proteins (see reference [285] and reference [286]). However, the application of a distribution function in this approach also introduces a physical bias because one must choose a mathematical function to describe the distribution. Modig *et al.*[284] recognized this limitation and proposed the Model Independent Correlation (MIC) time distribution. This approach has distinct advantages for it completely avoids the need to specify a functional form of the correlation time distribution, and the requirement of statistical independence of three types of motions. However, one issue that still remains and that is justifying increasing the number of correlation times or distributions in either approach based on the empirical information available from nuclear spin relaxation which is limited.

2.2.3. Reduced spectral density mapping.

As indicated above, the relationship between the spectral density values and the relaxation parameters is described by the following functions:

$$R_1 = d\,[J(\omega_H - \omega_N) + 3\,J(\omega_N) + 6\,J(\omega_H + \omega_N)] + c\,J(\omega_N)$$

$$R_2^0 = \frac{d}{2}[4\,J(0) + J(\omega_H - \omega_N) + 3\,J(\omega_N) + 6\,J(\omega_H) + 6\,J(\omega_H + \omega_N)] + \frac{c}{6}[4\,J(0) + 3\,J(\omega_N)]$$

$$R_2 = R_2^0 + R_{ex}$$

$$NOE = 1 + \frac{\gamma_H}{\gamma_N}\frac{d}{R_1}[6\,J(\omega_H + \omega_N) - J(\omega_H - \omega_N)]$$

Spectral density mapping is a method allowing the quantification of motions by reconstructing the site-specific J(ω) of the coupled ^{15}N-^{1}H spin systems over 5 frequencies, ω_N, ω_H, ω_H - ω_N and ω_H + ω_N and zero-frequency. It is not possible to unambiguously evaluate the 5 unknown spectral density values from only 3 measured relaxation parameters; at least 5 independent parameters would be needed. In order to overcome the limitation, Peng and Wagner proposed an expanded set of 6 relaxation parameters, including T_1, T_2, NOE and 3 more parameters, allowing the spectral density values at all five frequencies to be experimentally and uniquely determined[287, 288]. However, experiments to measure these rates can lead to multi-exponential decays from which relaxation rates are difficult to extract with precision.

The insight that led to the spectral density mapping method is the observation that the spectral density reaches a constant value at higher frequencies. Thus, the three spectral densities values at three highest frequencies (ω_H, ω_H - ω_N and ω_H + ω_N) are combined into a single value $\langle J(\omega_H)\rangle$, the spectral density at the effective proton frequency[273, 274, 289]. This corresponds to J(0.87ω_H) as described in Farrow *et al.*[273].

The reduced spectral density mapping approach determines the spectral density values at 3 frequencies, the zero frequency J(0), the nitrogen frequency $J(\omega_N)$, and the effective proton frequency $\langle J(\omega_H)\rangle$. The dependence of the three reduced spectral densities on the relaxation parameters T_1, T_2 and NOE are:

$$J(0) = \frac{-1.5}{3d+c}\left(\frac{R_1}{2} - R_2 + 0.6\sigma_{NOE}\right)$$

$$J(\omega_N) = \frac{1}{3d+c}(R_1 - 1.4\sigma_{NOE})$$

$$\langle J(\omega_H)\rangle = \frac{\sigma_{NOE}}{5d}$$

where σ_{NOE}, the cross-relaxation rate, is defined as follows:

$$\sigma_{NOE} = (NOE - 1)R_1\frac{\gamma_N}{\gamma_H}$$

Results from spectral density mapping indicates whether the motions of a particular bond vector are dominated by high- or low- frequency oscillations. It also helps to discriminate between motions either faster, slower or on a similar timescale than the global tumbling. Křížová *et al.* proposed a graphical approach to analyze spectral density mapping by plotting either $J(\omega_N)$ or $\langle J(\omega_H)\rangle$ as a function of J(0) and comparing the distribution to theoretical functions describing the simple case of a single motion defined as follows[290].

$$J(\omega_N) = \frac{J(0)}{1 + 6.25(\omega_N J(0))^2}$$

$$\langle J(\omega_H)\rangle = \frac{J(0)}{1 + 6.25(\omega_H J(0))^2}$$

The curve in Figure 2.6 was generated from the first equation above. Point P represents a residue in which relaxation is dominated by motion on the global tumbling timescale (ns); point Q represents a residue where relaxation is dominated by motion on a much faster internal correlation timescale (ps). Most relaxation data in proteins should be distributed between points P and Q around the connecting line, such as point R, depending on the contributions of the motions to the relaxation. When slower motions (μs-ms) and conformational exchange are present, the data are shifted to higher J(0) values in the direction of the red arrow toward point S.

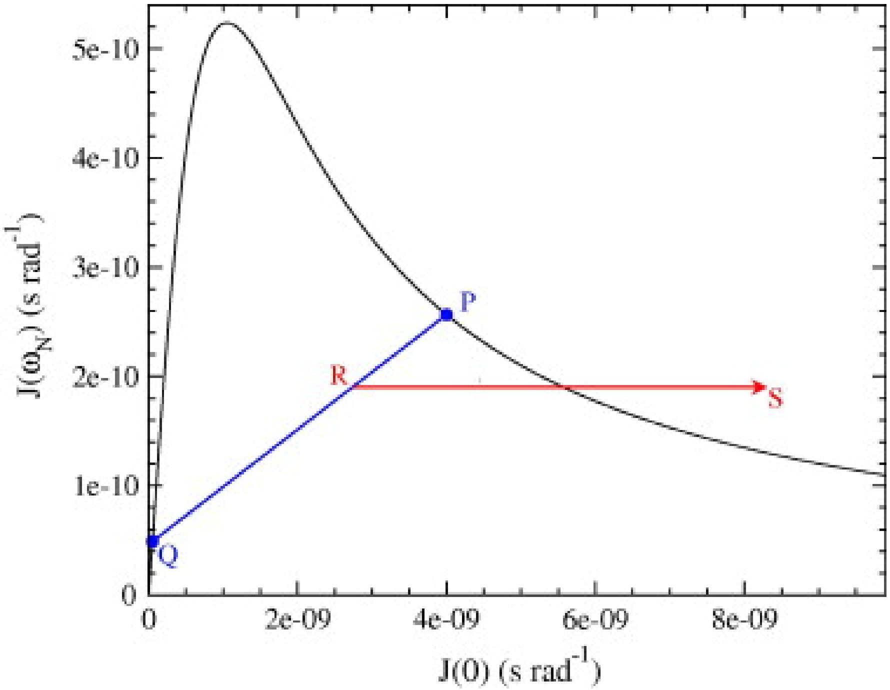


Figure 2.6. Graphical analysis of reduced spectral density mapping. Reprinted with permission from reference [268]*. Copyright © 2010 Elsevier.*

2.3. Relaxation dispersion.

Protein dynamics on slow timescales (tier-0 dynamics, Figure 1.18) are associated with many aspects of function, including molecular recognition and signaling, folding, enzymatic catalysis, and allostery[180, 291]. They normally involve the inter-conversion of low-populated and transient conformers that are commonly found in IDPs. NMR relaxation dispersion is a solution-based NMR technique that can provide atomic resolution information on millisecond timescale conformational transitions in proteins, including the effective relaxation rates, the populations of conformers, and the chemical shift differences and rates of exchange between the conformers. This section will provide a brief review of the effects of conformational exchange events on NMR spectra, introduce the NMR relaxation dispersion experiment using a magnetization vector model and discuss the analysis of relaxation dispersion data.

2.3.1. Effect of protein conformational exchange on NMR spectra.

Consider the exchange between the populated ground state A, and an excited state B of a protein:

$$A \underset{k_{BA}}{\overset{k_{AB}}{\rightleftharpoons}} B$$

with the exchange rate defined as $k_{ex} = k_{AB} + k_{BA}$ where k_{AB} is the rate constant for the conversion of A to B, and k_{BA} is the rate constant for the reverse reaction. The chemical shift difference between the two states is $\Delta\omega = \omega_A - \omega_B$. The populations of each conformer, p_A and p_B follow the Boltzmann distribution.

The relationship between the rate of exchange, k_{ex}, and chemical shift difference, $\Delta\omega$, will affect the observed spectrum as summarized in Table 2.1. Figure 2.7 shows the results of a simulation, conducted by Palmer *et al.*[292] on the effects of the exchange process on NMR lineshape. The integrated intensity of each line is equal to the population of spins in conformation A and B, respectively. For the very fast exchange limit, the spins do not exist in either states (A and B) long enough to give rise to separate resonance frequencies. Thus, on the spectrum, a single resonance line is observed representing the averaged resonance frequency. At the other limit, under the very slow exchange condition where the process is incapable of averaging the chemical shifts while the spins are precessing, two resonance lines are observed, representing the fractions of spins that are found in each conformation. When the exchange rate is similar to the chemical shift difference, the exchange causes significant line-broadening of signals from both states. The low-populated excited state resonance usually becomes too broad to be observed (i and j on the right panel of Figure 1.26). However, information about this invisible state can still be extracted from the broadened signal of the ground state by the relaxation dispersion method.

Table 2.1. Summary of the effects of chemical or conformational exchange on the properties of the NMR spectrum. The chemical shift separation is defined as $\Delta\nu = \frac{1}{2\pi}(\omega_A - \omega_B)$. *The timescale of exchange is characterized by the dimensionless parameter α* [242, 293]*, defined as* $\alpha = \frac{2/(k_{ex}/\Delta\omega)^2}{1+(k_{ex}/\Delta\omega)^2}$. *Adapted with permission from reference* [242]*. Copyright © 2006 Springer.*

Exchange Rate		*α*	*Observed spectrum*
Very slow	$k_{ex} << \Delta\nu$	0	Two resonances
Slow	$k_{ex} < \Delta\nu$	<1	Two broadened resonances
Intermediate	$k_{ex} \approx \Delta\nu$	1	Complex lineshape
Fast	$k_{ex} > \Delta\nu$	>1	Single broadened resonances
Very fast	$k_{ex} >> \Delta\nu$	2	Single resonance

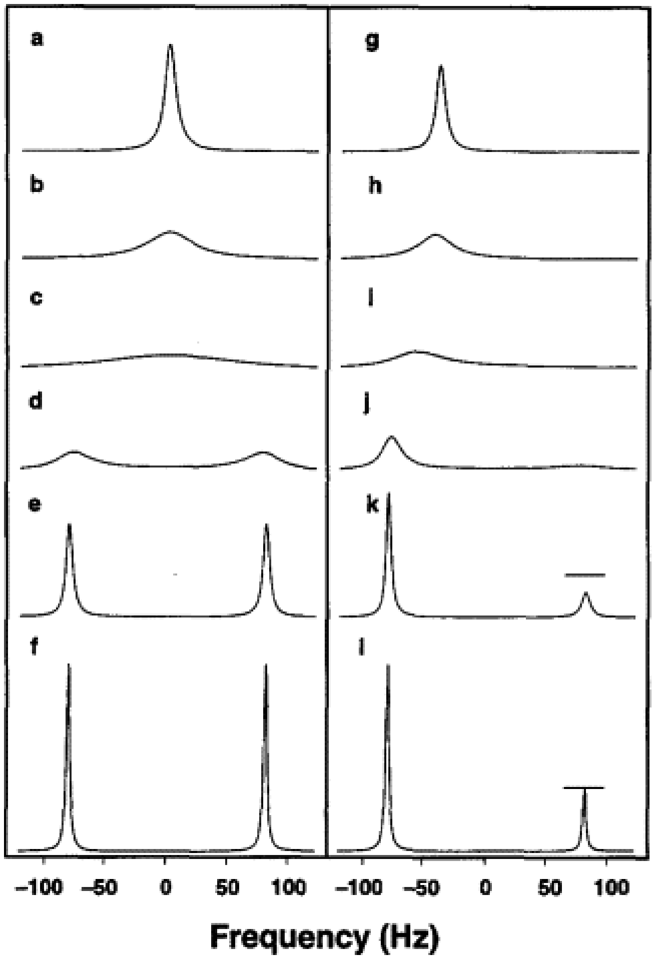


Figure 2.7. Chemical or conformational exchange lineshapes. (a–f) Symmetric exchange with $p_A = p_B = 0.5$. (g–l) Exchange with skewed populations $p_A = 0.75$ and $p_B = 0.25$. Values of k_{ex} are (a, g) 10,000, (b, h) 2000, (c, i) 900, (d, j) 200, (e, k) 20, (f, 1) 0.0 s^{-1}. In (k) and (1), the horizontal bar is drawn at 1/3 the height of the larger resonance. The spectra are simulated with $R^0_{2_A} = R^0_{2_B} = 10\ s^{-1}$ *and $\Delta\omega$ = 180 rad/sec. Reprinted with permission from reference* [292]*. Copyright © 2001 by Academic Press.*

2.3.2. Carr - Purcell - Meiboom - Gill (CPMG) pulse train and the relaxation dispersion experiment.

Relaxation dispersion uses the CPMG pulse train, which is built upon the idea of using a series of spin-echo pulses, to refocus inhomogeneous broadening of the nuclear spins[291, 294, 295]. First let's consider the case of a protein in solution that exists in one single conformation. After insertion into the external magnetic field of the NMR instrument, the bulk magnetization (red arrow in Figure 2.8) aligns with the field (z-axis). After the 90^0 pulse, it is rotated into the x-y plane and aligns with the x-axis. Because of the intrinsic transverse relaxation, after a short evolution period, the magnetization loses coherence, represented by the spreading colored vectors. The coherence can be restored by the spin-echo pulse as described in section 2.2.1, resulting in the refocusing of magnetization on the y-axis for detection.

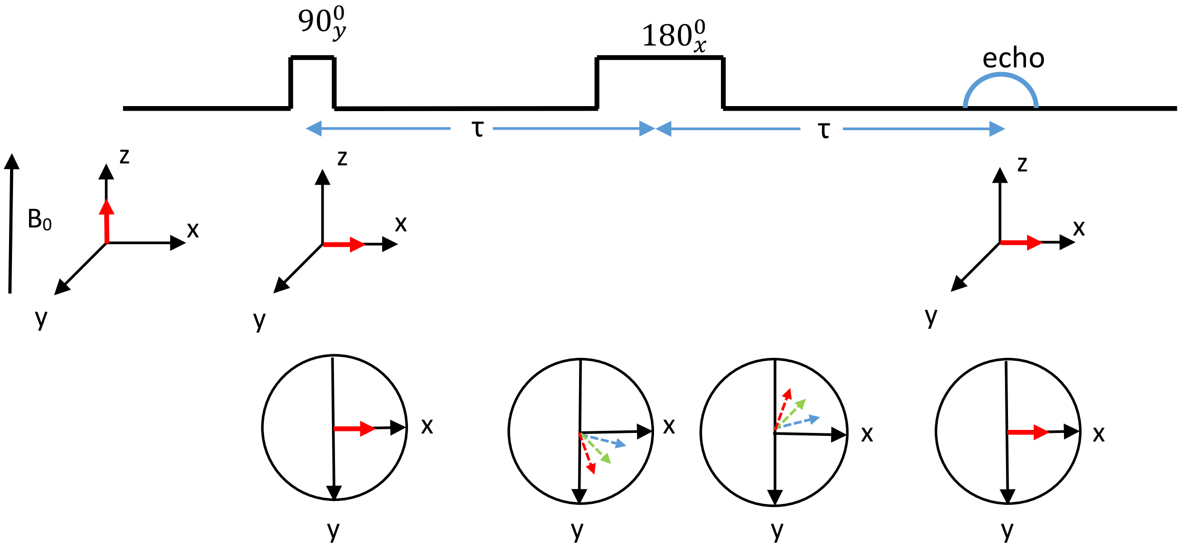


Figure 2.8. Behavior of protein magnetization where the protein exists in a single conformation during a spin-echo pulse sequence. The protein NMR sample is subjected to a static external magnetic field B_0. The spin-echo pulse sequence at the top of the figure consists of a 90^0

pulse that rotates the protein magnetization around the y-axis, represented as a small square, and a 180^0 pulse represented as a large square that "flips" the magnetization around the x-axis,. The protein magnetization is refocused after a delay time τ that is equal to the separation time between the two pulses. Evolution of the protein magnetization in a three dimensional Cartesian coordinate and on the x-y plane are depicted in the middle and at the bottom of the figure, respectively.

Now consider the case of a protein that exists in two discrete conformations, which do not interconvert. The bulk magnetization for each conformation is represented by a vector, red for conformation A and blue for conformation B in Figure 2.9. At equilibrium in the presence of a magnetic field B_0, both magnetization vectors align along the z-axis. For each conformation, the magnetization decoherence caused by the intrinsic transverse relaxation will be eliminated at the end of the applied spin-echo pulse sequence, therefore it is ignored in the following analysis. The 90^0 pulse puts both magnetizations along the x-axis. During the evolution period, each magnetization rotates around the z-axis with a characteristic angular frequency, accruing different phases $\varphi_A(t) = \omega_A t$ and $\varphi_B(t) = \omega_B t$. Next, a 180^0 pulse is applied to put the magnetization on the other side of the x-axis, while maintaining the direction of precession. After the second evolution period, magnetizations of both conformations return to the x-axis for detection, giving rise to 2 resonance lines. The intensities of the resonance lines provide a true readout of the relative populations of the conformations.

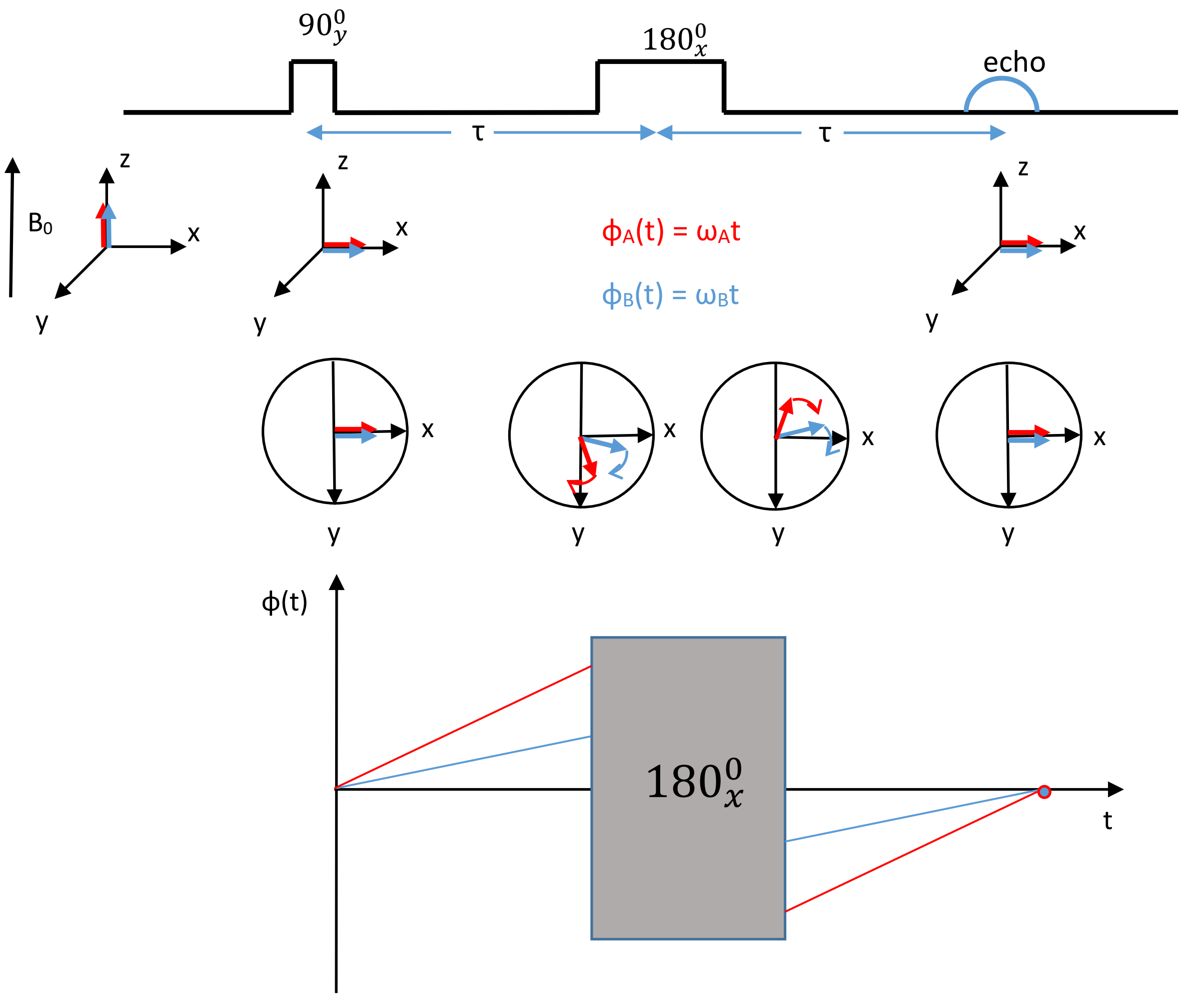


Figure 2.9. Behavior of protein magnetization where the protein exists in two conformations (A in red, and B in blue) not undergoing any exchange during a spin-echo pulse. The protein NMR sample is subjected to a static external magnetic field B_0. The spin-echo pulse sequence at the top of the figure consists of a 90^0 pulse represented as a small square that rotates the magnetizations of the two conformations around the y-axis and a 180^0 pulse represented as a large square that "flips" the magnetization around the x-axis. The protein magnetizations are refocused after a delay time τ equal to the separation time between the two pulses. Evolution of the protein magnetization in a three dimensional Cartesian coordinate and on the x-y plane is

depicted in the middle of the figure. The bottom panel shows the accrued phases of the magnetization over time. The gray box here represents the 180^0 pulse.

When exchange between the conformations is considered (Figure 2.10), the precession frequency of an NMR spin may stochastically vary between ω_A and ω_B. This makes the accrued phases during the two evolution periods, before and after the 180^0 pulses, different, leading to an incomplete refocusing of the magnetizations, so that the resonance lines of the 2 conformations are broadened and the intensities weaken. Thus, the exchange process also contributes to the relaxation process since it causes decoherence among the spins. The refocusing by the spin-echo element for an individual spin can be achieved to different extents because each spin executes a different trajectory (different times in each state, different numbers of conversions). Assuming $\omega_A > \omega_B$, the phase that each spin accrues after a τ – $180^0{}_x$ – τ block ($\varphi(2\tau)$) will be distributed within the range:

$$(\omega_A - \omega_B)\tau > \varphi(2\tau) > (\omega_B - \omega_A)\tau \qquad (8)$$

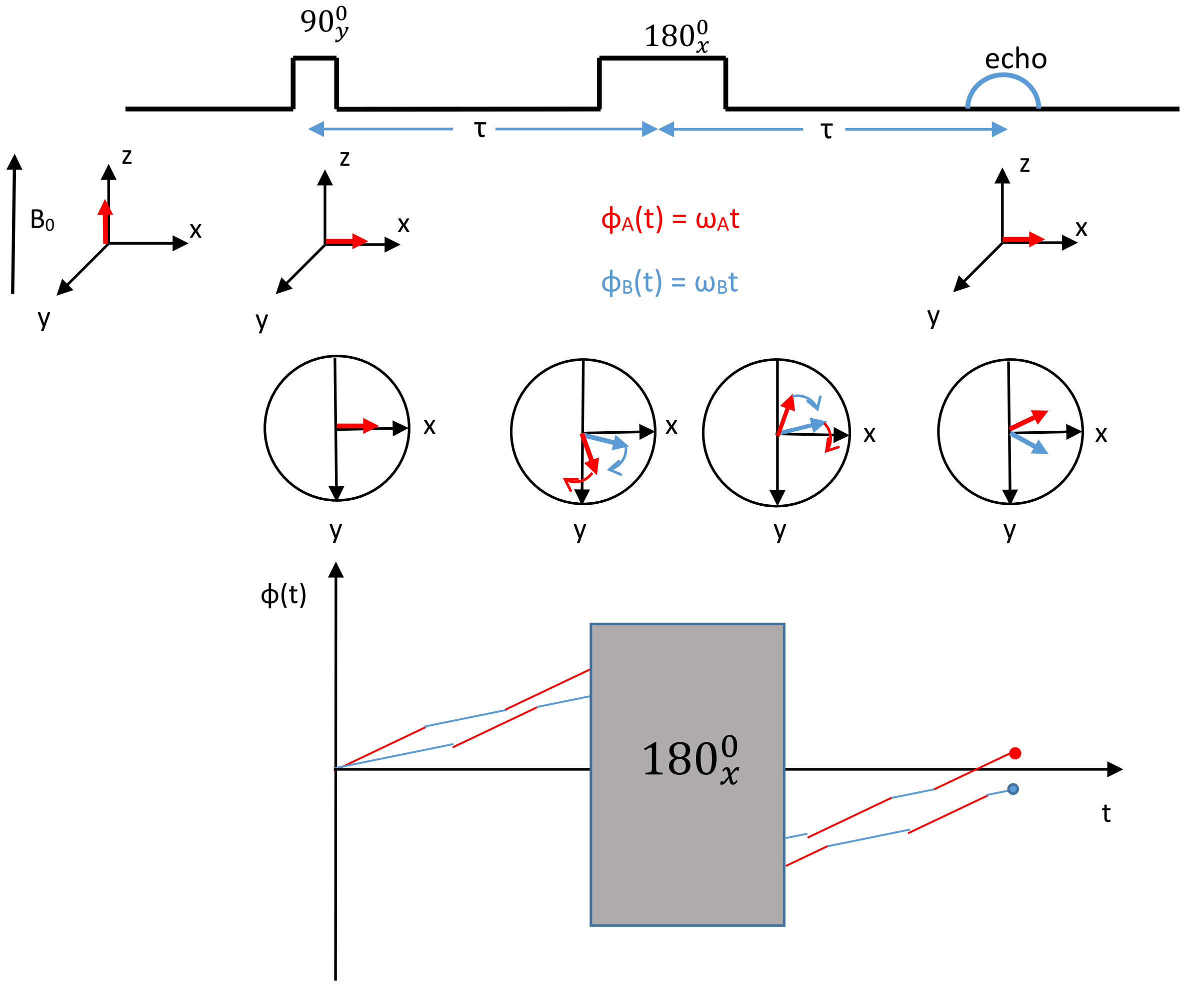


Figure 2.10. Behavior of protein magnetization where the protein exists in two conformations (A in red, and B in blue) in exchange on the same timescale as the repetition rate of the interpulse delays during a spin-echo pulse sequence. The protein NMR sample is subjected to a static external magnetic field B_0. The spin-echo pulse sequence at the top of the figure consists of a 90^0 pulse represented as a small square that rotates the magnetizations of the two conformations around the y-axis, and a 180^0 pulse represented as a large square that "flips" the magnetization around the x-axis. The protein magnetizations are not refocused at the end of the pulse sequence due to stochastic interconversion between the two conformations. Evolution of the protein magnetization in a three dimensional Cartesian coordinate and on the x-y plane is depicted

in the middle of the figure. Bottom panel shows the accrued phases of the magnetization over time. The gray box here represents the 180^0 pulse.

When multiple τ – 180$^0{}_x$ – τ blocks are applied within an amount of time, called the CPMG relaxation time τ_{CPMG}, we have a CPMG pulse train. If N τ – 180$^0{}_x$ – τ blocks are applied, the evolution time for each block will be calculated as $\tau = \frac{\tau_{CPMG}}{2N}$. By substituting this evolution time into equation (8), the phase that each spin accrues after one spin-echo block (φ(2τ)) is within the range:

$$(\omega_A - \omega_B)\frac{\tau_{CPMG}}{2N} > \varphi(2\tau) > (\omega_B - \omega_A)\frac{\tau_{CPMG}}{2N} \qquad (9)$$

Here we define a parameter, called the CPMG frequency, as $\nu_{CPMG} = 1/(2\tau)$. This parameter represents the rate at which the spin – echo blocks are applied. Among N blocks, only the ones that have conformational transitions taking place during the evolution time will have a reduction in their refocusing capability. The number of such blocks, here called defective blocks, is N if $\nu_{CPMG} < k_{ex}$ (multiple transitions in 1 spin – echo block, see Figure 2.11b), and is k_{ex} τ_{CPMG} (1 transition per block, see Figure 2.11c) if $\nu_{CPMG} \geq k_{ex}$. After the number of defective spin-echo blocks of the CPMG pulse, the phases accrued by the spins will be distributed within the range as follows:

- if $\nu_{CPMG} < k_{ex}$:

$$N(\omega_A - \omega_B)\frac{\tau_{CPMG}}{2N} > \varphi(2\tau) > N(\omega_B - \omega_A)\frac{\tau_{CPMG}}{2N}$$

Or: $$(\omega_A - \omega_B)\frac{\tau_{CPMG}}{2} > \varphi(2\tau) > (\omega_B - \omega_A)\frac{\tau_{CPMG}}{2}$$

- if $\nu_{CPMG} > k_{ex}$:

$$k_{ex}\tau_{CPMG}(\omega_A - \omega_B)\frac{\tau_{CPMG}}{2N} > \varphi(2\tau) > k_{ex}\tau_{CPMG}(\omega_B - \omega_A)\frac{\tau_{CPMG}}{2N}$$

Or: $$\frac{k_{ex}\tau_{CPMG}}{N}(\omega_A - \omega_B)\frac{\tau_{CPMG}}{2} > \varphi(2\tau) > \frac{k_{ex}\tau_{CPMG}}{N}(\omega_B - \omega_A)\frac{\tau_{CPMG}}{2}$$

When $\nu_{CPMG} > k_{ex}$, we have:

$$\frac{k_{ex}\tau_{CPMG}}{N} = \frac{k_{ex}\tau_{CPMG}}{\frac{\tau_{CPMG}}{2\tau}} = \frac{k_{ex}}{\nu_{CPMG}} < 1,$$

Therefore:

$$(\omega_A - \omega_B)\frac{\tau_{CPMG}}{2} > \frac{k_{ex}\tau_{CPMG}}{N}(\omega_A - \omega_B)\frac{\tau_{CPMG}}{2}$$

And: $$\frac{k_{ex}\tau_{CPMG}}{N}(\omega_B - \omega_A)\frac{\tau_{CPMG}}{2} > (\omega_B - \omega_A)\frac{\tau_{CPMG}}{2}$$

This means when $\nu_{CPMG} < k_{ex}$, the range of phase distribution, or equivalently magnetization distribution, is broader than when $\nu_{CPMG} \geq k_{ex}$. When $\nu_{CPMG} \geq k_{ex}$, the magnetization is nearly completely restored along the X-axis after each block, and the refocusing by the CPMG train is improved. In this case, the contribution of the exchange process to transverse relaxation (by the decoherence of spins) is limited, leading to a lower relaxation rate as shown in Figure 2.11c and 2.11d. On the other hand, when $\nu_{CPMG} < k_{ex}$, the CPMG sequence becomes less effective at refocusing, leading to a higher transverse relaxation rate (Figure 2.11b and d).

The observed transverse relaxation is contributed to by both intrinsic transverse relaxation and conformational exchange, therefore, the effective relaxation rate, $R_{2,eff}$, is a combination of an intrinsic transverse relaxation rate and a conformational exchange rate. The intensity of a signal at a relaxation delay time t is written as:

$$I(t) = I_0 e^{-R_{2,eff}t}$$

It is then referenced against the intensity at time 0:

$$\frac{I(t)}{I_0} = e^{-R_{2,eff}t}$$

Therefore,

$$R_{2,eff} = -\frac{1}{t} ln\left(\frac{I(t)}{I_0}\right)$$

At t = τ_{CPMG}, and with an increasing number of spin-echo blocks applied, or equivalently, an increasing value of ν_{CPMG}, the signal intensity is no longer a function of time, instead, it becomes a function of ν_{CPMG}, I(t) = I(ν_{CPMG}). The effective relaxation rate is written as:

$$R_{2,eff} = -\frac{1}{\tau_{CPMG}} ln\left(\frac{I(\nu_{\text{CPMG}})}{I_0}\right)$$

The relationship between $R_{2,\text{eff}}$ and ν_{CPMG} is represented by a relaxation dispersion profile as illustrated in Figure 2.11d. In the case of no exchange, the profile is flat, showing no reduction of $R_{2,\text{eff}}$.

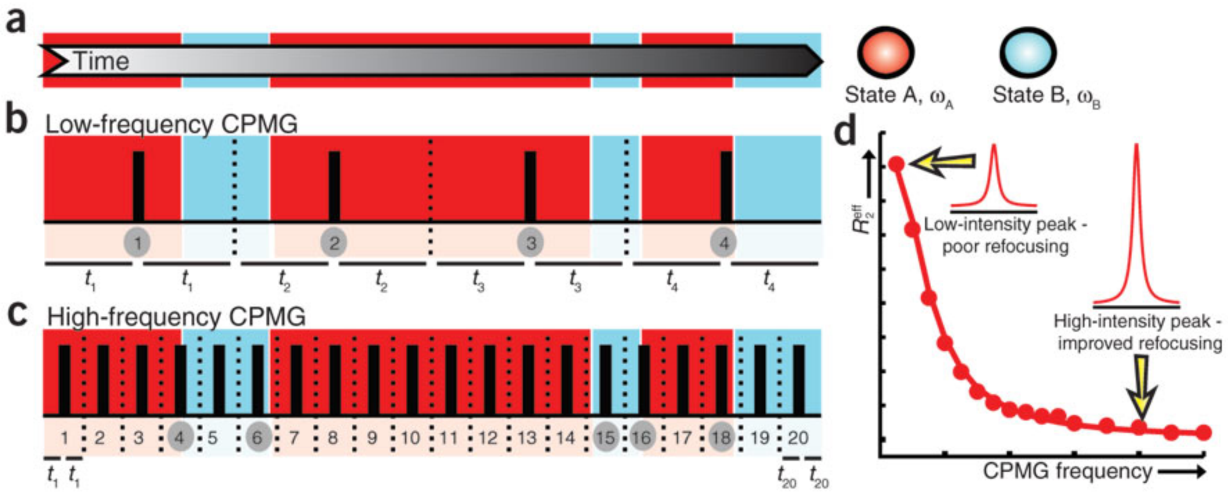


Figure 2.11. A simple schematic of a CPMG relaxation dispersion experiment. A molecule interconverts between two conformational states, A and B, as a function of time, with states A and B denoted by red and blue colors, respectively. Each molecule will interconvert with its own

trajectory, and a single example is shown. (b, c) During the trajectory, pulses are applied that lead to a modulation of the relaxation rates of NMR spins. (d) A relaxation dispersion profile is obtained from which details of the exchange reaction can be obtained. Reprinted with permission from reference [185]*. Copyright © 2009, Nature Publishing Group.*

In practice, conformational exchange can be characterized by following the signal intensities of data points on the transverse relaxation exponential decaying curve, corresponding to a relaxation time equal to τ_{CPMG}, when an increasing number of spin-echo blocks is used in the CPMG train. Relaxation dispersion profiles allow the identification of regions in a protein where μs-ms dynamics exist and can be used to track changes in the dynamics with respect to changing experimental conditions.

Extraction of information on the exchange process requires the fitting of experimental data to an equation describing a two-state exchange model. The Carver-Richards equations[296] are valid for any exchange regime:

$$R_2 = \frac{1}{2}\left[R_{2A}^0 + R_{2B}^0 + k_{ex} - \left(\frac{1}{2\tau}\right) cosh^{-1}(D_+ \cosh(\eta_+) - D_- \cos(\eta_-))\right]$$

$$D_\pm = \frac{1}{2}\left(\pm 1 + \frac{\psi\ +\ 2(\Delta\omega)^2}{\sqrt{\psi^2\ +\ \zeta^2}}\right)$$

$$\eta_\pm = \left(\frac{2\tau}{\sqrt{2}}\right)\left(\pm\psi + \sqrt{\psi^2\ +\ \zeta^2}\right)^{1/2}$$

$$\psi = (R_{2A}^0 - R_{2B}^0 - p_A k_{ex} + p_B k_{ex})^2 - (\Delta\omega)^2 + 4 p_A p_B k_{ex}^2$$

$$\zeta = 2\Delta\omega(R_{2A}^0 - R_{2B}^0 - p_A k_{ex} + p_B k_{ex})$$

The Ishima-Torchia equation [297]is valid for a skewed population ($p_B < p_A$)

$$R_2 = R_2^0 + \frac{p_A p_B (\Delta\omega)^2 k_{ex}}{k_{ex}^2 + \left(p_A^2 (\Delta\omega)^4 + \frac{144}{16\tau^4}\right)^{1/2}} = R_2^0 + \frac{\Phi_{ex} k_{ex}}{k_{ex}^2 + \left(p_A^2 (\Delta\omega)^4 + \frac{144}{16\tau^4}\right)^{1/2}}$$

The fast-exchange equation is valid when $k_{ex} > \Delta\omega$

$$R_2 = R_2^0 + \left(\frac{p_A p_B (\Delta\omega)^2}{k_{ex}}\right)\left(1 - \frac{2\tanh(k_{ex}\tau)}{2k_{ex}\tau}\right) = R_2^0 + \left(\frac{\Phi_{ex}}{k_{ex}}\right)\left(1 - \frac{2\tanh(k_{ex}\tau)}{2k_{ex}\tau}\right)$$

The analysis using Carver-Richards equations can easily be programmed for fitting *via* standard computational tools despite their apparent complexity and therefore it is the method of choice. Information on the exchange process, including the rate of exchange, the populations of the conformers and the absolute chemical shift differences between the conformers, can be extracted from the analysis. Such information gives insight into the secondary structures of the excited states or of transient secondary structures in the protein dynamic ensemble, which are commonly found in IDPs. Relaxation dispersion overcomes the limits usually encountered with low populations and short lifetimes (μs-ms) of excited states and transient conformations, which make them invisible to many of the tools of structural biology[185, 291].

2.4. Amide hydrogen exchange in proteins.

Hydrogen exchange (HX) is recognized as a powerful tool available to study proteins, especially to study protein conformational exchange. It began as a simple idea to pursue a straightforward goal, but later revealed both the unexpected complexity of the subject and the potential power of the method for probing deeply into how proteins work. In 1951, Pauling and Corey discovered the α-helix and β-sheet and postulated that the structures were stabilized by hydrogen bonds. The hydrogen exchange method was developed by Linderstrøm-Lang[298, 299] starting in 1954 in order to answer the question of how to identify protein hydrogen-bonded structures. He proposed that NH protons in peptides underwent exchange with H_2O by a mechanism that involves hydrogen bonding, whereas C-H protons did not exchange measurably in the same time frame, and that internally H-bonded peptide NH protons in proteins would exchange much more slowly than non-hydrogen bonded peptide NH protons such that they can be easily distinguished from each other. He further proposed that any given protein hydrogen-bonded amide may exchange by being exposed to different kinds of dynamic conformations ranging from small local fluctuations to whole molecule unfolding[300] and that there was a relationship connecting the hydrogen exchange rate with protein dynamics and energetics[301].

2.4.1. Chemistry of hydrogen exchange.

The chemical basis of the hydrogen exchange process between water and solute molecules involves reactions that are catalyzed by acids or bases[302, 303]. In the base-catalyzed reaction, an amide proton is removed to create the imidate anion (the conjugate base of the amide),

$$\mathrm{RC(=O)NHR'} \;+\; \mathrm{OH^-} \rightleftharpoons \mathrm{RC(–O^-)=NR'}\ \text{(imidate)} \;+\; \mathrm{H_2O}$$

The imidate anion then abstracts a H^+ from water to regenerate the amide:

$$\mathrm{RC(–O^-)=NR'} \;+\; \mathrm{H_2O} \rightleftharpoons \mathrm{RC(=O)NHR'} \;+\; \mathrm{OH^-}$$

The mechanism of the acid-catalyzed reaction[303] depends on the amide. For the $CONH_2$ side chains of asparagine and glutamine residues, the exchange occurs by protonation of the nitrogen, followed by removal of a different hydrogen as H^+:

$$\mathrm{RC(=O)NH_2} \;+\; \mathrm{H^+} \rightleftharpoons \mathrm{RC(=O)NH_3^+} \rightleftharpoons \mathrm{RC(=O)NH_2} \;+\; \mathrm{H^+}$$

For backbone CONH groups, the catalytic H$^+$ bonds with the oxygen atom, which is much more basic than the nitrogen, to produces a conjugate acid.

$$\mathrm{RC({=}O)NHR' + H^+ \rightleftharpoons RC({-}OH) = NHR'^+}$$

A H$^+$ from the nitrogen is removed to produce an imidic acid, an unstable tautomer of an amide.

$$\mathrm{RC({-}OH) = NHR'^+ \rightleftharpoons RC({-}OH) = NR'^+\ (imidic\ acid) + H^+}$$

Finally, a solvent H$^+$ bonds to the amide nitrogen, followed by removal of a H$^+$ from oxygen, and the amide is regenerated, but with its H having been exchanged:

$$\mathrm{RC({-}OH) = NR'^+ + H^+ \rightleftharpoons RC({-}OH) = NHR'^+ \rightleftharpoons RC({=}O)NHR' + H^+}$$

Because hydrogen exchange is catalyzed by H$^+$ and OH$^-$, a graph showing the log of the hydrogen exchange rate and pH is V-shaped, with a minimum rate occurring near 3 (Figure 2.12). Side chains of the nearest neighboring residues may exert inductive effects[304, 305] and shift the curve along the pH axis. Steric blocking effects from side chains may also decrease both acid and base catalyzed rates[305].

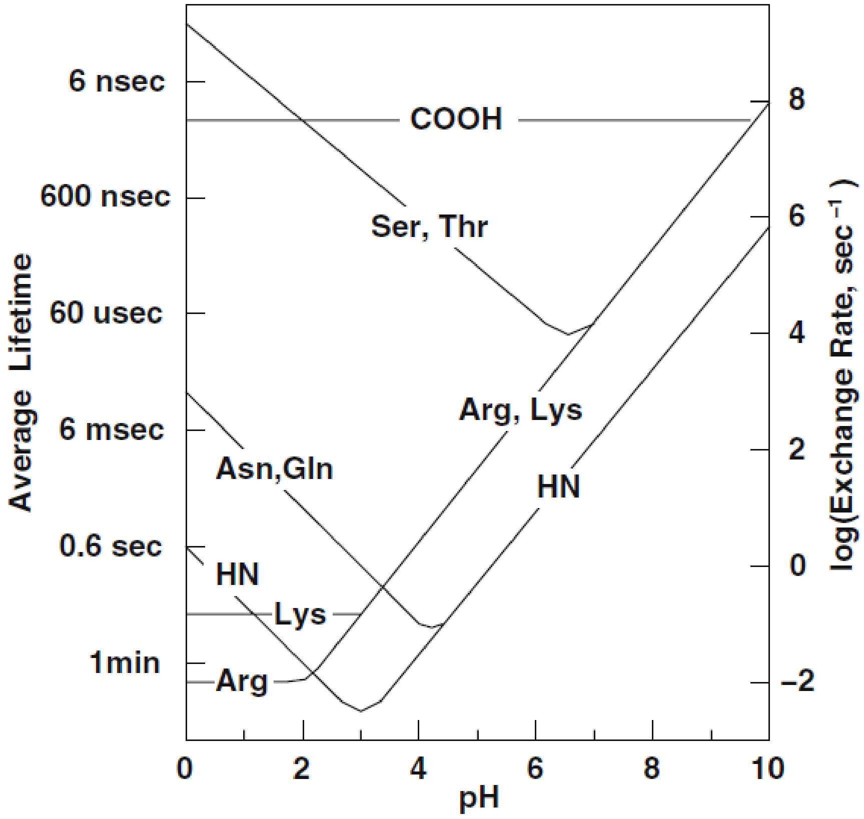


Figure 2.12. Effect of pH on hydrogen exchange rates. The hydrogen exchange rates are shown for several exchangeable groups in proteins. NH indicates the backbone amide proton. Other labels refer to side chains. The displayed rates assume that the group is fully exposed to solvent. The left axis indicates the mean lifetime of the proton while the right scale gives the exchange rate. Reprinted with permission from reference [242]*. Copyright © 2006 Springer.*

2.4.2. Mechanisms of protein hydrogen exchange.

Linderstrøm-Lang proposed a general model for hydrogen exchange as follow[306-309]:

$$\text{Closed} \underset{k_{cl}}{\overset{k_{op}}{\rightleftharpoons}} \text{Open} \xrightarrow{k_{int}} \text{Exchange}$$

The model supposes that protected hydrogens (closed state) cannot exchange at all. In order to be exchange competent, structural transitions are needed allowing the protein to change from a closed to an open state. Such a transition is kinetically defined by opening and closing rates, k_{op} and k_{cl}. In the exposed condition, an exchange event occurs at some rate, k_{int}. The measured hydrogen exchange rate, k_{HX}, is as follow:

$$k_{HX} = \frac{k_{op}k_{int}}{k_{op} + k_{cl} + k_{int}}$$

When $k_{int} >> k_{op} + k_{cl}$, the measured exchange rate approaches a maximum rate, $k_{HX} = k_{op}$. In this so-called EX1 limit, the hydrogen exchange experiment measures the opening rate, or the mean residence time of the closed state, $\tau_C = 1/k_{op}$. More commonly, hydrogen exchange experiments are performed in the condition where $k_{int} << k_{op} + k_{cl}$, or the EX2 limit, and the measured exchange rate is as follow:

$$k_{HX} = \frac{k_{int}}{\kappa + 1}$$

where $\kappa = \frac{k_{cl}}{k_{op}} = \frac{\tau_C}{\tau_O}$ is the effective equilibrium constant between open and closed states, and τ_O is the mean residence time of the open state.

In the open state, the hydrogen exchange process is free from any structural effects and only subjected to the inductive and steric effects from neighboring side chains, which is similar to the hydrogen exchange of a short peptide. The hydrogen exchange of peptide models were thoroughly investigated by Bai *et al.*[305, 310], and a method to precisely predict the exchange rates in any sequence, at various temperatures and pH values was derived.

Hydrogen exchange of the open state is assumed to proceed at the same rate as in model peptides with the same neighboring side chains, $k_{int} = k_{HX}^{0}$, so κ, the equilibrium constant between open and closed states, can be deduced from the measured protein hydrogen exchange rate and the rate of exchange of a model peptide:

$$\kappa = \frac{k_{HX}^{0}}{k_{HX}} - 1.$$

The protection factor, defined as $P = \frac{k_{HX}^{0}}{k_{HX}} = \kappa + 1$, represents the effect of structural protection on the hydrogen exchange rate. Amide protons with higher protection factors have lower exchange rates.

Hydrogen exchange can only occur if the amide proton is accessible to solvent and able to H-bond with solvent, so protein conformational fluctuations must be an integral part of the exchange mechanism. However, the protection factor, which greatly impacts the exchange rate, does not provide information on the structure and dynamics of the open state. Several attempts have been made to correlate experimental protection factors with physical attributes of the amides, such as solvent contact[311-315], burial depth[316], intramolecular H-bonds[313, 316-318], packing density[316, 319], and electric field[320]. However, upon conducting hydrogen exchange measurements on the doubly-mutated Staphylococcal Nuclease (SN) protein, the Englander group[300, 309] has drawn out some important conclusions:

- The solvent contact model suggests that hydrogen exchange is dependent on the relative accessibility to solvent of exchanging amides in the static protein structure, such that the more accessible the amide, the faster the exchange rate. This holds true only for hydrogens on dynamically unstructured protein segments in contact with solvent, as they exchange at the expected rate of unstructured peptide models. Unprotected hydrogens on structured segments exchange slower than expected by up to 40-fold. Hydrogens that are sterically protected by H-bonding exchange even more slowly by many orders of magnitude, suggesting that dynamic perturbations are required to separate the protecting H-bond and expose the hydrogen to attack by an exchange catalyst.

- Other models correlate the burial depth and the solvent accessibility of amides with the slow exchange of buried hydrogens, with an assumption that the hydrogen exchange process requires the penetration of catalyst into the protein matrix. However, the study shows that this model is incorrect because many hydrogens near the protein surface can exchange as slowly as the well buried ones and therefore the dependence of protection factors on depth of burial is in significant.
- H-bonding to water molecules that are held in place by protein interactions can also block exchange. Structural reorganization is required to displace such water molecules.
- In order to proceed, hydrogen exchange requires protein dynamics, ranging from transient global unfolding of a protein molecule, through smaller cooperative unfolding of secondary structural elements, and down to local structural fluctuations that involve as little as a single peptide group or side chain or water molecule.

Hydrogen exchange measurement is a valuable method to characterize IDPs. Typically, protection factors are on the order of $10^3 – 10^6$ for folded proteins, whereas, transient structural elements in IDPs can only provide a protection up to about 10-fold[321]. This approach has been successfully used for showing the lack of structure in the N- and C-terminal regions of Nogo[322] and segments of the HET-s prion amyloid[323], as well as for demonstrating the existence of transient helical segments in securin[324] and in the N-terminal domain of histone mRNA binding protein SLBP[325].

2.4.3. Hydrogen exchange measurement by NMR Spectroscopy.

For folded proteins, H-bonding in structural elements provides protection to hydrogens, leading to low exchange rates, and the time frame for the exchanges may range from minutes to months. Measurement of hydrogen exchange in such long time frames can be conveniently done by the deuterium substitution experiment. In a basic version of this experiment, a lyophilized protein is dissolved in D_2O and a series of 2D spectra is recorded to monitor the progressive loss of amide proton signals[326]. The accessible time frame of this experiment is limited by how fast the experiment can be repeated. Introduction of the accelerated-acquisition 2D experiments ("ultrafast" 2D NMR spectroscopy, SOFAST HMQC, and UltraSOFAST HMQC) with repetition rates approaching 1 Hz has extended the limits of measurement by NMR spectroscopy[327]. Proton–deuterium substitution can also be monitored in fast exchanging systems on timescales from ms to second by stop-flow type experiments[328].

Amide hydrogens of IDPs or unprotected amide hydrogens in folded proteins exchange with solvent at much higher rates, which can be measured by equilibrium NMR techniques. In these experiments, water protons are selectively excited, and the magnetizations are transferred to other sites due to the exchange of protons. The rate at which the magnetization, or equivalently the observed NMR signal, builds up on one site is equal to the rate of exchange[329]. While being simple and robust, this approach encounters a number of difficulties[330], including:

(a) Subtraction artefacts, and even false detection of hydrogen exchange, due to water signal radiation-damping;

(b) Long recycling delays between scans are required to restore saturated water magnetization to a true equilibrium;

(c) Initial degree of saturation of the water signal and the recovery rate of the water magnetization must be known for data analysis;

(d) Alternative magnetization transfer pathways make isolation and quantification of amide hydrogen exchange difficult. They include the intramolecular NOE and TOCSY (Total Correlation Spectroscopy) transfers between saturated $^1H^\alpha$ spins and $^1H^N$; the relayed transfer, where the saturation first propagates from water to hydroxyl and amine sites *via* chemical exchange and then onto $^1H^N$ spins *via* the NOE; and the intermolecular NOE transfer from water to $^1H^N$.

Pursuing the equilibrium approach, the CLEANEX-PM (Phase-Modulated CLEAN chemical Exchange),[331] has long been considered a benchmark experiment for accurate quantification of amide hydrogen exchange. In the beginning of the sequence, a (gradient)–(selective pulse)–(gradient) combination is used to destroy $^1H^N$ magnetization, while at the same time preserving H_2O magnetization. It avoids the interferences from unwanted magnetization transfer due to the Overhauser effect by using a spin-lock module to lock the proton magnetization during the mixing period, in which chemical exchange between water and amide protons takes place. Effects from radiation-damping are suppressed by using a small gradient applied through the whole mixing period. At the end of the mixing period, the FHSQC (Fast Heteronuclear Single-Quantum Correlation) sequence is used as a detection scheme to resolve peaks and to flip the water back to the z-axis before detection, thereby avoiding water saturation[329]. With the unwanted pathways blocked, CLEANEX-PM monitors the flow of magnetization from water to amides due to the solvent exchange. In practice, a series of spectra at different mixing times are acquired, along with a reference FHSQC spectrum. Exchange rates can be extracted by fitting signal intensities to the following equation[329]:

$$\frac{I}{I_o} = \frac{k}{\left(R_{1A,app} + k - R_{1B,app}\right)}\left[e^{-R_{1B,app}\tau_m} - e^{-\left(R_{1A,app}+k\right)\tau_m}\right]$$

where:

- I is signal intensity at a certain mixing time τ_m;
- I_o is signal intensity of a reference spectrum;
- k is the rate of exchange;
- $R_{1A,app}$ is a combination of amide proton longitudinal and transverse relaxation rates;
- τ_m is the mixing time;
- $R_{1B,app}$ is measured by separate experiments, in which the dependence of the water signal on mixing times is observed using 1D CLEANEX-PM without water suppression.

When applied to IDPs, CLEANEX-PM is potentially affected by some issues as pointed out by Chevelkov *et al.*[330]:

- Although Hwang *et al.*[331] intentionally minimize the effect of radiation damping by using a weak gradient during the mixing time, the uncertainty about the degree of water saturation due to the short recycling delays and the loss of magnetization during the selective pulse, especially in cases where the sample is heated up due to the rf pulse, may lead to a systematic error in the determined exchange rates.

- Unwanted exchange – relayed magnetization transfer from $^1H^\alpha$ to $^1H^N$ for residues experiencing fast internal motions (on the ps-ns timescale) cannot be compensated perfectly, as pointed out by the authors of CLEANEX-PM[331].
- Intermolecular NOE transfer from water to protein cannot be fully discounted.
- The powerful proton spin-lock applied may cause a significant amount of heating in samples at high salt concentration, which in turn speeds up the exchange rates.

Based on the above mentioned observations, Chevelkov *et al.*[330] proposed the SOLEXSY experiment that overcomes such limits particularly for the study of IDPs. The SOLEXSY experiment relies on the (HACACO)NH pulse sequence[332, 333], and monitors the interconversion between proton- and deuterium-bound ^{15}N spins in 50% H_2O–50% D_2O solution. The experiment begins with the magnetization transfer from $^1H^\alpha$ to ^{15}N. The two species, proton bound nitrogen $^{15}N^H$, and deuterium bound nitrogen $^{15}N^D$, are frequency-labeled during the first evolution time. The respective ^{15}N resonances will be resolved by the isotopic shift difference of 0.7 ppm[334]. During the following variable-length mixing time period, the $^{15}N^H$ moieties are partially transformed into $^{15}N^D$ and *vice versa* due to the exchange with solvent. The magnetizations are then transferred to $^1H^N$ for detection.

The detected signal includes 2 types of peaks (see Figure 2.13). The first group of peaks, called $^{15}N^H$ peaks, represent the amides that are initially protonated, and remain protonated after the exchange period. The second group of peaks, called $^{15}N^D$ peaks, represents the amides that are initially deuterated, and become protonated after the exchange period. When the duration of the mixing time is increased, the intensity of the $^{15}N^H$ signals drops due to the partial displacement of amide protons by deuterons whereas the intensity of $^{15}N^D$ signals increases as some of the initially deuterated amides become protonated. With sufficiently long mixing times, the intensities of the 2 groups become equalized since the solution contains 50% H_2O–50% D_2O.

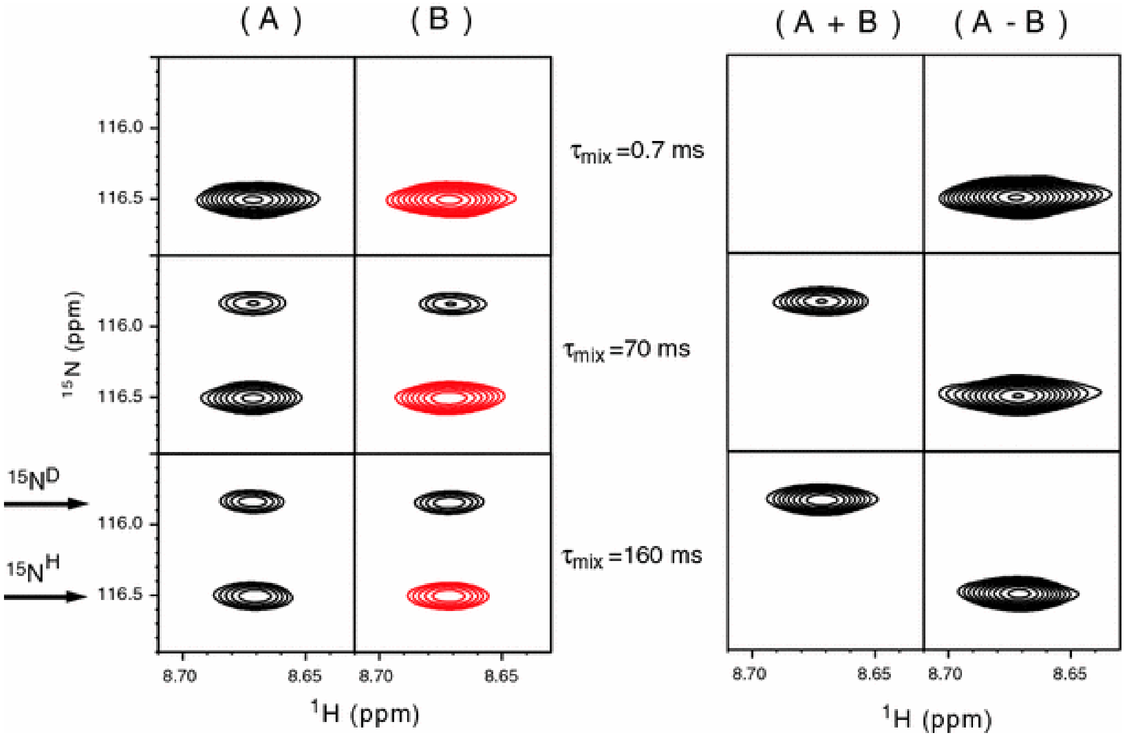


Figure 2.13. Regions from $^{15}N^{H/D}$–$^1H^N$ SOLEXSY spectra containing the signals from residue S34 of the chemically denatured drkN SH3 domain. The spectra were recorded using the SOLEXSY pulse sequence for three different values of τ_{mix}, as indicated in the plot. The columns A and B display the spectra obtained with the positive and negative proton frequency labeled magnetization of nitrogen (positive contour levels are plotted in black, negative contour levels in red). The results of the addition and subtraction of these spectra are shown in the columns labeled A + B and A – B. Reprinted with permission from reference [330]. Copyright © Springer Science+Business Media B.V. 2010

In the SOLEXSY experiment, for each exchanging site, 2 peaks representing the 2 frequency-labeled magnetizations are used, leading to a doubling of the number of peaks in an NMR spectrum. The following treatment is used to reduce the crowding: the magnetization of $^{15}N^{H}$ species is sign-encoded as positive or negative before the mixing period in 2 series of experiments, resulting in a positive or negative peak on the spectrum. As shown in Figure 2.13, the positive peaks of $^{15}N^{H}$ are in black in the A series, and the negative peaks are in red in the B series. Addition of the spectra of the same mixing time A+B results in the elimination of the $^{15}N^{H}$ peaks, whereas subtraction A – B removes the $^{15}N^{D}$ peaks. By following the intensities of the 2 series A+B and A-B, a hydrogen exchange profile of one site at different mixing times can be generated. The exchange rate is extracted from simultaneously fitting the NMR signal intensities of the 2 series to the McConnell equation[335, 336]:

$$\frac{d}{d\tau}\begin{pmatrix} N_z^H(\tau) \\ N_z^D(\tau) \end{pmatrix} = \begin{pmatrix} -R_1^{NH} - k_{HD} & k_{HD} \\ k_{HD} & -R_1^{ND} - k_{HD} \end{pmatrix} \begin{pmatrix} N_z^H(\tau) \\ N_z^D(\tau) \end{pmatrix}$$

where,

- τ runs from 0 to τ_{mix}, and τ_{mix} is the mixing time.
- k_{HD} and k_{DH} are rates of exchange from proton to deuterium and *vice versa*, respectively; and $k_{HD} = 1.1\ k_{DH}$;
- $N_z^H(\tau)$ and $N_z^D(\tau)$ are the magnetizations of nitrogen atoms labeled with proton and deuterium frequencies, respectively, as a function of τ.
- R_1^{NH}, R_1^{ND} are the relaxation rates of proton and deuterium frequency-labeled nitrogen. They, along with $N_z^H(0)$, $N_z^D(0)$, k_{HD} and k_{DH}, can be extracted as fitting parameters.

One important note is that the SOLEXSY experiment is suitable for measurements at or near physiological pH, but cannot be used with acid-denatured samples at pH ~2.

Chapter 3

Molecular dynamics simulations of disordered proteins.

Classical molecular dynamics (MD) simulations is one of the principal tools in the theoretical study of biological molecules and complexes in terms of structure, dynamics and thermodynamics. This computational method calculates the time dependent behavior of a molecular system and thus provides detailed information on the fluctuations and conformational changes of proteins and nucleic acids. Four decades after the first protein simulation of bovine pancreatic trypsin inhibitor[337], today in the literature, molecular dynamics simulations of solvated proteins, protein-DNA complexes as well as lipid systems are routinely found, addressing a variety of issues. The continuing growth of computing power and the appearance of supercomputers have allowed a great expansion of the simulation technique in terms of the level of modeling (the world's largest molecular dynamics simulation was the simulation of $4.125*10^{12}$ particles on the supercomputer SuperMUC, located at the Leibniz Supercomputing Centre, Garching, Munich[338]), simulation resolution (all atom, coarse grain), timescale (from sub-picosecond, up to ms), and degrees of freedom of interest (variation in temperature, pressure, pH). Simulators are computer programs that carry out the simulation calculation. They generally employ multithreading techniques to enhance simulation speed. Some popular packages can be named such as

CHARMM[339], LAMMPS[340], NAMD[341], OpenMM[342], and GROMACS[343], the program used in this project.

3.1. Modeling of proteins.

In order to describe the dynamics of a molecular system, a protein in particular, the most precise method is using quantum mechanics. However, due to the large number and the complexity of the interactions of the electrons and nuclei of atoms and molecules, it is a great challenge to apply a full quantum mechanical dynamics calculation on macromolecules like a protein system in water on a biological timescale level (ns – ms). Such molecular systems, therefore, need to be simplified. The first simplification is the employment of the Born-Oppenheimer approximation[344, 345]. This approximation assumes that electrons move much faster than nuclei because of being much lighter; therefore, they can quickly rearrange themselves to a stable configuration almost as soon as the nuclei move. As a consequence, the wave function that describes an electron configuration is parametrically dependent on the nuclear geometry. Each configuration of a nucleus is associated with a single electronic quantum state and a single potential energy level. Molecular motions change the atomic coordinates, thus giving rise to a potential energy surface (PES).

Further simplification is made by using classical mechanics to describe particles instead of quantum mechanics. In low-resolution (coarse grained) simulations, proteins are modeled as a chain of beads, where amino acid residues are treated as soft, interpenetrating spheres, with or without point charges at their centers, which are allowed to fluctuate during the simulation to account for protein charge regulation. Higher resolution simulations use atomistic modeling, where protein atoms are treated as effective point charges with mass, instead of nuclei with explicit electrons. Under classical treatment, a system of N particles is described by a position set, $\{x^1, y^1, z^1, \dots, x^N, y^N, z^N\}$, and a momentum set, $\{p_{x^1}, p_{y^1}, p_{z^1}, \dots, p_{x^N}, p_{y^N}, p_{z^N}\}$.

The total energy at one system state is the sum of the kinetic energy, $T(\{p^N\})$, and the potential energy, $V(\{r^N\})$:

$$E_{total} = T + V$$

The kinetic term is calculated as the sum of the kinetic energies of all the particles:

$$T = \sum_i^N \frac{|p_i|^2}{2m_i} = \sum_i^N \frac{1}{2} m_i v_i^2$$

Where m_i and v_i are respectively the mass and the velocity of atom i.

The potential energy term in a coarse grained simulation has the following form[346]:

$$V = V_{bonds} + V_{Pauli\ repulsion} + V_{hydrophobic} + V_{Debye-Huckel} + V_{titration}$$

$$+V_{Lennard-Jones} + V_{Gouy-Chapman}$$

In atomic simulations, the potential energy term has the following form[346]:

$$V(\{r^N\}) = \underbrace{V_{bonds} + V_{angles} + V_{proper\ dihedrals} + V_{impropper\ dihedrals}}_{\text{bonded interactions}}$$

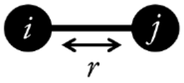
$$\underbrace{+ V_{Coulomb} + V_{van\ der\ Waals}}$$

non-bonded interactions

In the potential energy formalisms above, each term represents one type of interaction and is defined as follows[343, 346]:

- The bonding term V_{bonds} represents the covalent connection of the 2 particles i and j, while allowing their harmonic vibration around an equilibrium bond distance, r_0:

$$V_{bonds} = V\left(r_{ij}\right) = \sum_b \frac{1}{2} k_b (r_{ij} - r_0)^2$$



  k_b is a force constant where b stands for bond, meaning the sum is taken over all pairs of bonded particles.

- The Pauli repulsion term $V_{Pauli\,repulsion}$ represents the overlap of the beads:

$$V_{Pauli\,repulsion} = \sum_{i,j} 4\epsilon_r \left(\frac{\sigma_i + \sigma_j}{2r_{ij}}\right)^{12}$$

  ϵ_r is a parameter inversely proportional to the degree of overlap allowed between the beads. σ_i and σ_j are the radii of beads i and j, respectively.

- The hydrophobic term $V_{hydrophobic}$ represents the attraction of the beads within a certain cut off, r_{cutoff}, due to hydrophobicity:

$$V_{hydrophobic} = \sum_{i,j} \epsilon_h \quad \text{for} \quad r_{ij} \leq r_{cutoff}$$

- $V_{Debye-Huckel}$ represents the interaction of charged residues (in coarse grained simulations) according to Debye – Huckel theory:

$$V_{Debye-Huckel} = \sum_{i,j} \lambda_B \frac{z_i z_j k_B T}{r_{ij}} e^{-\kappa r_{ij}}$$

If q_i and q_j are the charges of residue i and j, then $q_i = z_i\ e_c$ and $q_j = z_i\ e_c$ where e_c is the elementary electron charge, k_B is the Boltzmann constant, T is the absolute temperature, λ_B is the Bjerrum length, defined as $\lambda_B = \frac{e_c^2}{4\pi\varepsilon_0\varepsilon kT}$ with ε_0 and ε respectively the vacuum permittivity and the dielectric permittivity of the medium, and κ is the inverse of the Debye screening length defined as $\kappa^{-1} = \left(\frac{\varepsilon_0\varepsilon kT}{2N_A e_c^2 I}\right)^2$ where N_A is Avogadro's number and I is the ionic strength.

- The $V_{titration}$ term represents the contribution from the protonated state of a residue to the potential energy function if the deprotonated state of the residue is chosen as the reference state with no overall contribution.

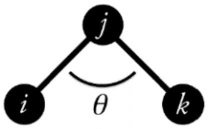
$$V_{titration} = \sum_i kTln\left[10\left(pH - pK_{a_i}\right)\right],$$

if residue i is protonated.

- The 2 terms $V_{Lennard-Jones}$ and $V_{Gouy-Chapman}$ only exist in surface adsorption simulations.

- V_{angles} represents the harmonic vibration of bond angles θ_{ijk} around the equilibrium angle θ_0:

$$V_{angles} = V\left(\theta_{ijk}\right) = \sum_{\theta} \frac{1}{2} k_\theta \left(\theta_{ijk} - \theta_0\right)^2$$



where k_θ is a force constant.

- $V_{proper\ dihedrals}$ and $V_{impropper\ dihedrals}$ represent torsional potentials of bonded interactions of 4 atoms. The torsional potential of a proper dihedral angle (Figure 3.1a) is the ability of a bond to rotate around its own longitudinal axis, while the torsional potential of an improper dihedral angle (Figure 3.1b) is mainly used to maintain planarity in a molecular structure and to prevent a molecule from flipping over to its mirror image.

$$V_{proper\ dihedrals} = V(\phi_{ijkl}) = \sum_{\phi} k_{\phi}[1 + \cos(n\phi_{ijkl} - \delta)]$$

$$V_{improper\ dihedrals} = V(\xi_{ijkl}) = \sum_{\xi} k_{\xi}(\xi_{ijkl} - \xi_0)$$

ϕ_{ijkl} and ξ_{ijkl} are proper and improper dihedral angles, respectively, k_{ϕ} is the height of the torsional energetic barrier, ξ_0 is the equilibrium dihedral angle, and k_{ξ} is the force constant.

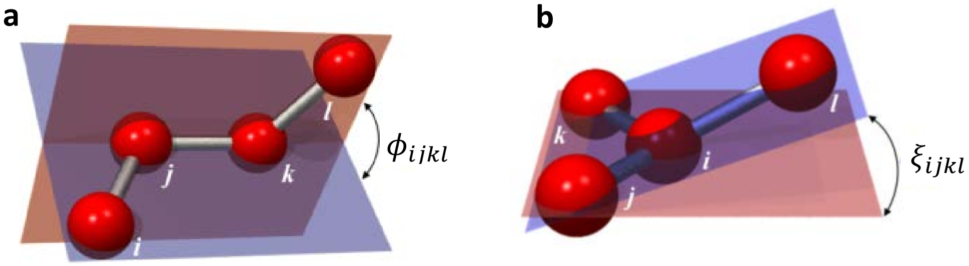


Figure 3.1. Proper dihedral (a) and improper dihedral (b) bond angles. The torsional angles (ϕ_{ijkl} and ξ_{ijkl}) are the angles between the plane going through the atoms i, j and k and the plane going through the atoms j, k and l. Adapted from webpage http://cbio.bmt.tue.nl/pumma/index.php/Theory/Potentials.

- $V_{Coulomb}$ represents the sum of all Coulombic interactions between any pair of charged particles (in atomic simulations) in the system:

$$V_{Coulomb} = U(r_{ij}) = \sum_{i,j} \frac{q_i q_j}{4\pi\varepsilon_0 \varepsilon r_{ij}}$$

- $V_{van\ der\ Waals}$ represents the contribution of the van der Waals interactions in the system:

$$V_{van\ der\ Waals} = U(r_{ij}) = \sum_{i,j} 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right]$$

where ϵ_{ij} is the depth of the potential energy well, and σ_{ij} is the finite distance at which the inter-particle potential is zero.

Other simplifications for protein modeling are that the electrons are highly localized and no bond formation or breaking can happen during the course of a simulation.

Simulations are normally done with a periodic boundary condition (section 3.3) where the molecular system within the boundary is considered isolated, therefore, the total energy is conserved. At a time t during the simulation, if the coordinate set of particles of the system is defined, the potential energy can be calculated, and the kinetic energy therefore is known. It needs to be noted that the initial kinetic energy, or equivalently the initial velocities of the particles, is generated as described in 3.5.a. By using the Lagrangian equation of motion[347], the force acting on each particle i in the system is derived as follow[348]:

$$f_i = -\frac{\partial V(\{r^N\})}{\partial r_i}$$

and can theoretically be calculated. Newton's second law of motion states that:

$$f_i = m_i a_i = m_i \ddot{r}_i$$

Therefore, the acceleration and velocity at time t can also be found and the new coordinates after a period of evolution Δt can be calculated. The same calculation process is again done on the new coordinates at time $t + \Delta t$. By repeating such calculations until the desired simulation time is reached, a trajectory of the molecule is generated since the coordinates of atoms are found.

In practice, there are several algorithms available to solve the derivative equations above. The leap-frog algorithm, presented pictorially below, is derived from the basic Verlet scheme[349], and is one of the most common methods typically used in molecular dynamics simulations.

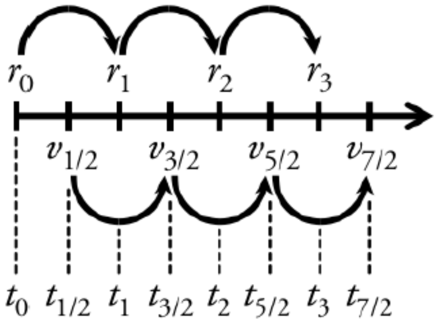


Figure 3.2. Schematic representation of the leap-frog algorithm.

At any step, a velocity v at time step $t + \frac{\Delta t}{2}$ is calculated from the velocity at time step $t - \frac{\Delta t}{2}$ and the acceleration a at time t:

$$v\left(t + \frac{\Delta t}{2}\right) = v\left(t - \frac{\Delta t}{2}\right) + \Delta t\, a(t)$$

The position r at time $t + \Delta t$ is calculated based on the velocity calculated above and the position at time t:

$$r(t + \Delta t) = r(t) + \Delta t\, v\left(t + \frac{\Delta t}{2}\right)$$

The velocity at time t is determined as:

$$v(t) = \frac{1}{2}\left[v\left(t - \frac{\Delta t}{2}\right) + v\left(t + \frac{\Delta t}{2}\right)\right]$$

The time step Δt is chosen to be relatively short (typically 2 fs). If the time step is too long, the change in energy is large enough to over-stretch the bond lengths / angles, which propagate during the simulation, and can lead to “blowing up” of the molecule.

3.2. Water model.

Water is the most common solvent used in the study of biological molecules. In practical experiments, how a protein behaves in solvent is of much interest, and studying the interactions between protein and solvent may reveal a lot of information about protein dynamics and functions. When doing a simulation, a single molecular system is simulated in a box, surrounded by a large number of water molecules. The appropriate treatment of solute–solvent and solvent–solvent interactions is also a key factor for the outcome of most atomistic simulations.

There are many water models that have been proposed, and the most commonly used in classical molecular dynamics simulations are the Single Point Charge (SPC) family (SPC[350] and SPC/E[351] models) and the TIPnP family (TIP3P[352], TIP4P[353], TIP4P/2005[354] and the modified version TIP4P-D[355]) (see Figure 3.3). All these models are non-polarizable in such a way that the partial charges and dipole moments are conserved throughout the simulation, regardless of the external electric fields they may eventually be exposed to.

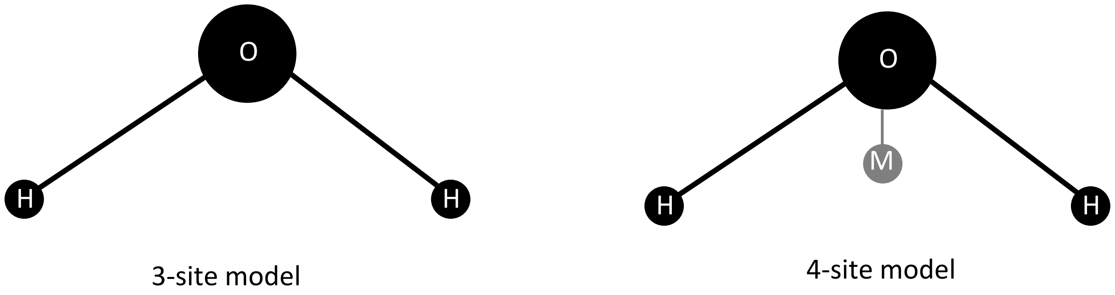


Figure 3.3. 3-site and 4-site water model.

The differences among the water models come from the number of interaction sites (or particles) in their construction, and from the parameters for equilibrium bond length, angle, partial charges and van der Waals interaction strength. The SPC family and the TIP3P are 3-site models, and the TIP4P, TIP4P/2005 and TIP4P-D are 4-site models. In the 3-site models, each atom has a point charge assigned whereas in the 4 –site models, the negative charge sits on a "dummy" atom M instead of on the oxygen, which improves the electrostatic distribution around the water molecule. Models with more parameters may describe the system better, but also demand higher computational costs; therefore only the 3-site models and 4-site models are commonly used in molecular dynamics simulations. Choosing a water model generally depends on the force field selection since each model is developed in and for a specific force field family. For example, the GROMOS force fields use the SPC and SPC/E models; AMBER force fields use TIP3P; OPLS force fields use TIP4P and TIP5P; CHARMM force fields use TIP3P or mTIP3P, a modified version of TIP3P. However, a certain water model developed for use in one force field can also be adopted for use in another force field.

The choice of water model becomes more important when running simulations using force fields (section 3.4) that are not optimized for disordered proteins. The resulting disordered-state ensembles tend to be structurally too compact compared to experiments[355]. The TIP4P-D water model, a modified version of the TIP4P/2005 model, is optimized for IDP simulations. When used with non-IDP-suitable force fields, it shows excellent agreement with the best performing IDP-suitable force fields[356]. TIP4P-D also shows good performance with the Amber99SB-ILDN force field, which is IDP-suitable. However, using the CHARMM22*[357] force field optimized for IDPs, with TIP4P-D yields poorer results compared with experimental data[356].

3.3. Periodic boundary condition.

In molecular dynamics simulations, periodic boundary conditions are used to avoid the boundary effects caused by the finite size of the simulation and to reduce the computational cost. It allows the simulation of a small part of a large system by enclosing particles in a box, which is replicated to infinity by rigid translation in all three Cartesian directions, completely filling space. Particles in the replicate image boxes move in a coherent fashion with particles in the original box. When a particle enters or leaves the simulation region, an image particle leaves or enters this region, such that the number of particles from the simulation region is always conserved (see Figure 3.4).

The 3 most useful box types for simulation of solvated systems are: the cubic box, the truncated octahedron box, and the rhombic dodecahedron box. Even though the cubic box is the simplest periodic system to visualize and to program, maximum performance is generally achieved by using the rhombic dodecahedron box. Its volume is 71% of the volume of a cube having the same image distance. By reducing the volume, the number of solvent particles is reduced, resulting in a reduction of about 29% in computational time[343].

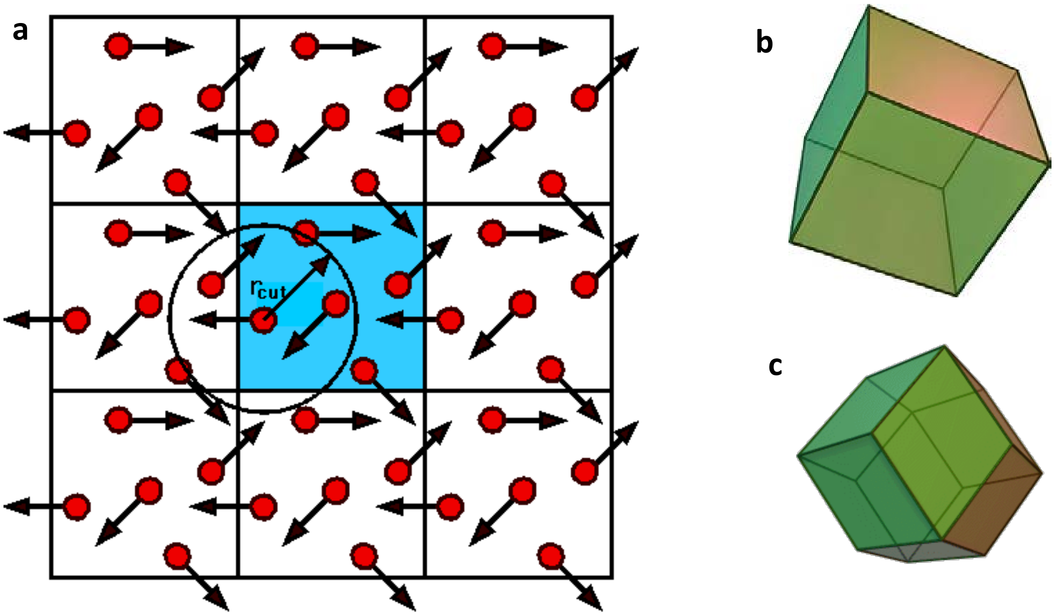


Figure 3.4. (a) Illustration of the periodic boundary conditions in 2D. The blue box represents the system in simulation, while the surrounding boxes are replicates. Both the number of particles and their velocities are identical between the simulation box and the replicate cells. Whenever a particle leaves the simulation box, it is replaced by another with exactly the same velocity, entering from the opposite cell face. (b) Cubic box and (c) Rhombic dodecahedron periodic boundary box in three-dimensional space. (a) is adapted from www.compsoc.man.ac.uk, (b) and (c) are adapted from www. wikipedia.org.

3.4. Force fields.

A force field is a mathematical expression describing the dependence of the energy of a system on the coordinates of the particles. It contains all the parameters allowing calculation of the potential energy terms as described in section 3.1. The quality of a force field is heavily determined by the quality of its parameters. These parameters are generally obtained from quantum mechanical calculations and/or by fitting experimental data, such as neutron, X-ray and electron diffraction, NMR, infrared, Raman and neutron spectroscopy, *etc.*, for a finite set of related compounds. Although a force field cannot completely reproduce all properties of a system, nor is it possible to obtain general parameters that satisfy all molecular structures, most force fields are designed to handle a series of related molecules. Specifically, simulation of globular proteins and IDPs requires different force fields, accompanied by a specific water model as mentioned in section 3.2. There are 4 commonly used force field families to simulate proteins including: CHARMM, AMBER, GROMOS and OPLS. Details on the development, parameterization and current status of the force fields are provided in references [358, 359]. For IDP simulation, CHARMM22*[357], Amber ff99IDPs[360] and Amber ff14IDPs[361] are optimized for disordered protein. The CHARMM36m[362] force field can be used for both folded proteins and IDPs.

3.5. Running a simulation.

In practice, a simulation consists of the following steps:

3.5.1. Setting up the initial conditions:

In this step, initial positions and velocities of the particles, force field and the force-field-associated water model are provided to the simulation package. The initial positions of the particles generally comes from an NMR or X-ray structure. In the case of IDP simulation, linear starting structures may be used to avoid structural bias. The initial particle velocities are randomly generated so that the resulting velocity set $\wp(v_i)$ follows the Maxwell-Boltzmann distribution at a given absolute temperature T:

$$\wp(v_i) = \left(\frac{m_i}{2\pi k_B T}\right)^{\frac{1}{2}} e^{-\frac{m_i v_i^2}{2k_B T}}$$

Generating the velocities randomly may lead to a nonzero net momentum of the system, which would create a systematic translational drift of the system. Thus, we typically adjust the velocities so as to remove this effect.

The periodic boundary of the system is set and the boundary box is then filled with solvent molecules. Counter-ions are added by replacing the same number of random solvent molecules to maintain the neutral charge of the system.

3.5.2. Energy minimization and equilibration.

The energy minimization in this step is to ensure that the solvated molecular system has no steric clashes or inappropriate geometry.

Particles in the system are assigned velocities associated with the low absolute temperature at the initial condition setup step. In order to simulate a system at a desired temperature, it needs to be heated up. The temperature of a system is directly related to the velocities of the particles, and it can be changed by scaling the particle velocity vectors. This method is called the Berendsen thermostat[363]. In order to raise the temperature, a modified method of the Berendsen temperature coupling is used. New velocities corresponding to a slightly higher temperature are periodically assigned to particles, and the simulation is allowed to continue to let the system equilibrate to this new value. This is repeated until the desired temperature is reached. After the temperature stabilizes, the pressure of the system is then equilibrated by using pressure coupling that allows the simulation box (volume of the system) and the particle coordinates to scale.

3.5.3. Production simulation

After the system is stabilized, production simulation is carried out by the simulation package to generate a trajectory of the molecule over a desired simulation time. Theoretically, if we run the simulation long enough, the trajectory will cover the whole potential energy surface (PES), providing a full description of the protein's motion. However, the simulation time is normally limited, therefore only a limited region on the PES is explored. For folded protein, the starting structure for simulation is generally the stable folded conformation residing in the energy well of the PES. Regular simulation times (up to a millisecond) is sufficient to characterize the PES minimum region. A folded protein may change to different conformations but it is required to overcome a large energy barrier, which can only be achieved *via* binding with ligand or a change in experimental conditions. A whole new simulation is needed in this case. When studying IDPs, the proteins may have multiple transient conformations, and the PES may have multiple minima distributed over a broad area. In this case, the simulation may not be able to sample all the conformations of the protein in the study, leading to incorrect interpretation of the dynamic behavior of the protein. To overcome the problem, several production simulation methods have been proposed.

The first and foremost method to enhance sampling is to extend the simulation time by increasing the computation performance with Graphical Processing Units (GPU), a hardware initially developed for graphical processing and the video gaming industry. Simulations on GPUs have resulted in tremendous acceleration of simulations at very low cost[364].

Other approaches involve algorithmic advances. Multicanonical molecular dynamics[365] (McMD) and trivial trajectory parallelization of multicanonical molecular dynamics (TTP-McMD)[366] enhance sampling by doing multiple independent short simulations on different conformations of the system, and aggregating the resultant trajectories in a statistical fashion, resulting in a complete model of the system dynamics. For simulations on a supercomputer cluster with multiple nodes, the later method has resulted in better performance for avoiding the frequent communication between nodes, which is a highly time-consuming step within the simulation process.

Replica exchange molecular dynamics (REMD) (or temperature replica exchange method, tREMD) is another method in which a multiple of chemically identical systems (replicas) are simultaneously simulated at different temperatures. Occasionally, the temperatures of neighboring systems are exchanged (see Figure 3.5). Replicas may overcome conformational energy barriers when their temperatures are high. Although simulations are done at higher than room temperature, the resulting trajectory is a room temperature trajectory.

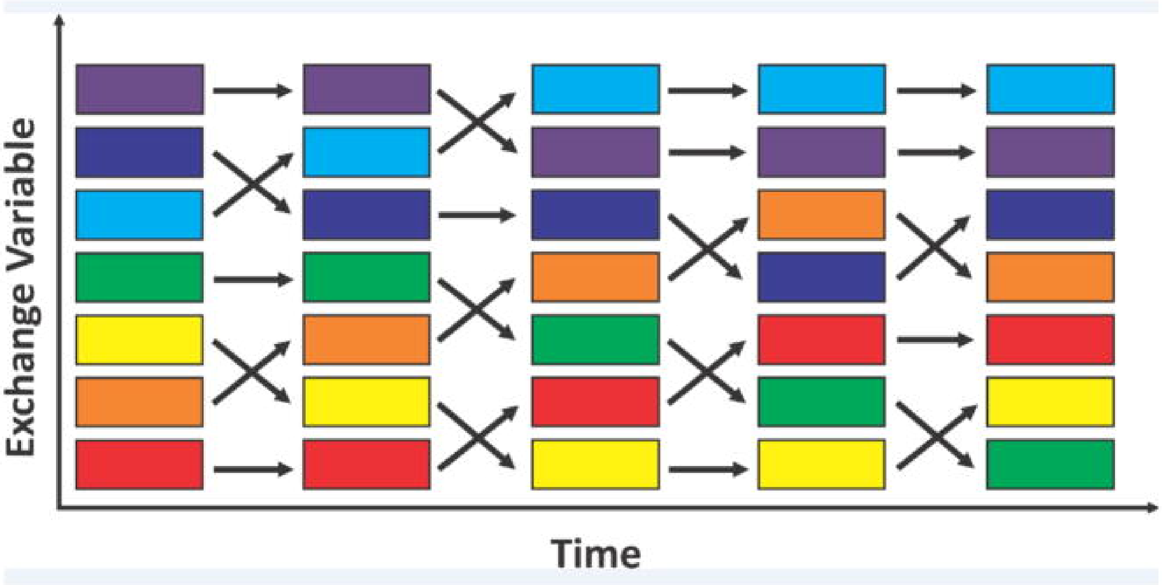


Figure 3.5. Illustration of the replica exchange molecular dynamics (REMD) method. A set of non-interacting replicas runs at different values of an exchange variable, usually temperature (t-REMD). At specific intervals, replicas at neighboring values for the exchange variable are swapped based on a Monte Carlo acceptance criterion. In an efficient run, all trajectories will experience changing of the exchange variable value. At each value for the exchange variable, the trajectories will be discontinuous, but follow a proper Boltzmann distribution for the specific value being exchanged. Reprinted with permission from reference [367]*. Copyright © 2014 Elsevier B.V.*

Metadynamics techniques enhance sampling by adding "memory" into the sampling process, thus preventing oversampling of local energy minima. Once a state has been sampled, a positive Gaussian potential is added to the real energy landscape to discourage the re-sampling of previously visited states. This can be thought of as “filling the free energy wells with computational sand”, and is illustrated in Figure 3.6.

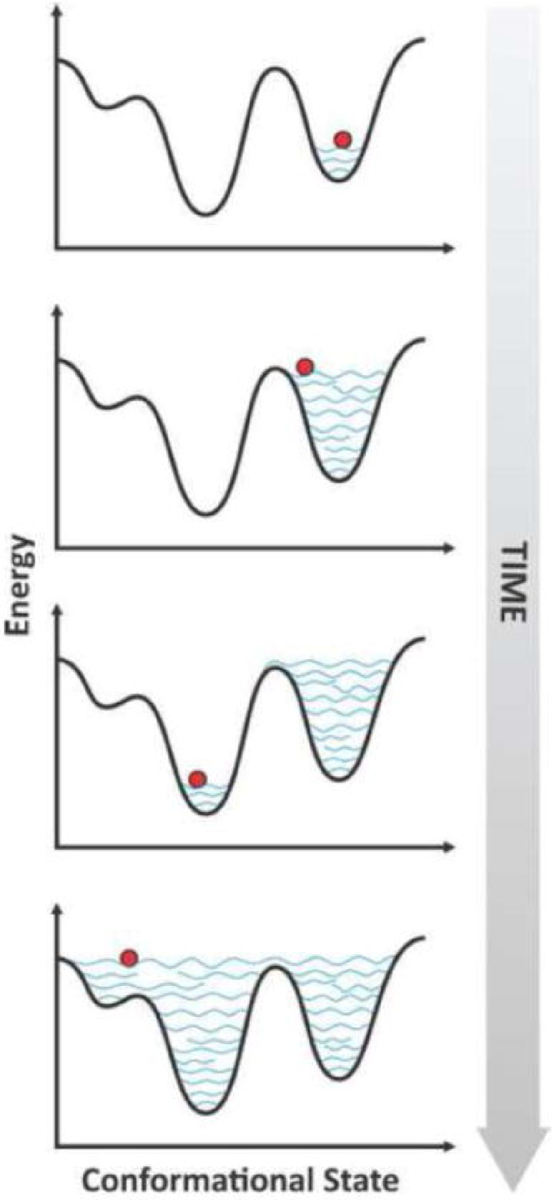


Figure 3.6. Illustration of the metadynamics method. Described as "filling the free energy wells with computational sand"[367]*, the metadynamics method allows the search inside each energy well avoiding an oversampling of the same conformations. When the system reaches a point where the energy is higher than a barrier separating two minima the system goes into a state of lower energy in the new minimum, again searching many possible conformations there. Reprinted with permission from reference* [367]*. Copyright © 2014 Elsevier B.V.*

3.6. Limits of molecular dynamics simulations.

Although it is a powerful tool, molecular dynamics simulation also has some drawbacks. The method is based on classical mechanics. It is acceptable for most atoms at normal temperatures. However, hydrogen is an exception. Hydrogen protons sometimes express quantum mechanical character by tunneling through a potential barrier in the course of a transfer over a hydrogen bond. Such processes cannot be properly treated by classical dynamics. Chemical reactions which involve formation and breaking of bonds cannot be simulated. In classical MD, electronic motions are not considered, the electrons remain in their ground state, and electron transfer processes and electronically excited states cannot occur. Electrons are highly localized meaning that polarization is neglected, and electrons in atoms cannot provide a dielectric constant as they should. Classical simulations do not explicitly include the hydrophobic effect.

Chapter 4

Materials and Methods

4.1. Construction of the full-length Tat$_{101}$ plasmid.

A pET28b plasmid with the *Escherichia coli* codon-optimized gene insert of full-length Tat$_{101}$ protein was constructed by Edis Dzananovic from Dr. Sean McKenna's research group. The insert was created by adding the codon-optimized gene of Tat's second exon to the Tat$_{72}$ gene, which was provided in the plasmid pSV2tat72 by Dr. Alan Frankel[102, 368], using PCR. The DNA was ligated into a pET28b (Novagen, Madison, WI) vector that was then transformed into Top10 Cells (Invitrogen) for storage, and into *E. coli* BL21(DE3) (Invitrogen) for protein expression. The plasmid was sequenced at Manitoba Institute of Cell Biology (MICB) and the sequence of the insert is provided in the Appendix B. The expressed protein has an N-terminal hexahistidine tag that adds an additional 20 residues to the 101-residue Tat. The protein sequence is given below:

MGSSHHHHHH SSGLVPRGSH MEPVDPRLEP WKHPGSQPKT

ACTNCYCKKC CFHCQVCFIT KALGISYGRK KRRQRRRPPQ

GSQTHQVSLS KQPASQPRGD PTGPKESKKK VERETETDPV D

where the his-tag region, the exon 1 and the exon 2 products are in green, red and blue, respectively.

4.2. Expression and purification of full-length Tat_{101} protein.

4.2.1. Culture media and protein expression.

Expression and purification of His-tagged Tat_{101} were adapted from the previously published method[102]. For expression of non-labeled protein, rich medium that contains 25 g/L of LB Broth (Sigma) and 34 μg/ml of kanamycin was used. For protein used in NMR experiments, in addition to uniform labeling, Tat_{101} was prepared with ^{15}N site-specific labeling[225] of lysine, methionine, leucine, and valine, and a ^{14}N unlabeling method[369, 370] was used for lysine, tryptophan, and arginine. In each labeling scheme, the culture was grown in a specific M9 minimal medium. All minimal media were prepared by diluting a 5x M9 salt solution, which contains 15 g of KH_2PO_4, 35 g of Na_2HPO_4, and 2.5 g of NaCl per liter, to working concentration, then autoclaving. For each 1 L of autoclaved solution, 34 mg of kanamycin, 10 mL of MEM Vitamin Solution (Gibco), 2 mL of 1 M $MgSO_4$, 100 μL of 1 M $CaCl_2$, 1 mL of 1000× Trace Metals (Teknova), and other appropriate supplements were added as follow:

- For uniform [^{15}N] Tat, 1 L of M9 minimal medium was supplemented with 1 g of $^{15}NH_4Cl$ and 4 g of D-glucose per liter.

- For uniform [^{13}C, ^{15}N] Tat, M9 minimal medium was supplemented with 1 g of $^{15}NH_4Cl$ and 2 g of [$^{13}C_6$]-D glucose per liter.

- In unlabeling experiments, M9 minimal medium was supplemented with 1 g of $^{15}NH_4Cl$, 4 g of D-glucose, and 0.1 g of the amino acid to be unlabeled per liter.

- For site-specific labeling, M9 minimal medium was supplemented with 0.1 g of a ^{15}N-labeled amino acid and 1 g of each of the 19 other amino acids per liter.

All solid materials were dissolved in water, and filter-sterilized before adding to the medium. In the case of site-specific labeling, amino acids were added to the medium without autoclaving, and the medium was filter-sterilized. All isotope-enriched chemicals were from Cambridge Isotope Laboratories Inc. (Andover, MA).

For non-labelled protein, an overnight culture was grown from a single colony after transformation or from a glycerol stock in 25 ml of rich medium. One ml of overnight culture was added to 1 L of rich medium in a 4 L baffled flask. The culture was grown in an orbital shaking incubator at 37^0C, with shaking at 300 rpm for aeration. When the OD_{600} reached 0.8 - 1.0, 60 mg of isopropyl-beta-D-thiogalactopyranoside (IPTG) (Goldbio) were added to induce protein expression. Cells were harvested 5 hours after induction by centrifugation for 15 minutes at 4^0C and 5000 xg.

For all labeling and unlabeling experiments, 1 ml of overnight culture was added to each of two 4 L baffled flasks that contained 1L of rich medium each. The culture was grown in an orbital shaking incubator at 37^0C and 300 rpm until the OD_{600} was 1.0 - 1.2, following which the cells were harvested by centrifugation for 20 minutes at 4^0C and 2600 xg. Harvested cells were resuspended in and transferred to 1 L of appropriate minimal medium in a 4 L baffled flask that was pre-warmed to 37^0C. The culture was incubated in an orbital shaking incubator at 37^0C and 300 rpm for 30 minutes to let the cells adapt to the new medium. After that, 60 mg of IPTG were added to induce the expression of Tat. Cells were harvested after 5 hours by centrifugation for 15 minutes at 4^0C and 5000 xg.

4.2.2. Purification of full-length Tat$_{101}$ protein.

Harvested cells were resuspended in 100 ml of resuspension buffer at pH 7.2 that contains 100 mM phosphate buffer, 200 μg of DNAse and RNAse each, and 10 mg of lysozyme. This cell suspension solution was subjected to 2 freeze – thaw cycles, and then 3 cycles of sonication at 35 % power with 30 second bursts and 30 seconds between bursts using a sonic dismembrator (Fisher Scientific, Model 500) equipped with a solid disruptor horn. 90 g of guanidine-HCl (Sigma) was added so that the concentration of guanidine was about 6 M (the change in the volume of the solution upon adding a large amount of guanidine was accounted for). Reducing agent, tris (2-carboxyethyl) phosphine hydrochloride (TCEP-HCl or TCEP) (Goldbio), was added to a final concentration of 10 mM, and the pH of the solution was readjusted to 7.2 since it was significantly reduced by both TCEP and guanidine-HCl. The cell lysate was centrifuged for 30 minutes at 4^0C and 12,000 g and the supernatant was kept for purification by metal-affinity chromatography.

A 10 mL polypropylene gravity flow column (QIAGEN Inc.) was packed with 4 mL of Talon™ (cobalt-Superflow™) metal affinity resin (Clonetech). The column was pre-equilibrated with 50 mL of extraction buffer at pH 7.2 that contained 100 mM phosphate buffer, 6 M guanidine-HCl and 10 mM TCEP before addition of the cell lysate supernatant described above. The resin was then washed with 20 mL of extraction buffer, followed by 30 mL of washing buffer at pH 6.4 containing 50 mM phosphate buffer, 6 M guanidine-HCl and 10 mM TCEP. Tat$_{101}$ protein was eluted from the column with 20 mL of elution buffer at pH 4.0, containing 20 mM acetate buffer and 10 mM TCEP.

Protein in elution buffer was dialyzed against 1 L of degassed acetate buffer at pH 4.0 at concentrations of 0.1 M, 0.1 M, 0.05 M, and 0.01 M (approximately 6 hours each). The dialysis buffer in the first step was supplemented with up to 10 mM EDTA to chelate any cobalt metal that might have leaked from the column. This adaptation came from the observation that resonances in the His-tag region of HSQC spectra of the Tat_{101} protein that had not been dialyzed against EDTA-containing buffer were broadened in comparison with the resonances of those dialyzed against EDTA. After the dialysis step, the dialysate was frozen and freeze-dried. Dry protein was stored at -20^0C before being used in later experiments.

An attempt to remove the hexahistidine affinity tag segment was made by treating the purified protein with Thrombin protease. However, besides the primary cutting site (L_{14}VPRGS_{19}), there, apparently, was a non-specific secondary cutting site for Thrombin within the protein sequence as the sample treated with this enzyme showed multiple bands on an SDS gel (data not shown). In another attempt to remove the affinity tag, the encoding gene was modified to change the cleavage site into a Tobacco Etch Virus (TEV) protease cutting site. However, the TEV protease performance at low pH condition (pH 4.0) was too low, making it impossible to remove the affinity tag from Tat protein with this enzyme. Therefore, all of the following studies were conducted on Tat protein with the His-tag present.

4.3. Expression and purification of cyclin T1 for Tat_{101} – cyclin T1 complex formation attempt.

An *Escherichia coli* codon-optimized synthetic gene of human cyclin T1 (1-266)[165] was purchased from GenScript, and ligated into the pET28a vector (Novagen, Madison, WI). The plasmid was then transformed into Top10 Cells (Invitrogen) for storage, and into *E. coli* BL21(DE3) (Invitrogen) for protein expression. The expressed protein has an N-terminal hexahistidine tag, a short linker and a TEV digestion site that adds an additional 44 residues to the 266-residue cyclin T1. The protein sequence is given below:

MGSSHHHHHHSSGLVPRGSHMASMTGGQQMGRGSDGPENLYFQGMEGERKNN
NKRWYFTREQLENSPSRRFGVDPDKELSYRQQAANLLQDMGQRLNVSQLTINTAIVYM
HRFYMIQSFTQFPGNSVAPAALFLAAKVEEQPKKLEHVIKVAHTCLHPQESLPDTRSEAY
LQQVQDLVILESIILQTLGFELTIDHPHTHVVKCTQLVRASKDLAQTSYFMATNSLHLTTF
SLQYTPPVVACVCIHLACKWSNWEIPVSTDGKHWWEYVDATVTLELLDELTHEFLQILE
KTPNRLKRIWNWRACEAAKK

An overnight culture of transformed cells was grown from a single colony on an LB-agar plate after transformation or from a frozen glycerol stock in 25 ml of rich medium that contained 25 g/L of LB broth (Sigma) and 34 µg/ml of kanamycin. One ml of overnight culture was added to 1 L of rich medium in a 4 L baffled flask. The culture was grown in an orbital shaking incubator at 37^0C, shaking at 300 rpm for aeration. When the OD_{600} reached 0.8 - 1.0, 60 mg of IPTG (Goldbio) were added to induce protein expression. Cells were harvested 5 hours after induction by centrifugation at 5000 xg for 15 minutes at 4^0C.

Harvested cells was resuspended in 50 ml of resuspension buffer at pH 7.2 that contained 50 mM HEPES buffer, 300 mM NaCl, 2 mM TCEP, 200 µg of DNAse and RNAse each, and 10 mg of lysozyme. This solution was subjected to a freeze – thaw cycle, and then 10 cycles of sonication on ice at an amplitude of 50 % with 30 second bursts and 30 seconds between bursts using a sonic dismembrator (Fisher Scientific, Model 500) equipped with a solid disruptor horn. The cell lysate was centrifuged at 4^0C and 12,000 xg for 30 minutes. The supernatant was then loaded onto a 10 mL polypropylene gravity flow column (QIAGEN Inc.) packed with 4 mL Talon™ (cobalt-Superflow™) metal affinity resin (Clonetech), that had been pre-equilibrated with

50 mL of extraction buffer. Extraction buffer was at pH 7.2 and contained 100 mM HEPES buffer, 300 mM NaCl and 2 mM TCEP. The resin was then washed with 50 mL of extraction buffer, followed by washing buffer at pH 7.2, containing 20 mM HEPES buffer, 300 mM NaCl, 30 mM of imidazole and 2 mM TCEP, until the eluate OD_{280} reached about 0. Cyclin T1 protein was eluted from the column by addition of 20 mL of elution buffer at pH 7.2, containing 20 mM HEPES buffer, 300 mM NaCl, 150 mM of imidazole and 2 mM TCEP.

4.4. NMR measurements and data analysis.

4.4.1. NMR sample preparation.

Freeze-dried protein was dissolved in degassed NMR buffer containing 10 mM acetic acid-d_4, pH 4.0, 7% D_2O, 2 mM TCEP, and 75 µM 2,2-dimethyl-2-silapentane-5-sulfonate (DSS). For NMR samples of Tat_{101} prepared at neutral pH, the NMR buffer contained 10 mM HEPES instead of acetic acid-d_4. In all NMR experiments, DSS served as an internal reference in such a way that the ^1H water signal at 293 K resonates at 4.821 ppm relative to DSS. ^{15}N and ^{13}C referencing were done indirectly relative to DSS[242]. All NMR experiments reported in this thesis were done on samples at pH 4.0, unless specifically indicated otherwise. The low pH was necessary to suppress inter- and intra-molecular disulphide bond formation in Tat.

Tat_{101} protein was added into an NMR tube and carefully degassed and purged with argon for 15 minutes. The NMR tube cap was sealed with Teflon tape. A ^{13}C and ^{15}N-labeled sample was prepared for SOLEXSY hydrogen-deuterium exchange measurements[330] by dissolving the protein in 50% H_2O/D_2O. The pH of this sample was corrected by 0.2 units (reading pH of 4.2 for a pH 4.0 sample) to account for the glass electrode isotope effect in the 50% H_2O – 50% D_2O solvent[330, 371, 372].

The protein concentration of NMR samples was estimated to be 500 μM using UV spectroscopy on a NanoDrop 2000c spectrophotometer (Thermo Scientific™). The extinction coefficient at 280 nm was estimated to be 8250 M^{-1} cm^{-1} using Protein Calculator 3.4 (http://protcalc.sourceforge.net).

4.4.2. NMR experiments for backbone assignment and $^3J_{HNH\alpha}$ measurement.

All NMR experiments for backbone assignments and $^3J_{HNH\alpha}$ measurements were conducted on a Varian INOVA 600 MHz NMR spectrometer equipped with a triple-resonance probe head, at 25 °C, on a reduced uniform ^{13}C- and ^{15}N-labeled Tat_{101} sample at pH 4.0, using the following standard Agilent/Varian BioPack pulse sequences: HSQC, HNCO, HN(CA)CO, HNCA, CBCA(CO)NH and HNHA[373-376] (see Table 2.1). All backbone assignment experiments were sensitivity-enhanced and used pulse-field gradients for coherence selection and water suppression. Radiation damping was suppressed by a water flip – back pulse. ^{15}N decoupling during acquisition was done with the WALTZ-16 sequence[377]. Recycling delays for all experiments were 0.7 s. The number of transients, complex points, and sweep widths for the experiments are provided in Table 4.1. NMR data were processed with NMRPipe[378]. All spectra were apodized using a squared cosine-bell function, zero-filled to twice the data set size, and linear predicted before Fourier transformation. Sequential backbone assignments were done manually, using SPARKY[379, 380]. $^3J_{HNH\alpha}$ coupling constants[381] were calculated as follows:

$$\frac{I_{cross}}{I_{diag}} = -tan^2\left(2\pi J_{H_N H_\alpha}\zeta\right)$$

where I_{cross} and I_{diag} are the intensities of cross peaks and diagonal peaks, respectively, in the HNHA experiment, and 2ζ=26.3 ms which is the rephasing delay.

Chemical shift differences for each residue along the protein sequence were determined by subtracting the chemical shifts from the backbone assignment by the random coil values, taken from the ncIDP (neighbor corrected IDP) Library by Tamiola *et al.*[256]. Secondary structure propensity was calculated using the SSP program[264], kindly provided by J. Forman-Kay's group, Hospital for Sick Children Research Institute, Toronto.

Table 4.1. Parameters for backbone assignment NMR experiments.

Experiment	Transients	Complex points	SW[^{1}H] ppm	SW[^{13}C] ppm	SW[^{15}N] ppm
^{1}H/^{15}N HSQC	16	1024x128	16		32
HNCO	4	1024x64x64	16	25	32
HN(CA)CO	8	1024x64x32	15	16	33
HNCA	8	1024x64x64	16	30	32
CBCA(CO)NH	8	1024x64x64	16	80	32
HNHA	8	1024x128x32	16/16		32

4.4.3. Nuclear spin relaxation.

Relaxation data were collected on reduced, uniform ^{15}N His-tagged Tat, using an Agilent/Varian INOVA 600 MHz NMR spectrometer at the University of Manitoba equipped with a triple resonance probe head, and on an Agilent/Varian INOVA 800 MHz NMR spectrometer at the University of Alberta (NANUC, Edmonton, AB) using a triple-resonance cold probe with the assistance of P. Mercier. For R_1 measurements, 10 data sets were acquired using relaxation delays of 10, 200, 400, 600, 800, 1000, 1200, 1600, 1800, and 2200 ms. For R_2 measurements, 12 data sets with relaxation delays of 10, 30, 50, 70, 90, 110, 130, 150, 170, 190, 210, and 230 ms were acquired. The post-acquisition delay for each experiment was 5s. For NOE measurements, 2 spectra were collected with and without the 5 s proton pre-saturation period at the beginning of the experiment. All experiments used pulse-field gradients for coherence selection and water suppression. The number of transients, number of complex points, and sweep widths for the experiments are provided in Table 2.2.

Table 4.2. Parameters for nuclear spin relaxation NMR experiments.

Experiment	Transients	Complex points	SW[^{1}H] ppm	SW[^{15}N] ppm
R_1	16	1024x128	16	32
R_2	16	1024x128	16	32
NOE	128	1024x64	16	32

NMR data were processed with NMRPipe[378]. All spectra were apodized using a squared cosine-bell function and zero-filled to twice the data set size before Fourier transformation. Cross-peak intensities were measured as peak heights and fit to a two parameter exponential decay using the Relax[281] program to extract relaxation rates (R_1 and R_2). The steady-state $^1H–^{15}N$ NOE values were obtained from the ratios of peak heights from experiments with (I_{NOE}) and without (I_{noNOE}) proton saturation. Errors were calculated using the signal-to-noise ratios of the NMR peak heights where the noise level was estimated using SPARKY[379, 380].

Data from the R_1, R_2, and NOE experiments were analyzed by reduced spectral density mapping[273, 290] using *Mathematica*© notebooks kindly provided by L. Spyracopoulos[382] (University of Alberta) and by Model-Free analysis[221, 222] using a *Mathematica*© notebook provided by S. Shojania (University of British Columbia, Vancouver, BC). The errors in R_1, R_2, and NOE measurement were estimated from the spectral baseline noise. Error analysis was done using 500 iterations of a Monte Carlo simulation for the spectral density mapping and 100 iterations for the Model-Free analysis.

4.4.4. Relaxation dispersion.

Relaxation dispersion experiments were conducted at 600 MHz (University of Manitoba), and 800 MHz (NANUC) with the assistance of P. Mercier. The constant relaxation time delay (T_{cp}) at 600 MHz was 40 ms and at 800 MHz was 80 ms. The Carr− Purcell−Meiboom−Gill pulse chain was applied with rates (ν_{CPMG}) of 50, 100, 200, 300, 400, 500, 600, 700, 800, 900, and 1000 Hz[383, 384]. Each data point was collected with 16 transients and 1024x64 complex points; the sweep widths on the proton and nitrogen dimensions were 16 and 32 ppm, respectively. Data were processed with NMRPipe[378], and the extracted peak heights were used in the analysis using the relaxation dispersion program NESSY[385].

4.4.5. pH titration.

In the pH titration experiment, the starting NMR sample was prepared at pH 7.0 (see 2.3.1), and the titration was done by gradually reducing the pH of the sample using acetic acid-d_4. The reason for doing the titration by reducing the pH was because the protein readily aggregates upon the addition of base (NaOH). HSQC spectra were acquired on samples at pH 7.0, 6.0, 5.0 and 4.0. Data were processed using NMRPipe[378], and overlaid using SPARKY[379, 380].

4.4.6. Hydrogen exchange.

CLEANEX-PM hydrogen exchange measurements[329, 331] were made on a Varian 600 MHz NMR spectrometer with eight time delays of 10, 20, 35, 50, 75, 100, 125, and 150 ms. SOLEXSY hydrogen exchange measurements were made on a Bruker 800 MHz NMR spectrometer at the University of British Columbia with the assistance of S. Shojania. Eight data points were collected corresponding to time delays of 1.2, 31.2, 61.2, 101.2, 141.2, 201.2, 501.2, and 1001.2 ms. Parameters for the NMR experiments are provided in Table 2.3.

Table 4.3. Parameters for NMR experiments to measure fast hydrogen-deuterium exchange.

Experiment	Transients	Complex points	SW[^{1}H] ppm	SW[^{15}N] ppm
CLEANEX-PM	32	1024x128	16	32
SOLEXSY	16	1024x386	16	24

By using interleaved acquisitions, each collected data point contained 2 data sets (A and B) corresponding to the positive and negative proton-frequency labeled magnetization of nitrogen. The separation, addition (A+B), and subtraction (A-B) of the 2 data sets were done using NMRPipe. Axial and cross-peak resonances (N^H and N^D peaks) were quantified by measuring peak heights.

For both experiments, exchange rates were calculated by fitting the peak heights to appropriate hydrogen exchange models (see 2.4.3) using *Mathematica*© notebooks that I authored. Error analysis was done using 500 iterations of Monte Carlo simulation and spectral noise. Rates of exchange for unstructured Tat, corrected for sequence-dependent inductive effects and under the same conditions of pH and temperature that were used in the hydrogen exchange measurements, were calculated using a program provided by W. Englander (University of Pennsylvania, Philadelphia, PA)[305, 310]. Protection factors were calculated by taking the ratio of the predicted exchange rates for unstructured Tat_{101} to the measured rates.

4.5. Molecular dynamics simulations.

Simulation of the first exon product of Tat (Tat_{72}) was done using GROMACS 5.1.2 software with the CHARMM22* force field and modified TIP3P water model[357] (mTIP3P). A linear starting structure of the molecule was used to avoid structural bias during the simulation. The structure was built using the Avogadro program[386].

Protein was solvated in a rhombic dodecahedron box of mTIP3P water and Na^+ or Cl^- as the counter ions. The electrostatic interactions were treated using the PME method[343, 387] with a grid spacing for Fast Fourier Transform (FFT) of 0.16. The short-range van der Waals interaction cutoff was 1.0 Å. The energy and pressure were corrected for long-range dispersion interactions. The system was coupled to the modified Berendsen thermostat and Parrinello-Rahman barrostat as described in section 3.5. The integration step time was set to 2 fs. All bonds were constrained using the LINCS algorithm. After a 5000-step energy minimization, the system was equilibrated at 300 K and 1 bar. The last snapshot was used as the initial structure for subsequent productive NPT simulation where the number of atoms (N), pressure (P) and temperature (T) remained constant throughout the simulation. Structures were stored every 10 ps. The simulation was done in duplicate resulting in 2 trajectories of 100 ns each.

The radius of gyration profiles and the secondary structure for each time frame were calculated from the resulting trajectories using the built-in tools of GROMACS (*gmx gyr*ate and *gmx do_dssp*, respectively). The autocorrelation function profile that describes the internal motion of each N-H bond vector was also calculated by using the GROMACS tool *gmx rotacf* after removing the rotational and translational motions of the protein by aligning each frame to a reference structure. The autocorrelation function values were fit to the Lipari - Szabo model-free model, using a *Mathematica*$^{©}$ note book that I authored, to extract the per-residue order parameters which were then compared with NMR results.

Chapter 5

Results and Discussion

5.1. Protein expression and purification.

The full-length Tat_{101} gene was cloned into the pET28b vector, enabling protein expression with an N-terminal, thrombin-cleavable, hexahistidine purification tag. The yield of unlabeled protein was typically about 20 mg per liter of culture. For labeled protein, the yield was reduced to about 15 mg per liter of minimal medium culture. High purity level was achieved with metal affinity chromatography method. The SDS-PAGE gel image in Figure 5.1 shows a thick band at about 13.6 kDa, corresponding to the molecular weight of His-tagged Tat_{101}. The second band at around 27 kDa corresponds to the oxidized His-tagged Tat_{101} which seems to form a dimer. The formation of the oxidized protein is because the gel was prepared and run at neutral pH, allowing the oxidation of a small portion of the protein. A cobalt affinity column was used instead of a nickel column because this reduced the non-specific binding of other proteins to the resin. An attempt was made to purify Tat_{101} protein on a zinc column; however, there was no sign of protein coming off the column during the elution step (OD_{280} = 0). The combined use of a strong reducing agent (TCEP), denaturing conditions (6 M guanidine-HCl) and elution at pH 4.0 instead of using buffer that contains imidazole effectively prevented protein aggregation and disulphide bond formation. The modified dialysis step that added 10 mM EDTA to the buffer helped by chelating the cobalt metal that leaked from the column. This adaption came from the observation that resonances in the His tag region of HSQC spectra of the protein that had not been dialyzed against

EDTA-containing buffer were broadened in comparison with the resonances of those dialyzed against EDTA. This also prevented the possibility of interaction between the zinc-finger in the cysteine rich region and the leaked metal.

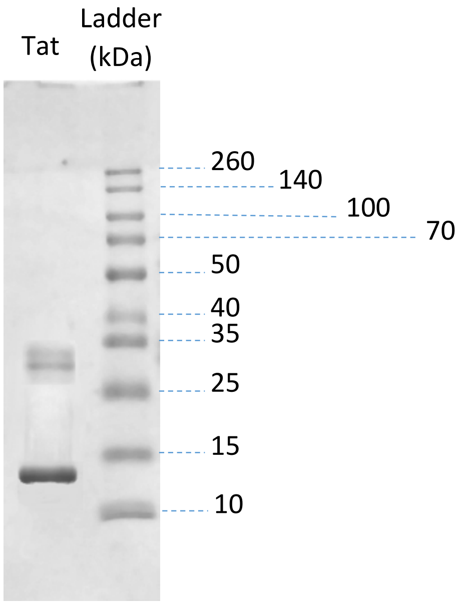


Figure 5.1. SDS-PAGE gel image. Purified Tat$_{101}$ protein in 10 mM acetate buffer at pH 4.0 is on the left lane and protein ladder is on the right lane.

Interestingly, the shorter version of Tat lacking the second exon product, Tat$_{72}$, more readily aggregated during the final dialysis step that was therefore limited to 4 hours. In contrast, the last dialysis step for full-length Tat$_{101}$ could be left for 2 days without visible aggregation. When preparing solutions of high concentrations of Tat$_{72}$ and Tat$_{101}$, large amounts of each protein were dissolved in 10 mM HEPES buffer at pH 7.0; then the solutions were subjected to 30 minutes of centrifugation at 15,000 xg. OD$_{280}$ measurement of Tat$_{101}$ solution showed higher absorption (OD$_{280, \text{Tat101}}$ = 6.389) than Tat$_{72}$ solution (OD$_{280, \text{Tat72}}$ = 4.504). The second exon product, therefore, appears to increase the solubility of the protein. The expression and purification of Tat_{72} was adopted from reference [102].

5.2. NMR resonance assignment.

Figure 5.2 shows an assigned two-dimensional (2D) ^1H–^{15}N HSQC spectrum of fully reduced ^{13}C- and ^{15}N-labeled Tat_{101} at pH 4.0 and 298 K. The low dispersion of resonances and the presence of proline residues in relatively high abundance made the assignment of the backbone chemical shifts challenging. Site-specific labeling and unlabeling methods were used as a complement to three-dimensional (3D) experiments (see Figures 5.3, 5.4, and 5.5) to support the sequential assignment process (Figure 5.6) that was also aided by the previously assigned first-exon product[160]. The well-known site-specific labeling method[388] greatly reduces the extent of spectral crowding and provides unambiguous residue-type assignments. ^{15}N-Val and ^{15}N-Leu labeling provided keystones for the assignment of full-length Tat. Difficult-to-assign residues, including M21 and several lysine residues, were also identified using this method (Figure 5.3). In unlabeling experiments[369], peaks of amino acids to be unlabeled disappear or have intensities much weaker than those of the peaks in the spectra of the uniformly labeled protein. This strategy is very inexpensive in comparison with the site-specific labeling experiments of the same amino acid but still provides a similar amount of information about the residue types of the peaks. Unlabeling can also reduce the extent of spectral crowding and reveal overlapped peaks. W31 and several arginine and lysine residues were identified using this method. Spectra can be found in Figures 5.4 and Figure 5.5.

As a result of these experiments, 104 of 106 expected backbone ^1H and ^{15}N amide resonances (97.8%) were identified, not including the 13 Pro residues and the two N-terminal amino acids (M1 and G2) that have proton amine and amide exchanging too fast to be observed. Other backbone atoms were assigned at 98.3% of C_α, 83.9% of C_β, and 95.8% of C' resonances. Missing ^{15}N-assignments correspond to K71 and R72. Backbone N, H_N, H_α, C_α, and C′ and side-chain C_β chemical shift assignments are listed in Table A.1 of the Appendix.

In general, the ^1H–^{15}N HSQC spectrum of full-length Tat_{101} protein has typical characteristics of denatured and disordered proteins. Cross peaks are clustered in several distinct regions: a Gly region, a Ser/Thr region, and a region containing the rest of the backbone amides. The spectral dispersion of the resonances is also typical for proteins lacking regular secondary structure in that all backbone resonances lie within a 1.4 ppm range in the ^1H dimension and within 18 ppm in the ^{15}N dimension.

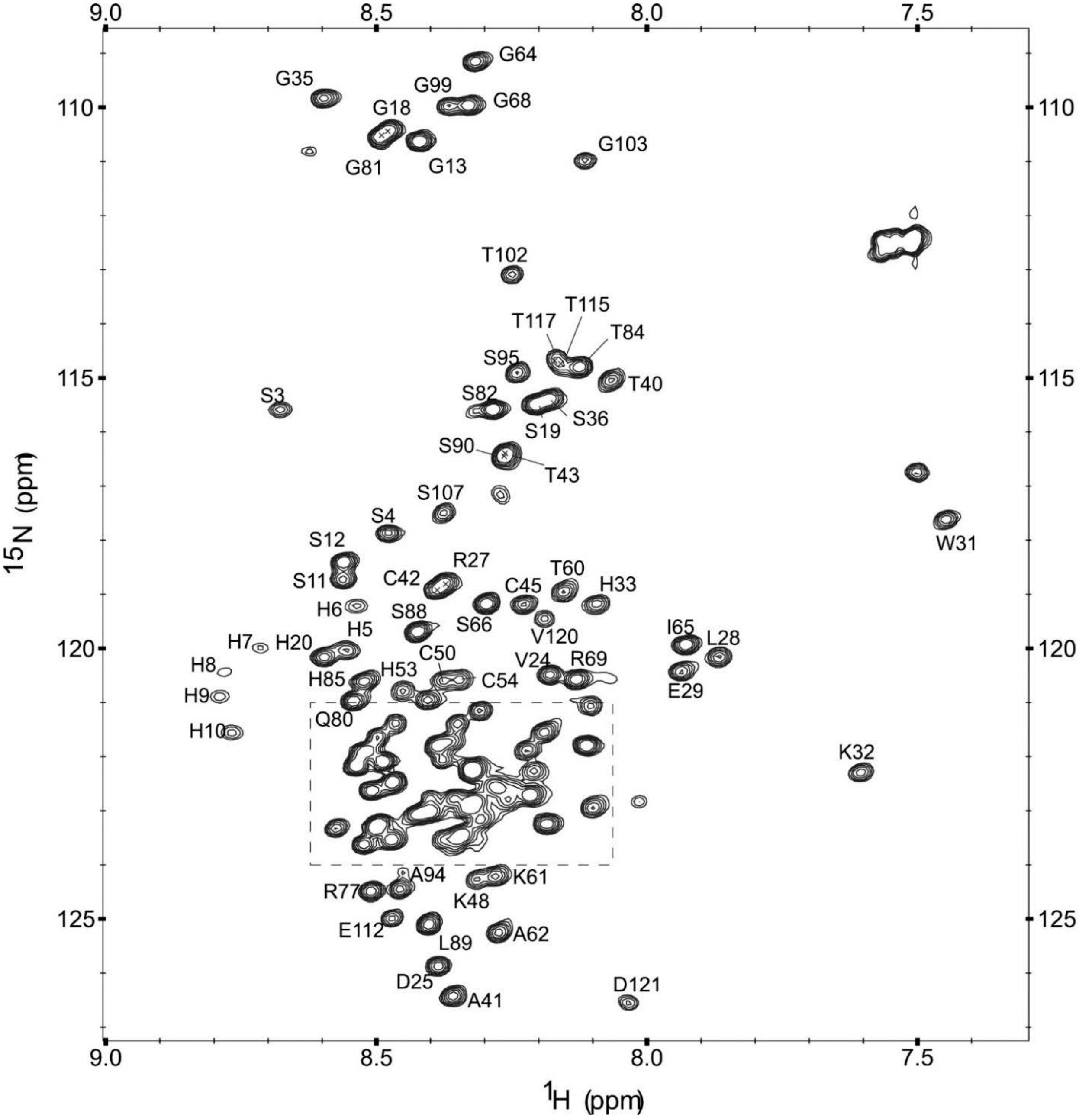


Figure 5.2a. Backbone ^{15}N and $^{1}H^{N}$ assignments for full-length Tat_{101} protein from HIV-1. ^{1}H–^{15}N HSQC spectrum of the fully reduced, uniformly ^{13}C- and ^{15}N-labeled protein sample at pH 4.0 and 298 K.

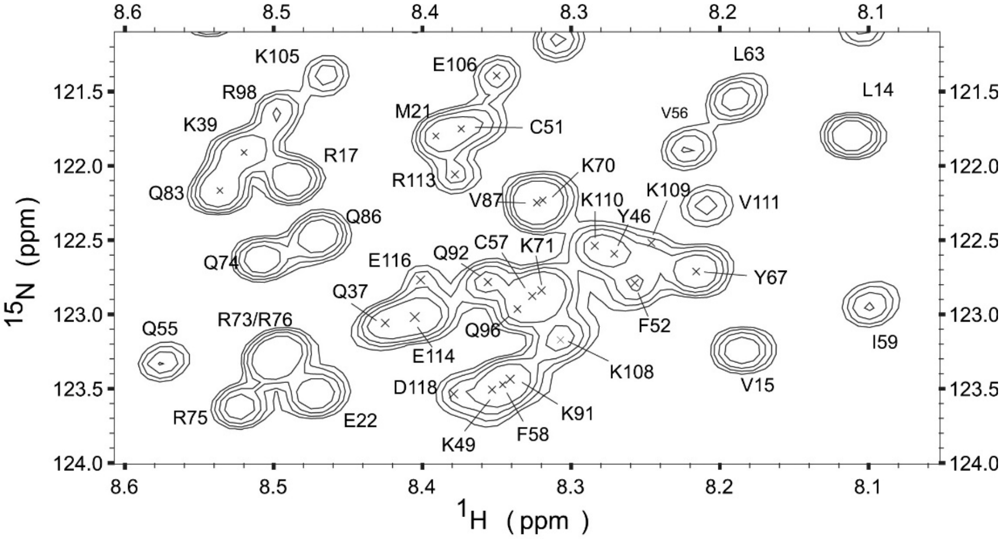


Figure 5.2b. Backbone ^{15}N and ^{1}H^{N} assignments for full-length Tat$_{101}$ protein from HIV-1. The dashed region of Figure 3.2a is expanded to allow labelling of the crowded central region of the spectrum.

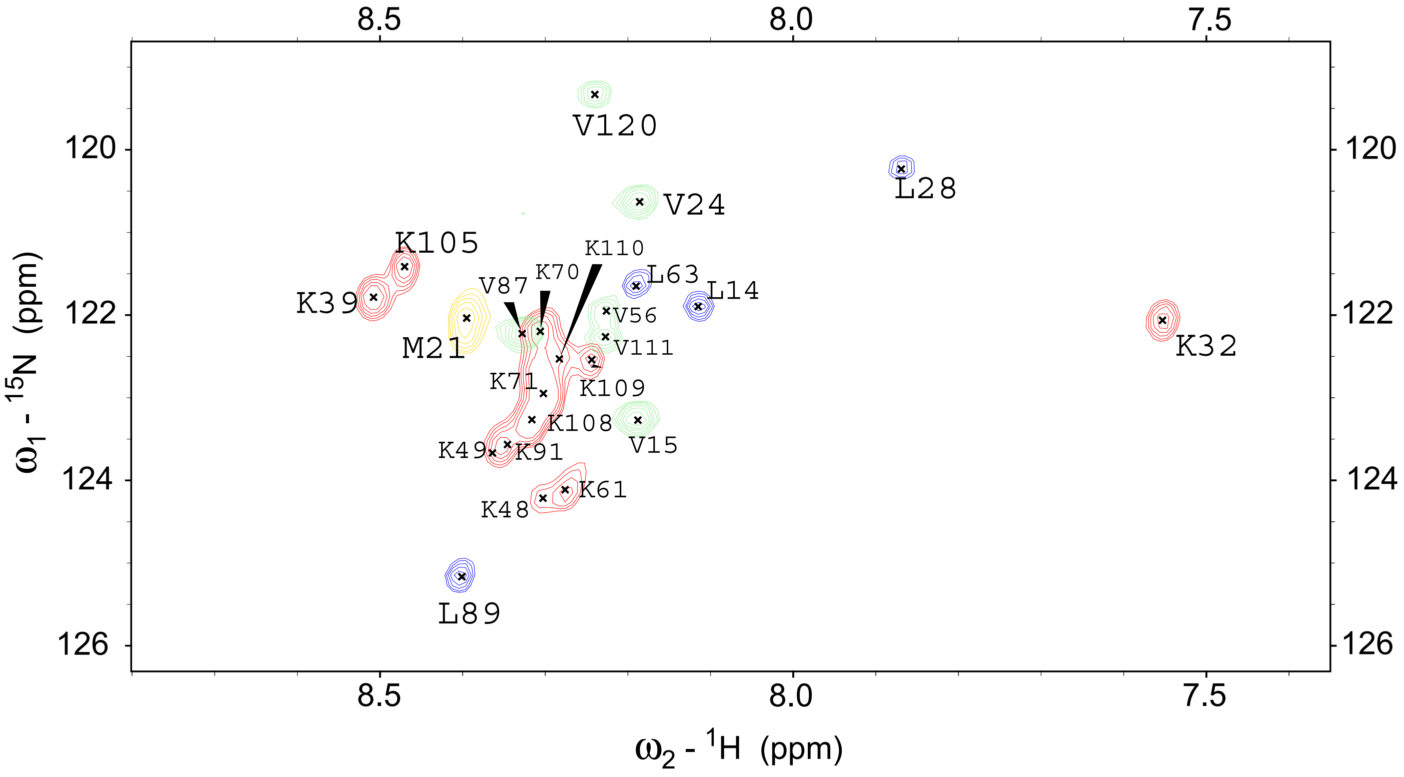


Figure 5.3. Overlay of the HSQC spectra of site-specifically labeled ^{15}N-Leu (blue), ^{15}N-Lys (red), ^{15}N-Met (yellow), and ^{15}N-Val Tat$_{101}$ (green).

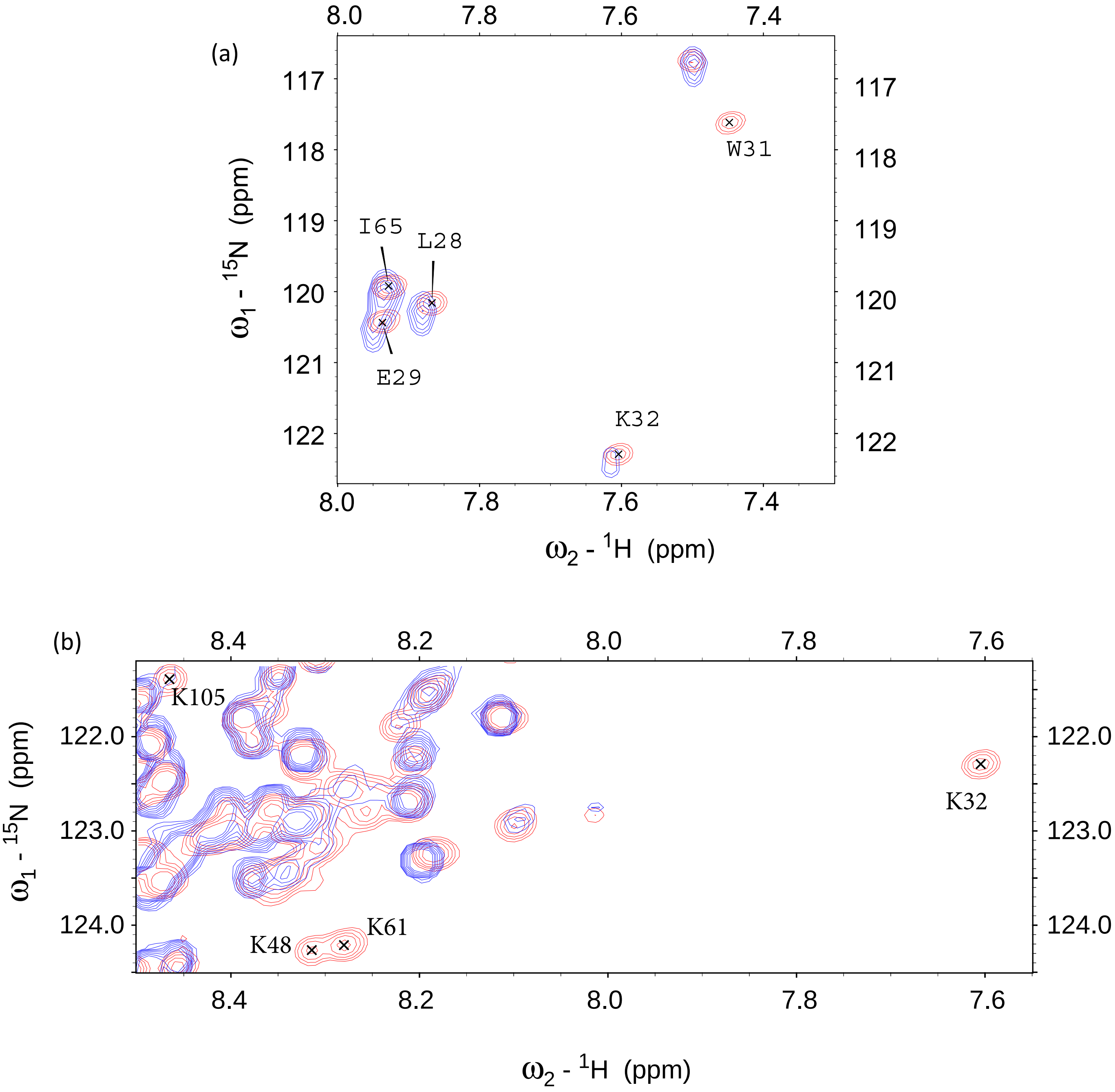


Figure 5.4. The blue peaks are from ^{1}H-^{15}N HSQC spectra of Tat_{101} protein unlabelled with (a) ^{14}N-Trp and (b) ^{14}N-Lys overlaid onto a uniformly ^{15}N-labelled spectrum in red. The corresponding peaks are completely eliminated or have intensities lower in the unlabeled spectrum compared to the uniformly ^{15}N-labelled spectrum.

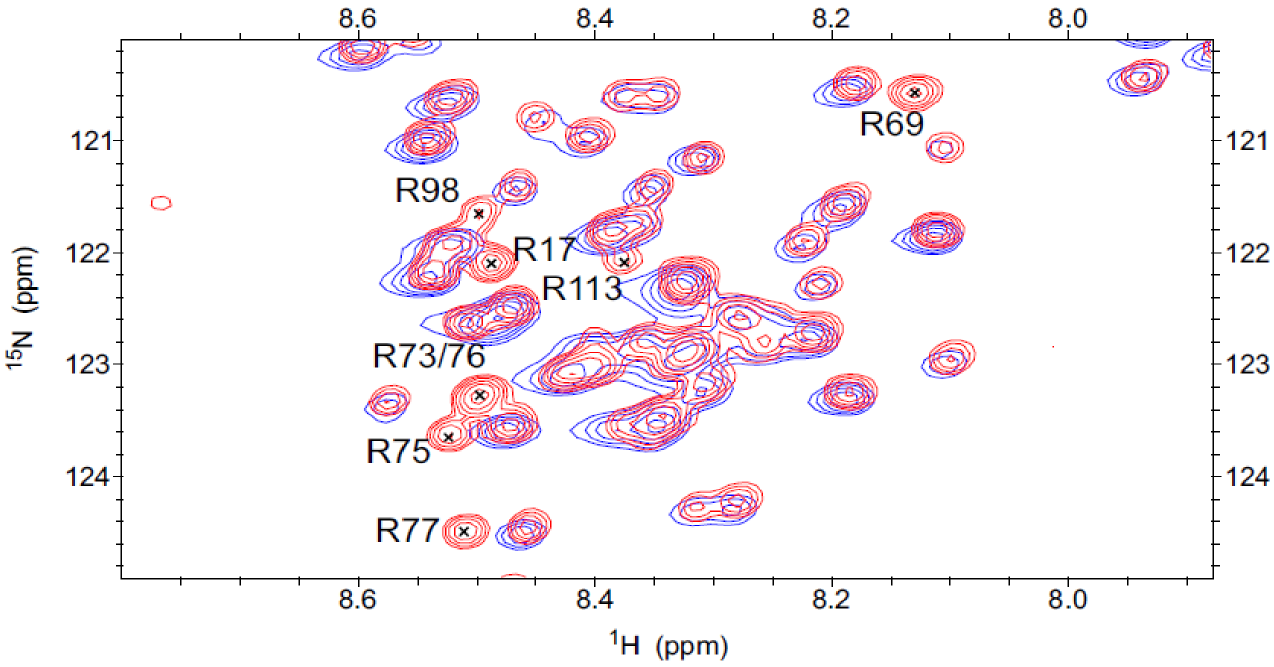


Figure 5.5. Regions of ^{1}H-^{15}N HSQC spectra of fully reduced full-length Tat$_{101}$ protein uniformly ^{15}N-labelled and ^{14}N–Arg unlabelled experiments. The unlabeled Tat$_{101}$ spectral region (in blue) is overlaid onto the same spectral region of a uniformly labeled sample (in red).

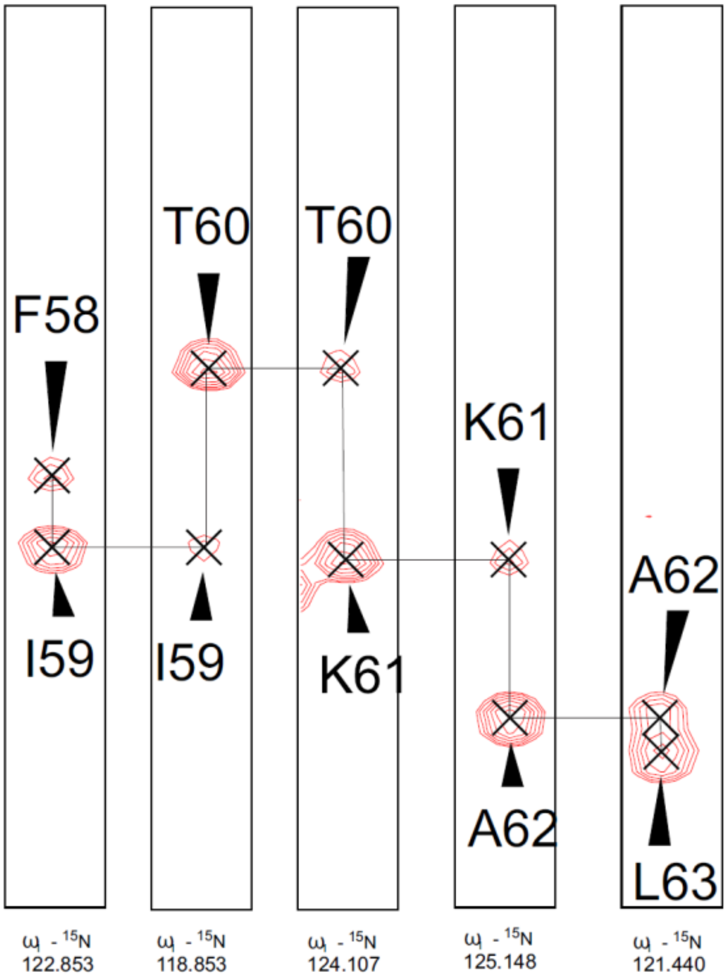
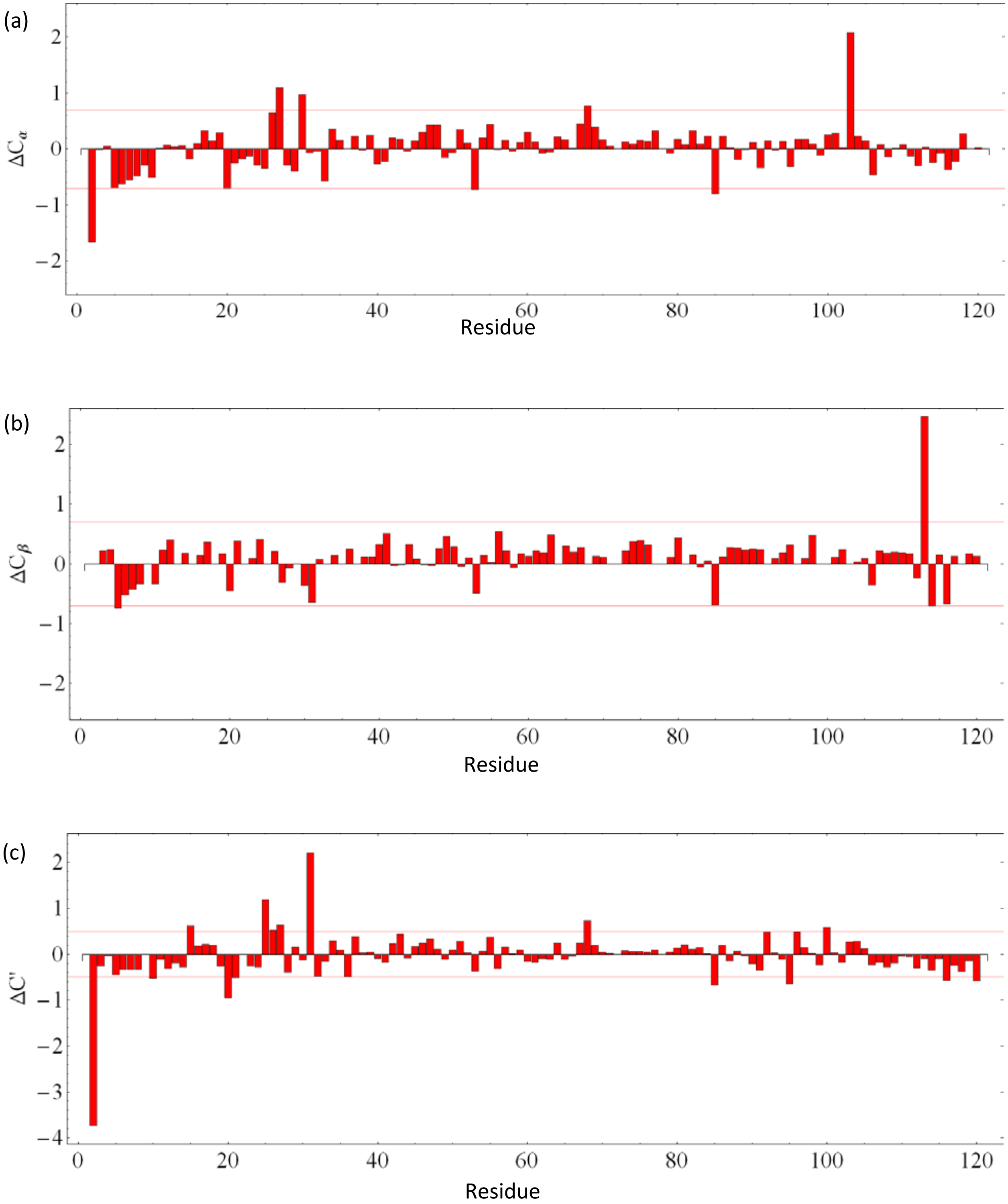


Figure 5.6. Sequential assignment using 3D - NMR experiments. Strip plots from an HN(CA)CO spectrum of Tat$_{101}$ *protein taken at pH 4.0 and 293K shows the intra-residual correlations between HN(i) and N(i) with C'(i) and C'(i-1) resonances. The strips correspond to different* ^{15}*N-planes in the 3D experiment. A segment of Tat*$_{101}$ *is shown depicting the connectivity between residues 59-63. All NMR spectra were obtained on the Varian 600 MHz spectrometer at the University of Manitoba.*

5.3. Chemical shift differences, J-coupling constants and secondary structure propensities.

Protein backbone chemical shifts are a highly reliable indicator of protein secondary structure[244]. The secondary chemical shift, the deviation of measured chemical shifts from their random coil values, has long been considered an indication of the tendency of a protein to adopt a local conformation[243-245, 256, 262, 263, 389]. In a biased disordered state, a polypeptide may occupy some secondary structures more frequently than others. This structural propensity can sometimes also be detected by changing the temperature of a protein or by adding cosolvents[390] .

Figure 5.7 shows the individual secondary chemical shift differences (C_α, C_β, C′, ^{15}N, H_α, and H_N) from the random coil values obtained from the neighbor-corrected Intrinsically Disordered Protein Library (ncIDP)[256]. The results indicate that a majority of the resonances in Tat_{101} are within the random coil range and there are no regions with more than three consecutive resonances in the α-helix or β-sheet chemical shift ranges. The results are in good agreement with the measurements on Tat_{72}[102]. This shows that the second exon product, residues 73-101, has little effect on the conformation on the first exon product. These results also show, for the first time, that the second exon product adopts a predominantly disordered conformation as expected from the prediction algorithms (Section 5.4). These conclusions are in good agreement with a prediction done on the CSI 3.0 server[263] indicating a random coil structure and fully flexible backbone for HIV-1 full-length Tat_{101} (Figure 5.8).


(a)
ΔCα
2
1
0
−1
−2
0
20
40
60
80
100
120
Residue
(b)
ΔCβ
2
1
0
−1
−2
0
20
40
60
80
100
120
Residue
(c)
ΔC'
2
1
0
−1
−2
−3
−4
0
20
40
60
80
100
120
Residue

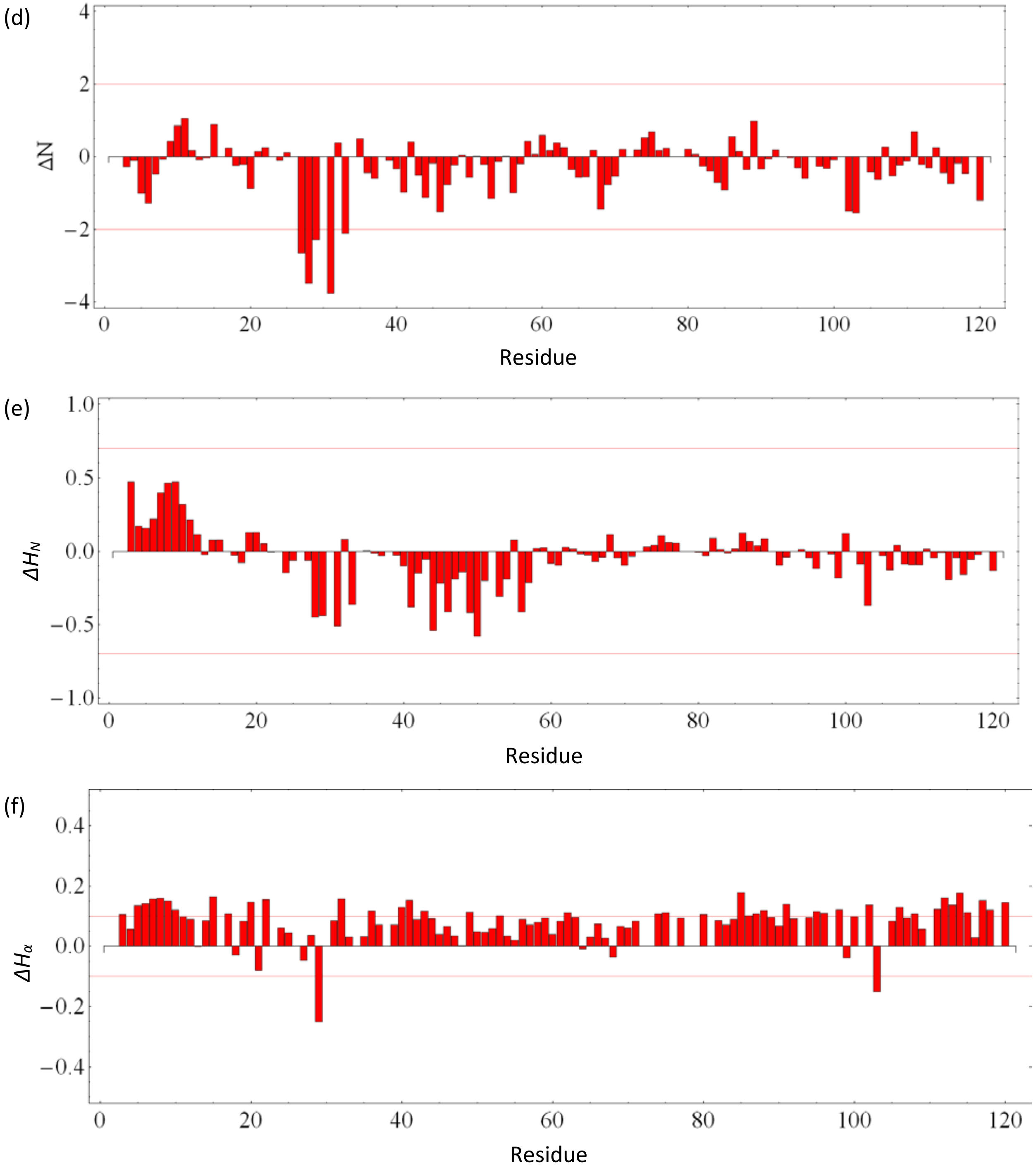


Figure 5.7. Chemical shift difference plots of (a) C_α, (b) C_β, (c) C', (d) ^{15}N, (e) H_N and (f) H_α. The random coil values for chemical shifts were adjusted for sequence dependence[256]. Reference lines are thresholds where differences begin to reflect the possibility of secondary structure formation adapted from references [245, 262].

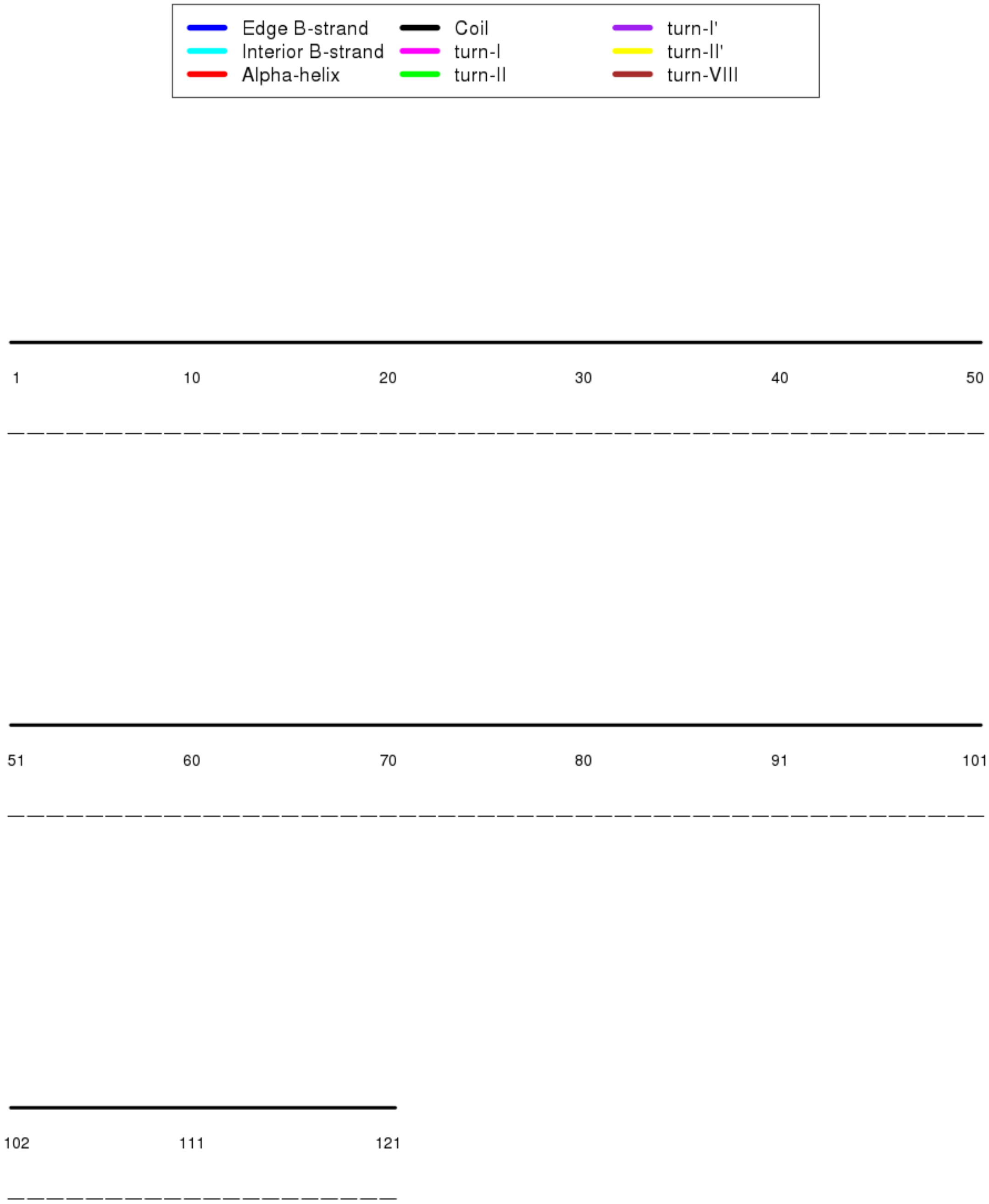


Figure 5.8. Secondary structure prediction result from CSI 3.0 server[263] *using the chemical shifts from the backbone assignments indicates a random coil structure and fully flexible backbone for HIV-1 full-length Tat.*

The secondary structure propensity (SSP) algorithm aims to detect structural propensity in IDPs using NMR backbone chemical shifts[264]. It combines several chemical shifts (C_α, C_β, and H_α) into a single residue-specific score representing the expected fraction of α- or β-structure. Even though the SSP value is not strictly quantitative, it provides valuable information about the secondary structure propensity of IDPs at each residue along the protein sequence. The method has been successfully applied to several IDPs, for example, in studies of α- and γ-synuclein[264], the C-terminal V5 domain of protein kinase C_α[391], and members of the dehydrin protein family[392, 393].

The absolute values of the SSP scores for Tat_{101} are uniformly below 0.5 (Figure 5.9), supporting the previous analyses of the individual chemical shifts showing that the protein is in a disordered conformation. However, Tat_{101} contains two regions with SSP scores with a modest propensity to fold into an α-helix (SSP scores between 0.2 – 0.3): (a) the region around the only tryptophan residue (R27–K32) and (b) the cysteine-rich region (A41–I59). Note that the residue numbers used here are higher than those of the native Tat_{101} sequences because of the addition of the 20-residue purification tag. There are also three regions with a modest propensity to fold into extended or β-sheet conformations (SSP scores between 0.15 – 0.3): (c) the arginine-rich region (K70–G81), (d) the RGD motif region (P93–G99), and (e) the ESKKKVE motif region (E105–E112). In both the Tat-P-TEFb X-ray structure[165] and Fab′-Tat X-ray structure[177], residues 29–32 form a β-turn whereas those regions have a propensity to form an α-helix according to the NMR chemical shift data in Figure 5.9 (region a). This suggests that interactions with a binding partner determine the bound structure in this region favoring an induced-fit mechanism over conformational selection. In contrast, Tat forms a 3_{10} helix between residues 49 and 53 and an α-helix between residues 54 and 62 in the Tat-P-TEFb complex[165], and these regions form the second segment of the protein that has helical propensity in Figure 5.9 (region b). The built-in helical propensity makes conformational selection a favored mechanism for partner-induced folding in this case. In the Tat-TAR-cyclin T1 structure[394], residues 61–67 are extended, and residues 68–79 form the TAR-binding α-helix that is followed by a turn at residues 81 and 82, and an extended segment from residue 83 to 89. Interestingly, the region that forms the TAR-binding helix shows a propensity to form β-structure in Figure 5.9 (region c), which means that the final folded state of the protein is determined by interactions with RNA and cyclin T1. The structures of the two remaining regions (d and e) have not been observed in any crystal structures making it impossible to draw any conclusions about binding mechanism on the basis of the secondary structure propensities.

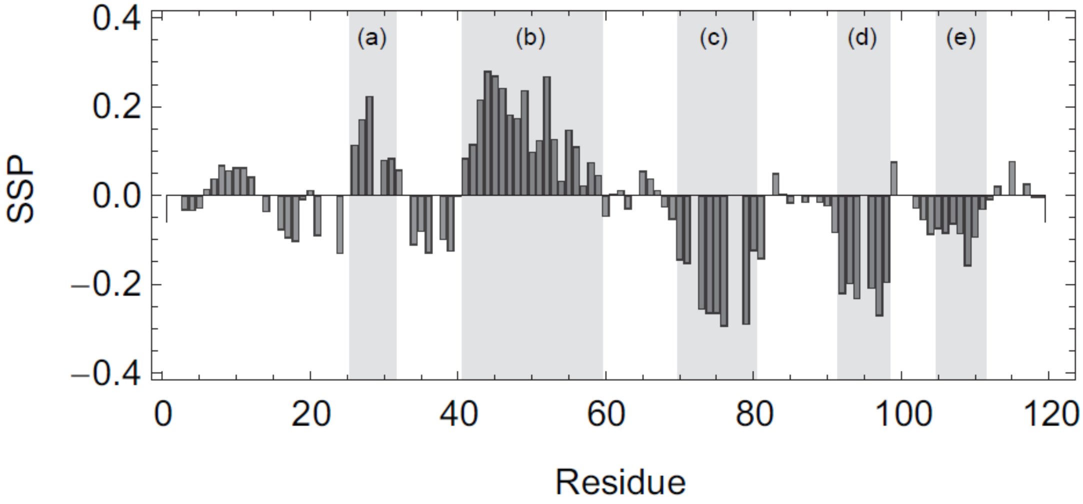


Figure 5.9. Secondary structure propensity of full-length Tat$_{101}$ protein from HIV-1. The five regions with relatively higher absolute SSP scores are (a) the region around the only tryptophan residue (R27–K32), (b) the cysteine-rich region (A41–I59), (c) the arginine-rich region (K70–G81), (d) the RGD motif region (P93–G99), and (e) the ESKKKVE motif region (E105–E112).

In addition to chemical shifts, scalar couplings can be used to obtain information about residual secondary structure. The backbone dihedral angles ϕ and Ψ can be derived from three-bond scalar couplings based on the Karplus equations [265] and, due to the dependence of the angles on the conformation of the polypeptide backbone, secondary structure elements can be inferred. $^3J_{HNH\alpha}$ of α-helices are below 6 Hz and above 8 Hz for β-strands. Most of the $^3J_{HNH\alpha}$ coupling constants (see Figure 5.10) for full-length Tat_{101} fall in the random coil range of 5.9 – 7.7 Hz suggesting that the backbone is disordered in good agreement with the chemical shift differences and SSP values measured above.

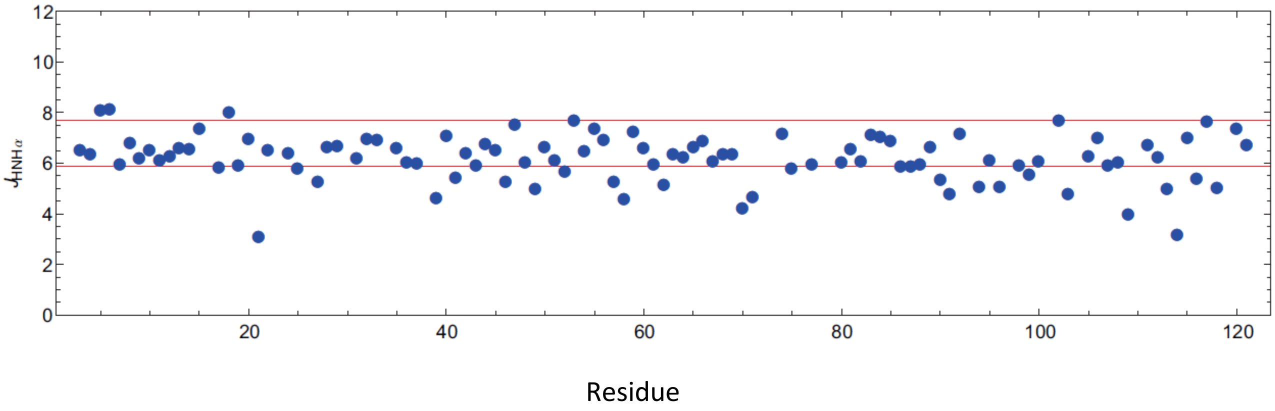
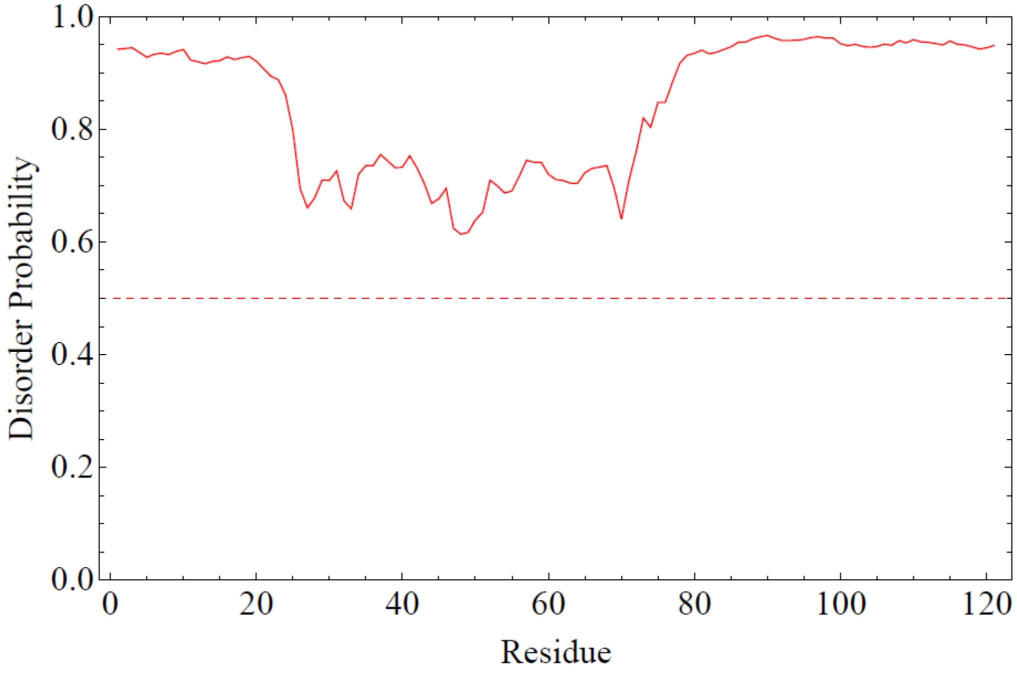


Figure 5.10. Coupling constant $^3J_{HNH\alpha}$ of full-length Tat_{101} protein from HIV-1. Most of the $^3J_{HNH\alpha}$ coupling constants for full-length Tat_{101} fall in the random coil range of 5.9 – 7.7 Hz (the red lines) suggesting that the backbone is disordered.

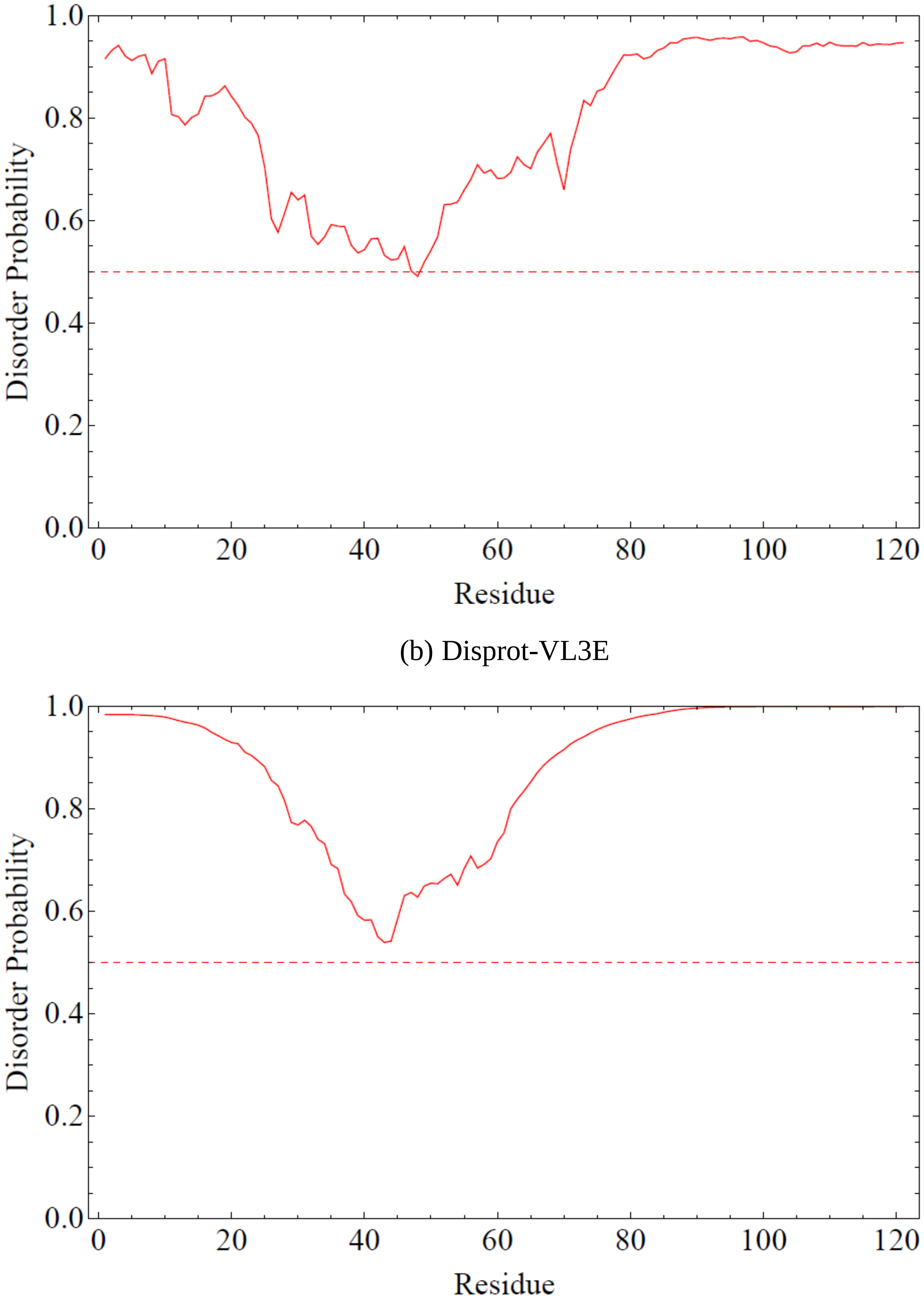
5.4. Predictions of Tat_{101} protein disorder.

DisProt[395-398], IUPreD[40, 399], and PrDOS[400] disorder prediction programs were used to predict disordered regions in full-length Tat_{101} protein. The disorder prediction programs in the DisProt family use machine – learning algorithms trained on the datasets of disordered regions of different length characterized by various methods. DisProt VL3-H and DisProt VL3-E programs are trained on and thus, give accurate prediction results only on long regions of disorder (longer than 30 residues)[395]. The DisProt VSL2P program is trained on a dataset of combined long and short disordered regions and therefore it achieves better-balanced prediction accuracies on both types of regions[397]. The algorithm used in the IUPreD program[40] was discussed in Section 1.1.2. The PrDOS prediction program[400] consists of two predictors, one based on the local amino acid sequence information and the other based on the template proteins. First, the input amino acid sequence is converted into a position-specific score matrix (PSSM). Then, two predictions are performed using the PSSM. The first predictor uses a machine – learning algorithm which is trained on a large dataset from the PDB. The second predictor is based on template proteins and uses the alignments of a query sequence with structures that are known. To combine the results of two independent predictors, the weighted average between the results of two predictors is calculated.

Prediction scores above the dashed lines in Figure 5.11 and Figure 5.12 indicate a high probability of disorder, while scores below the lines indicate order in the sequence. All programs predicted that the entire full-length Tat_{101} protein, except for the cysteine-rich region (residue 42-57), is disordered at high confidence. Some of the algorithms predicted that the cysteine-rich region has some degree of order due to the fact that the prediction methods expected the cysteines to be involved in forming intramolecular disulfide bonds or in coordinating with a metal ion (zinc finger). For the Pro-rich region (residue 20 – 41) and the basic region (residue 68-77), the lower disorder scores in all disorder prediction results suggest the possibility of formation of transient structures in these regions, in agreement with the SSP result measured by NMR (see Section 5.3 and Figure 5.9). In most of the prediction results (except for the DisProt-VSL2P), the small segment consisting of residues 105 – 110 has modestly lower disorder scores compared to the rest of the second exon product of Tat, which is also in agreement with the SSP result, suggesting a structural propensity, albeit at low probability, in this segment



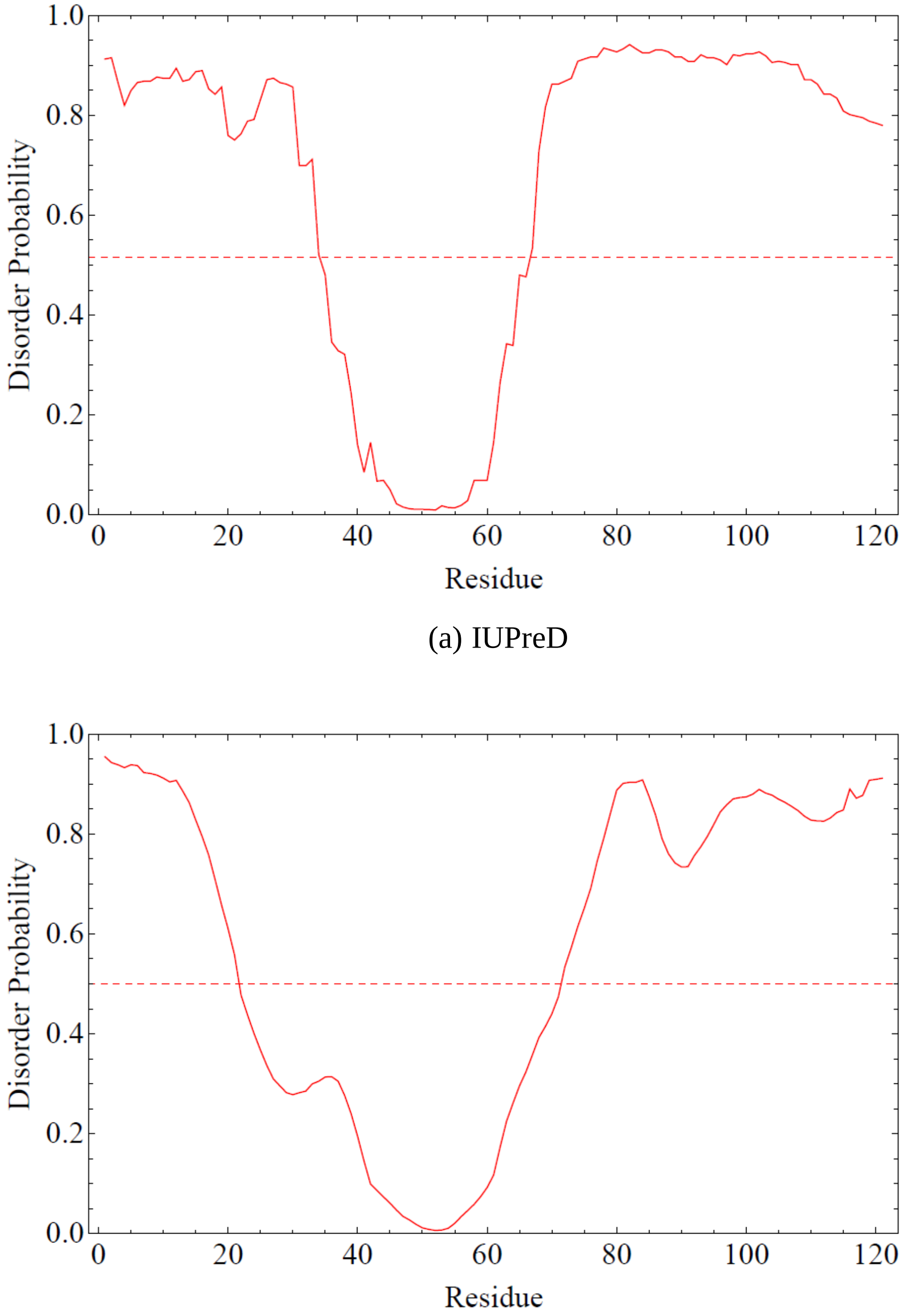
(a) Disprot-VL3H



(b) Disprot-VL3E

(c) Disprot-VSL2P

Figure 5.11. DisProt disorder predictions for the His-tagged full-length Tat$_{101}$ protein sequence using the algorithms: (a) VL3H[395, 396]*, (b) VL3E*[396]*, and (c) VSL2P*[397, 398]*.*



(a) IUPreD

(b) PrDOS

Figure 5.12. (a) IUPred[40, 399] and (b) PrDOS[400] disorder predictions for the His-tagged full-length Tat$_{101}$ protein sequence.

5.5. NMR relaxation.

5.5.1. ps – ns timescale.

5.5.1.1. Relaxation parameters R_1, R_2, and NOE

Figures 5.13 and 5.14 show sample spectra for the R_1 and R_2 relaxation measurements of the His-tagged Tat_{101} protein at pH 4.0 and 293 K acquired on the Varian INOVA 600 MHz spectrometer at University of Manitoba. As described in section 2.2.1, relaxation rates (R_1 and R_2) were extracted from the NMR relaxation experiments by fitting the experimental data to exponential decay models. Figure 5.15 shows sample fits for residue Val-111.

Figures 5.16 and 5.17 show spectra of the heteronuclear ^1H-^{15}N NOE measurements acquired on the Varian INOVA 800 MHz spectrometer at NANUC. The steady-state heteronuclear ^1H-^{15}N NOE values were obtained from the ratios of the peak heights from experiments with (I_{NOE}) and without (I_{noNOE}) saturation of the protons for 5 s at the beginning of the experiment. The heteronuclear NOE values were then obtained from (I_{NOE} - I_{noNOE})/I_{noNOE}.

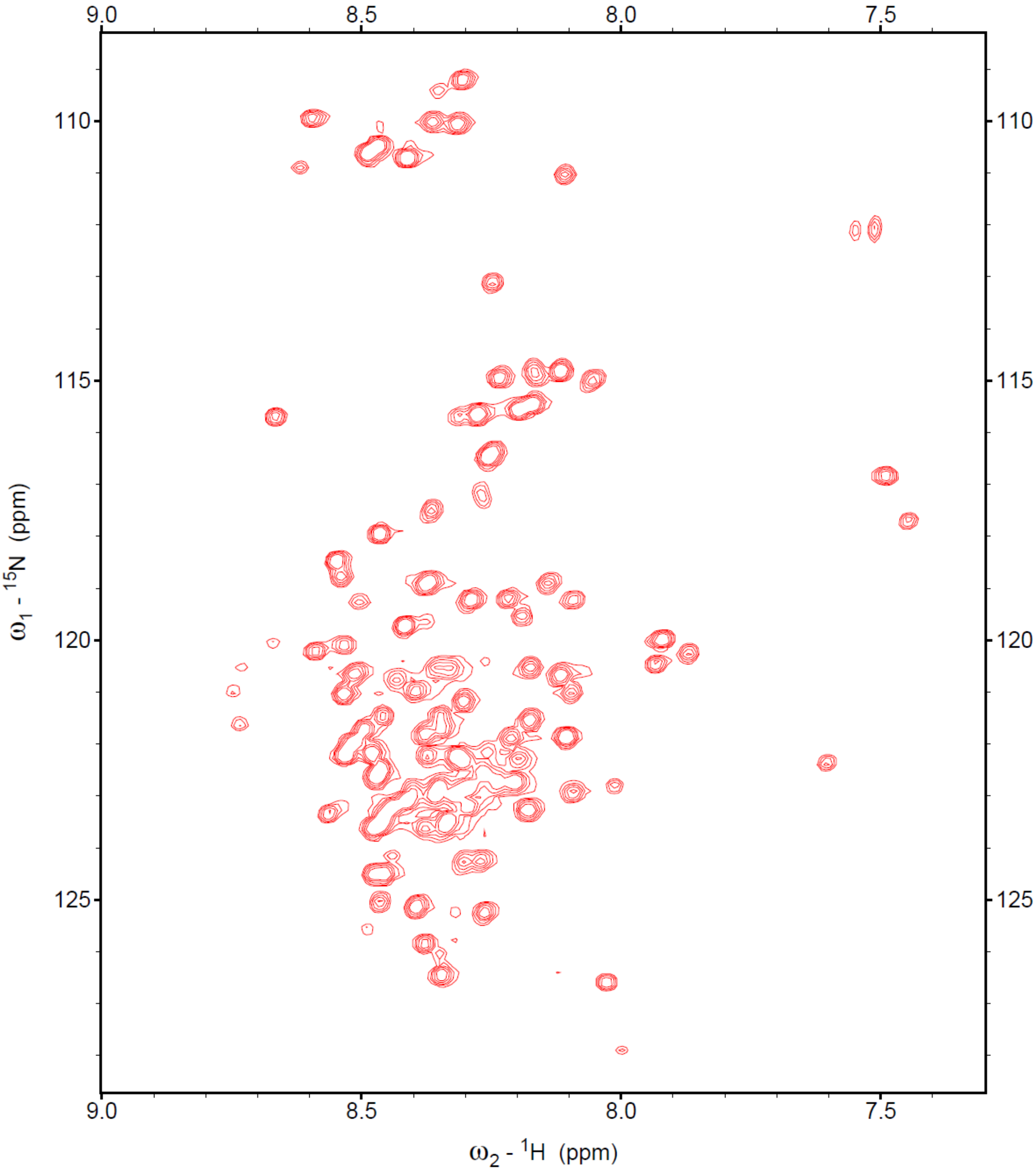


Figure 5.13. Sample spectrum (at 600 ms relaxation time) for R_1 relaxation measurement of His-tagged Tat_{101} at pH 4.0 and 293 K. The spectrum was recorded on a Varian INOVA 600 MHz spectrometer.

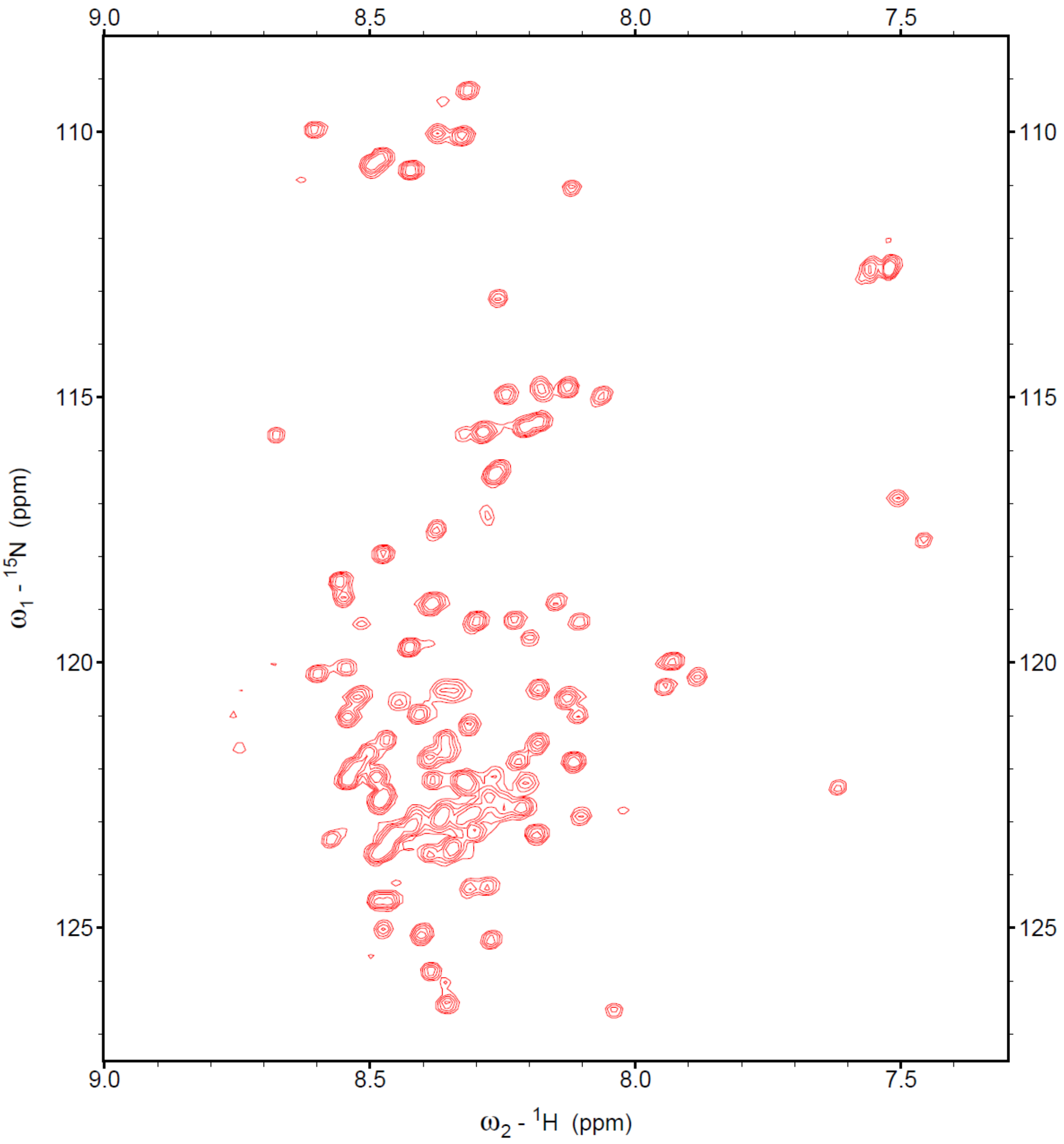
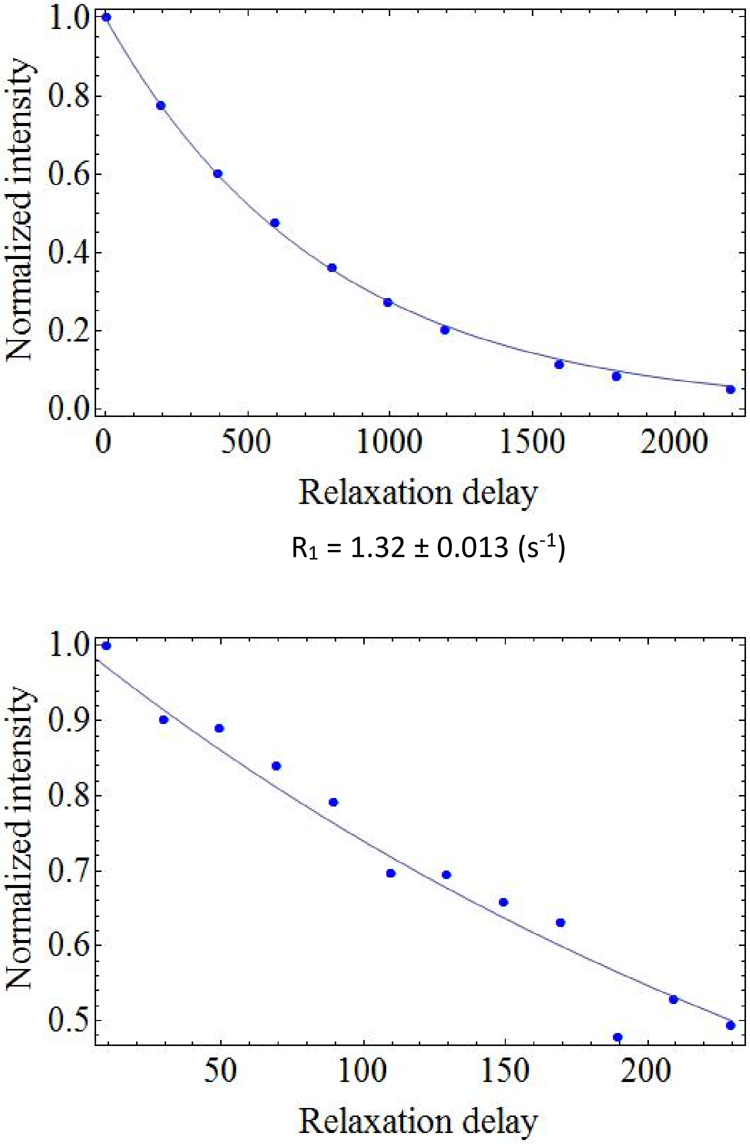


Figure 5.14. Sample spectrum (at 50 ms relaxation time) for R_2 relaxation measurement of His-tagged Tat_{101} at pH 4.0 and 293 K. The spectrum was recorded on a Varian INOVA 600 MHz spectrometer.



R_1 = 1.32 ± 0.013 (s^{-1})

R_2 = 2.80 ± 0.107 (s^{-1})

Figure 5.15. Sample fits for extraction of longitudinal (R_1, top panel) and transverse (R_2, bottom panel) relaxation rates of residue Val111, measured on the INOVA 600 MHz NMR spectrometer.

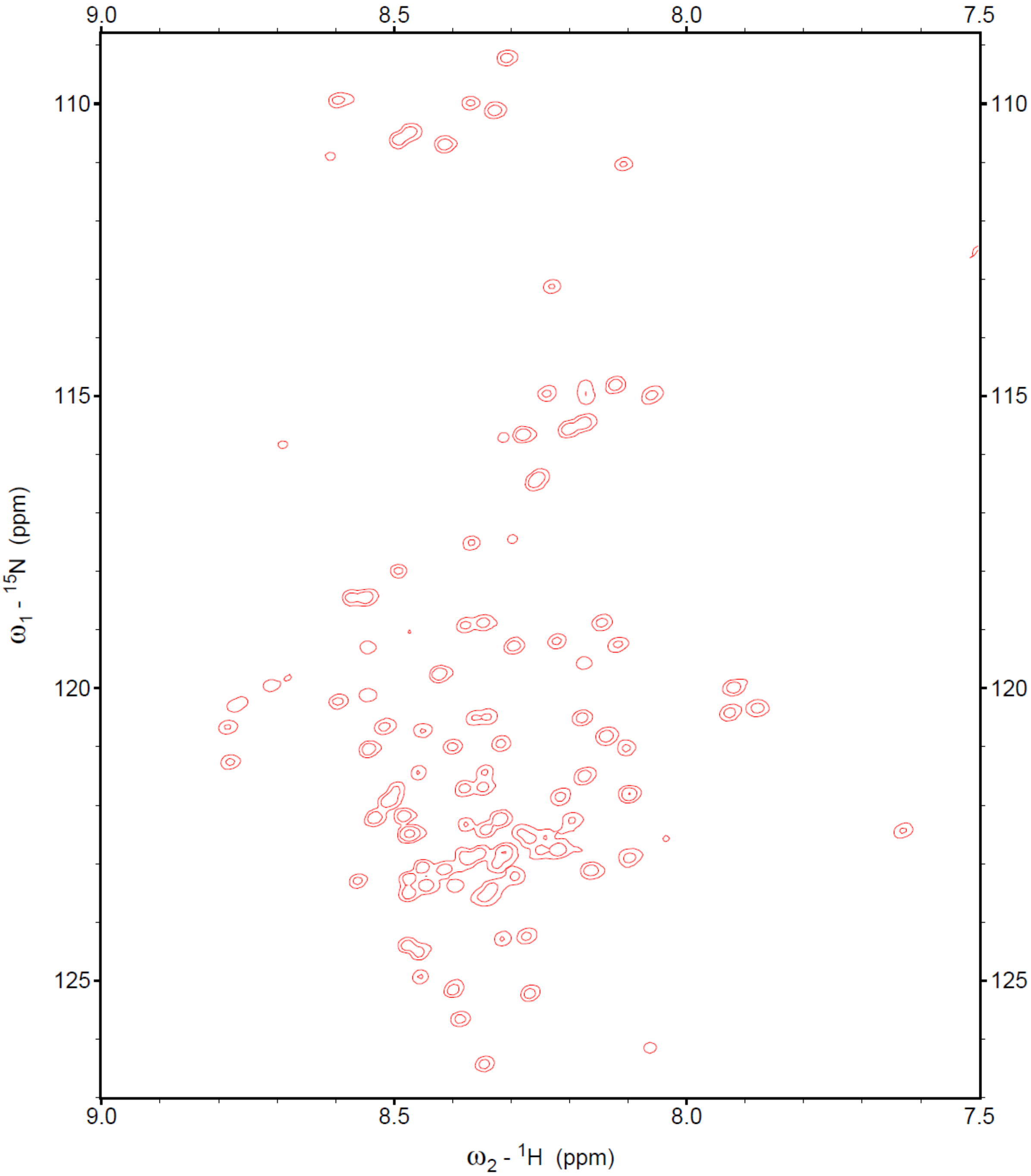


Figure 5.16. Steady-state heteronuclear ^{1}H-^{15}N NOE measurement spectrum with no saturation period of His-tagged Tat_{101} at pH 4.0 and 293 K. The spectrum was recorded on a Varian INOVA 800 MHz spectrometer.

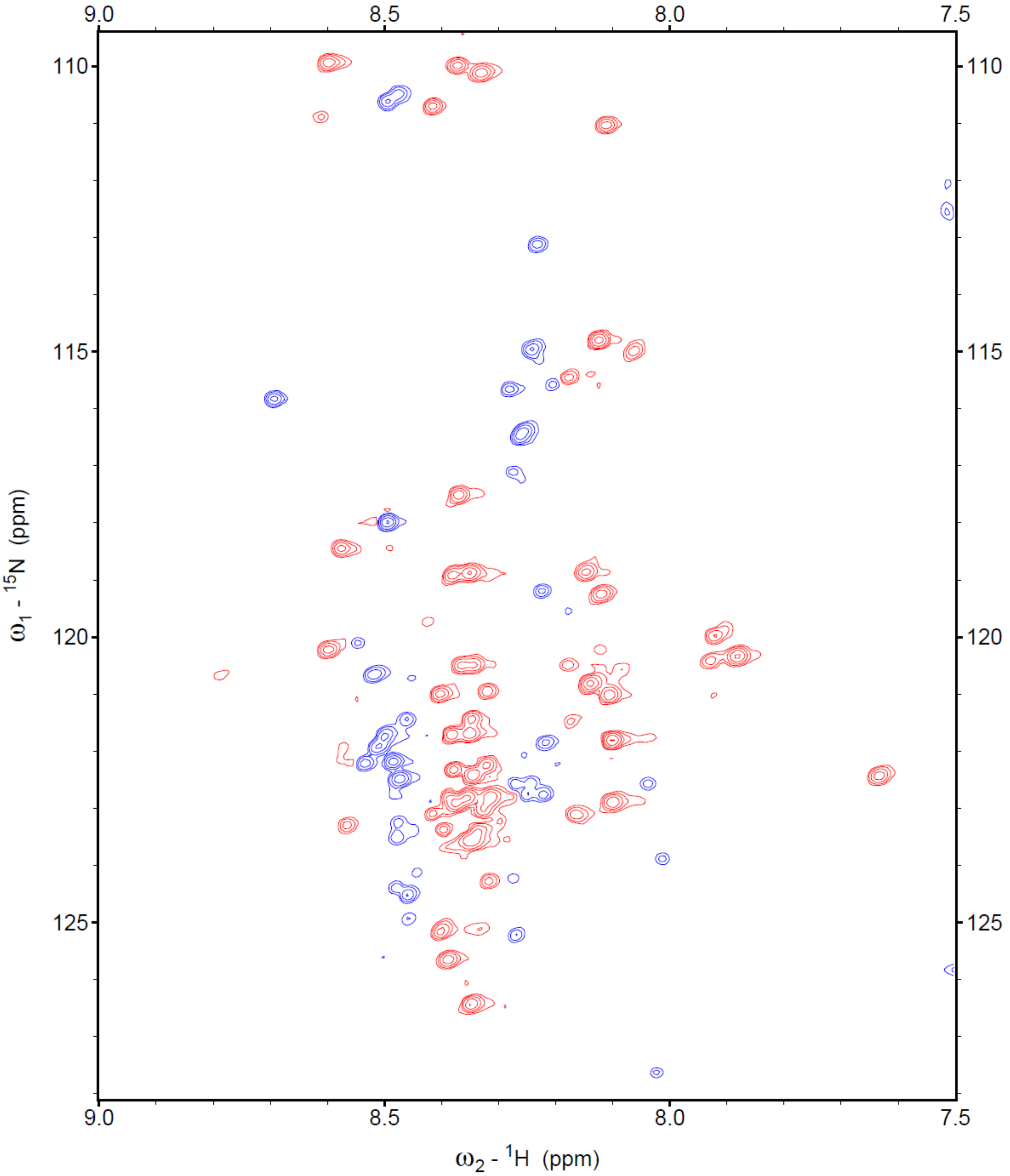


Figure 5.17. Steady-state heteronuclear ^{1}H-^{15}N NOE measurement spectrum with a 5 second saturation period of His-tagged Tat$_{101}$ at pH 4.0 and 293 K. Positive and negative peaks are in red and blue, respectively. Spectrum was recorded on a Varian INOVA 800 MHz spectrometer.

Relaxation data (R_1, R_2, and NOE) at two fields for non-proline residues are plotted against the protein sequence in Figure 5.18, showing relatively uniform relaxation rates and steady-state heteronuclear NOEs across the sequence. The mean values and standard deviations for R_1, R_2, and NOE at both fields are listed in Table 3.1 where they are also compared to the measurements on Tat_{72}[102]. The negative values of the NOEs mean there is very little restriction of dynamics on the nanosecond to picosecond timescale, supporting the conclusion that, in general, the protein is intrinsically disordered throughout the sequence. Residues at the N-terminus have the most negative NOEs and the lowest relaxation rate values, in agreement with the expectation of less restricted dynamics at protein termini[102]. The relaxation rates R_1 and R_2 are fairly low, consistent with expectations for a disordered protein[102, 401]. The mean values of the relaxation data of the full-length protein (Tat_{101}) are close in comparison with the first exon (Tat_{72}) which is also disordered[160]. This is also consistent with literature reports, for example, a recent NMR relaxation study[401] on the Merozoite Surface Protein 2 (MSP2), a disordered protein found in early stages of a protozoan parasite that causes malaria in humans shows R_1 relaxation rates distributed in the range of 1.0 - 1.5 s^{-1}, whereas the R_2 relaxation rates distribute in the range of 1.5 – 3.0 s^{-1}.

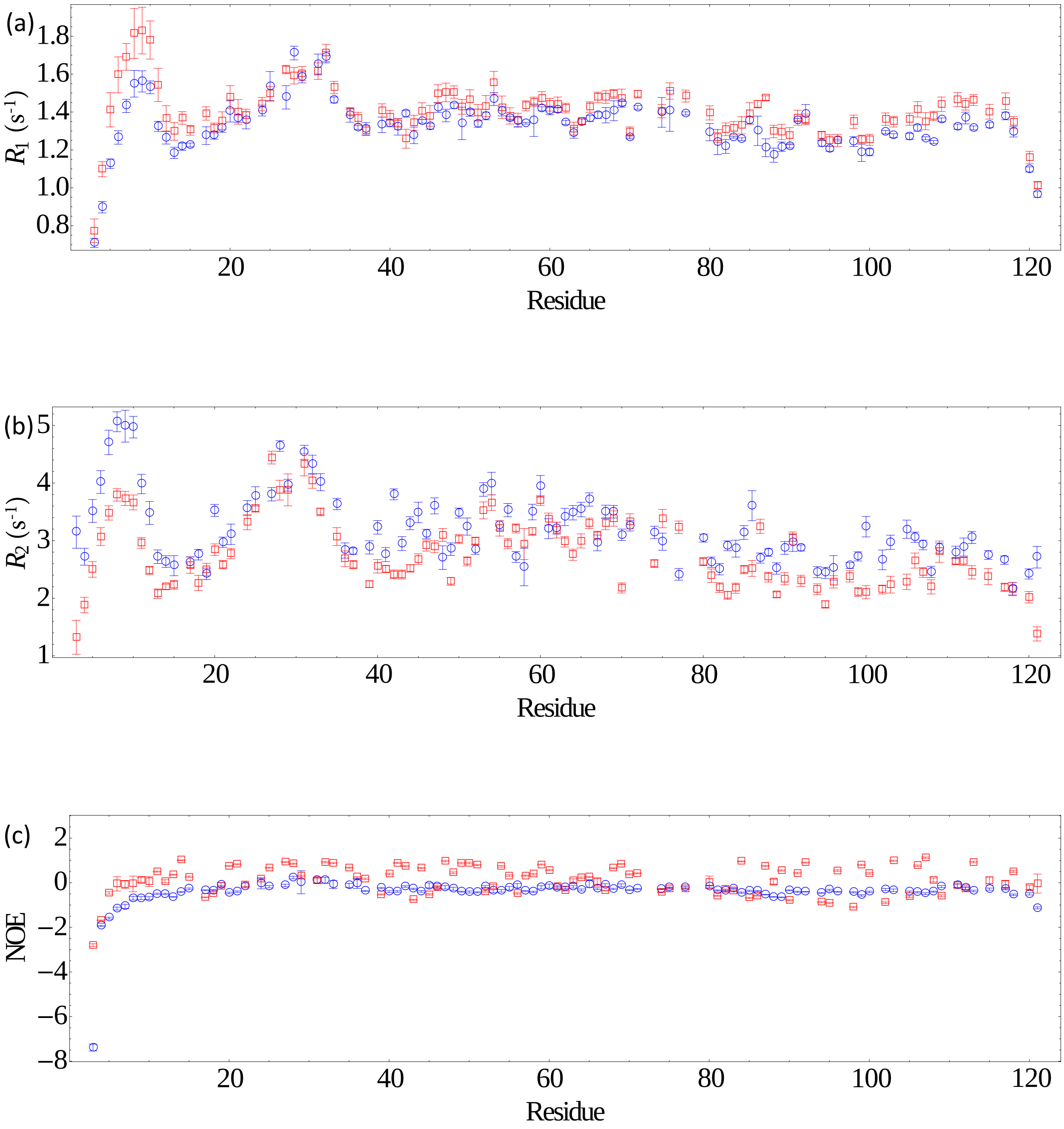


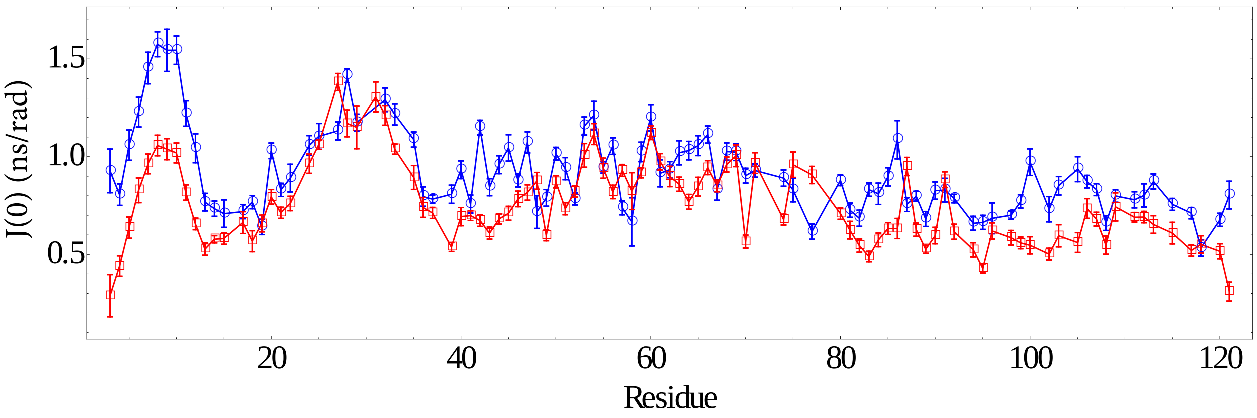
Figure 5.18. Relaxation measurements of full-length Tat_{101} protein from HIV-1 at pH 4.0 and 298 K: (a) longitudinal relaxation rates, R_1; (b) transverse relaxation rates, R_2; and (c) heteronuclear NOE values. Results are determined using relaxation data measured on 600 MHz (blue ○) and 800 MHz (red □) spectrometers.

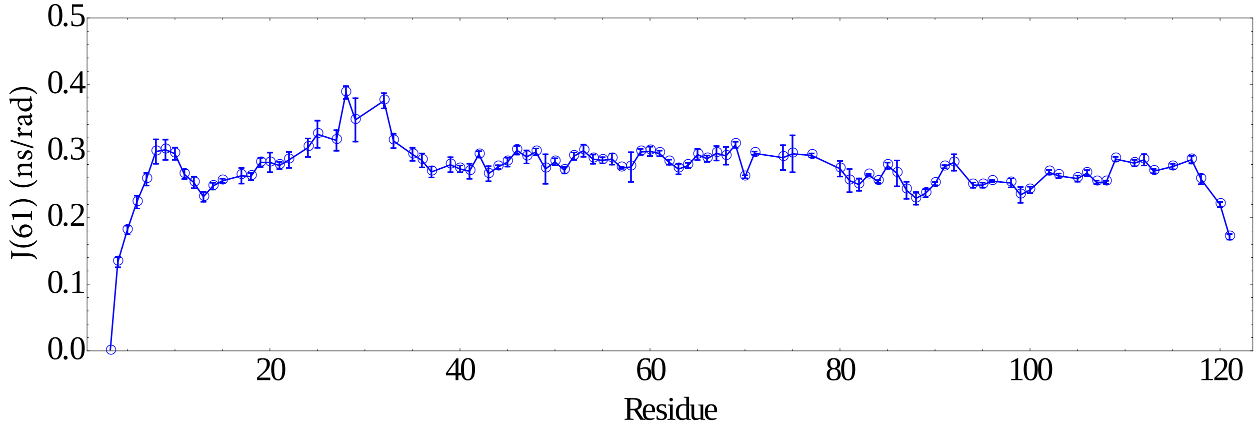
Table 5.1. Means and Standard Deviations of Relaxation Data of full-length protein, Tat$_{101}$, and the first exon product, Tat$_{72}$.

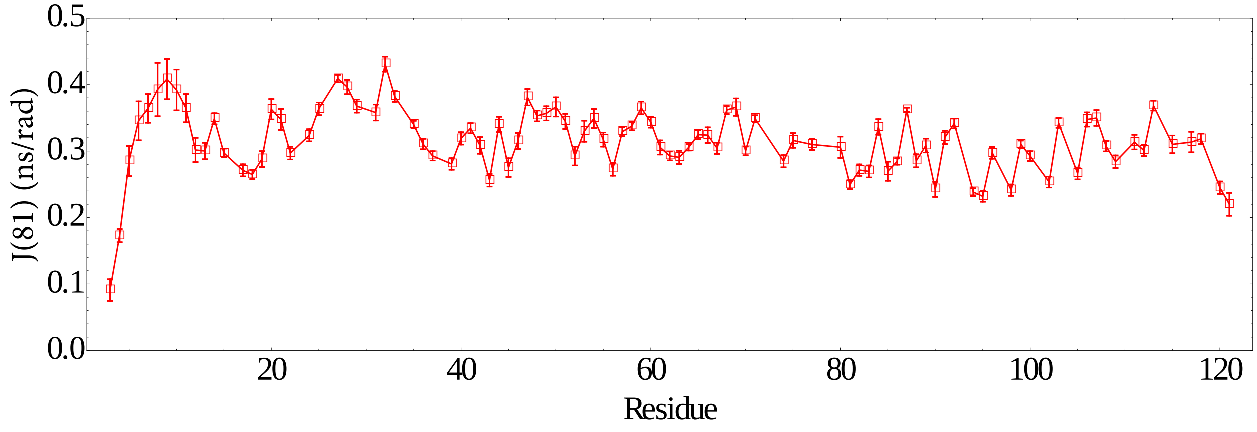
	Tat$_{101}$		Tat$_{72}$	
	600 MHz	800 MHz	600 MHz	800 MHz
R$_1$ (s^{-1})	1.33 ± 0.139	1.41 ± 0.145	1.5 ± 0.2	1.4 ± 0.2
R$_2$ (s^{-1})	3.20 ± 0.616	2.76 ± 0.602	3.8 ± 1.3	N/A
NOE	-0.432 ± 0.764	0.121 ± 0.654	-0.27 ± 0.46	-0.07 ± 0.33

5.5.1.2. Reduced spectral density mapping.

Reduced spectral density mapping was used to analyze the relaxation data. This method allows the spectral density of motions $J(\omega)$ to be sampled at different frequencies, zero, ω_N, and $0.87\omega_H$, corresponding to relaxation contributions from the motion of ^{15}N–^{1}H bond vectors on slow, intermediate, and rapid timescales, respectively[273]. Values of the spectral density at five frequencies, 0, 61, 81, 522, and 696 MHz, are presented in Figure 5.19. In general, the data show fairly uniform contributions to relaxation throughout the protein at all frequencies with a few minor exceptions.


J(0) (ns/rad)
1.5
1.0
0.5
20
40
60
80
100
120
Residue


J(61) (ns/rad)
0.5
0.4
0.3
0.2
0.1
0.0
20
40
60
80
100
120
Residue


J(81) (ns/rad)
0.5
0.4
0.3
0.2
0.1
0.0
20
40
60
80
100
120
Residue

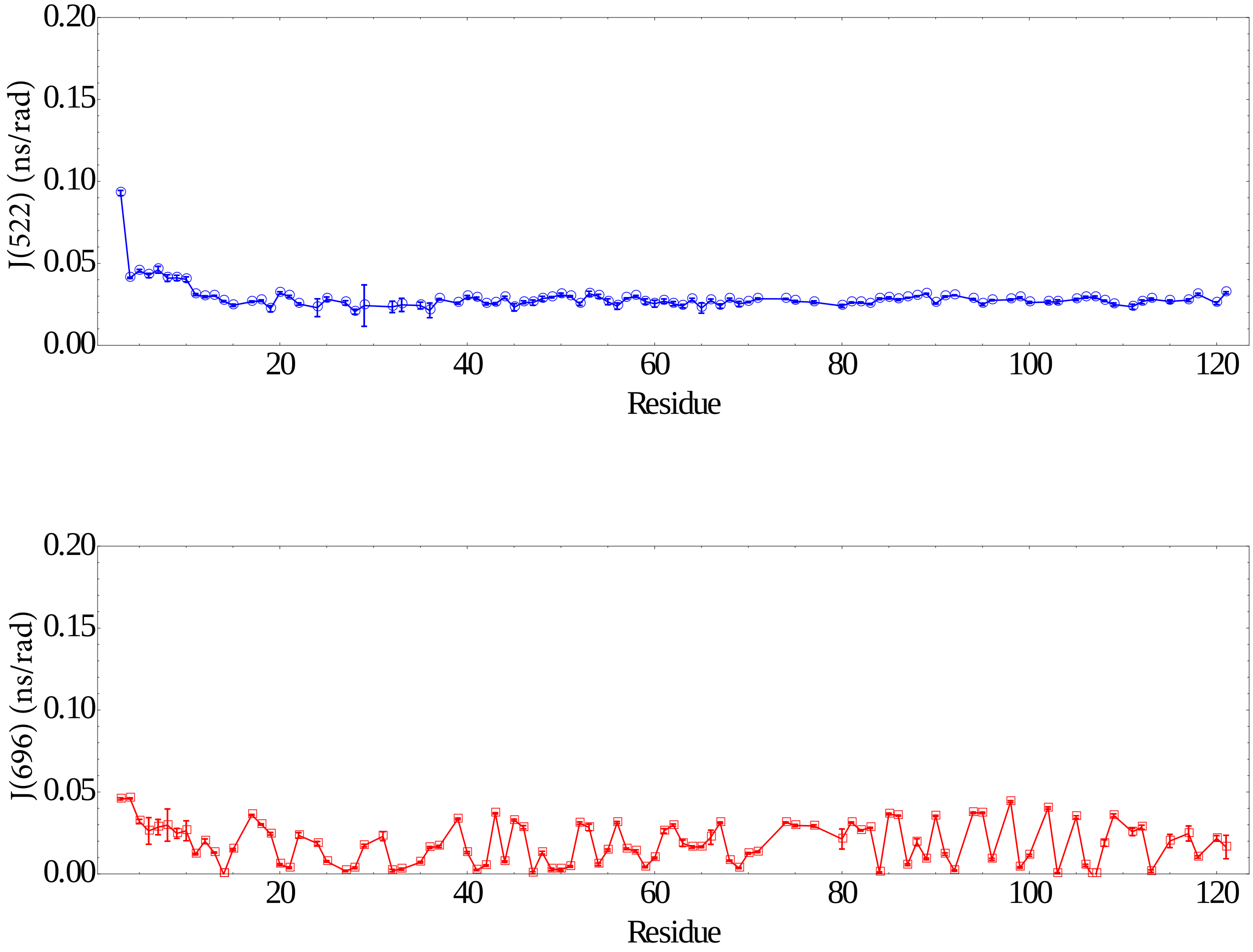


Figure 5.19. Reduced spectral density mapping at five different frequencies (0, 61, 81, 522, and 696 MHz) of full-length Tat$_{101}$ protein from HIV-1 at pH 4.0 and 298 K. Results are determined using relaxation data measured on 600 MHz (blue ○) and 800 MHz (red □) spectrometers.

In general, when analyzing spectral density mapping data, if the amide bond vectors are highly mobile, the internal motions will greatly contribute to the protein relaxation and, therefore, increase the value of the spectral density at high frequencies ($J(0.87\omega_H)$, *i.e.* $J(522)$ and $J(696)$), and decrease its magnitude at low ($J(0)$) and mid-frequencies ($J(\omega_N)$, *i.e.* $J(61)$ and $J(81)$) [402]. As expected for a disordered protein, the full-length His-tagged Tat_{101} shows relatively high contributions at high frequencies and smaller contributions at low frequencies. The tag region has higher values of $J(0.87\omega_H)$, and significantly lower values of $J(\omega_N)$ compared to the rest of the protein, strongly suggesting the high flexibility of this region. The protein C-terminus also has lower values of $J(\omega_N)$, but the $J(0.87\omega_H)$ values are similar to the other parts of the protein. This suggests that the C-terninus is flexible but to a lower extent compared to the N-terminus (the His-tag region).

On the other hand, if the amide bond vectors are motionally restricted, the limited internal motion has little effect on relaxation, and therefore the high-frequency contribution to the spectral density will be minor. In this case, relaxation mostly comes from conformational exchange, and the greatest contributions to the spectral density function will come from the low-frequency components. This is observed at the region near the W31 residue where $J(0)$ and $J(\omega_N)$ values are slightly higher , while $J(0.87\omega_H)$ values are similar compared to those of other parts of the protein. This observation suggests that the protein backbone undergoes significantly more local backbone fluctuations on a timescale between nanoseconds and milliseconds compared to other regions of the protein. This may indicate the formation of transient structures in this region on the microsecond timescale. Folding nuclei around hydrophobic aromatic residues have been observed previously[403].

In general, the field dependence of $J(0)$ can arise from conformational exchange contributions to R_2 and micro- to millisecond timescale motions[286], with the expectation of an increase in $J(0)$ when going from 600 to 800 MHz when exchange is present and/or there are slow motions. The observation of a minimal field dependence of $J(0)$ indicates that most of the residues are not undergoing conformational exchange on the millisecond timescale.

5.5.1.3. Model-free analysis.

Model-Free analysis is another method for analyzing relaxation data. The Lipari–Szabo Model-Free approach[221, 222] and its extension, the Extended Model-Free approach by Clore *et al.*[275], quantify the local backbone motions through the following parameters: the residue order parameter (S^2), the molecular rotational correlation time (τ_m), the effective residue internal rotational correlation time (τ_e), and the residue rate of slow conformational exchange (R_{ex}). In fitting the measured relaxation rates to the Lipari–Szabo spectral density

$$J(\omega) = \frac{2}{5}\left[\frac{S^2\tau_m}{1+\tau_m^2\omega^2} + \frac{(1-S^2)\tau}{1+\tau^2\omega^2}\right]$$

where $1/\tau = 1/\tau_m + 1/\tau_e$, the method of Schurr *et al.* was used[404], whereby S^2, τ_m, and τ_e are optimized for each residue individually. The residue - specific rotational correlation times (τ_m) help to avoid the use of a single global correlation time, or equivalently a single diffusion tensor, which is insufficient to describe the rotational motion of a disordered protein. The order parameter is a measure of the degree of spatial restriction of backbone amide motion, ranging from 0 (unrestricted motion) to 1 (rigid). The rate of exchange may be included or excluded in the analysis to verify the effect of exchange on relaxation, depending on model selection criteria[276, 277]. Model-Free analysis is based on the assumption that local backbone motions are independent of overall molecular rotation. Even though this assumption may not hold when applied to highly disordered proteins[405], IDPs exhibit parameters that are instructive.

The mean value for the order parameters of Tat_{101} (Figure 5.20) is 0.75 with a relatively high standard deviation of 0.24, suggesting the protein is very flexible. Interestingly, the average order parameter for Tat_{101} is significantly higher than that measured for the Tat_{72} protein[160] (0.55). The highest order parameters are observed for the His-tag region, suggesting that charge repulsion forces the backbone into a rigid extended conformation there. The R_{ex} values are low, showing that the protein has little slow motion on the micro- to millisecond timescale, in agreement with the relaxation dispersion results presented below. In general, the internal rotational correlation times are 10–100-fold faster than the overall rotational correlation times. This is in contrast to the case for the Tat_{72} protein, for which most of the internal correlation times were ∼10-fold faster than the overall molecular reorientation[160]. Thus, the addition of 29 residues appears to move the internal and overall dynamics onto timescales that permit better separation of the two regimes. However, it must be pointed out that the errors in the fits to the full-length protein are significantly greater than those measured for the first exon product[102].

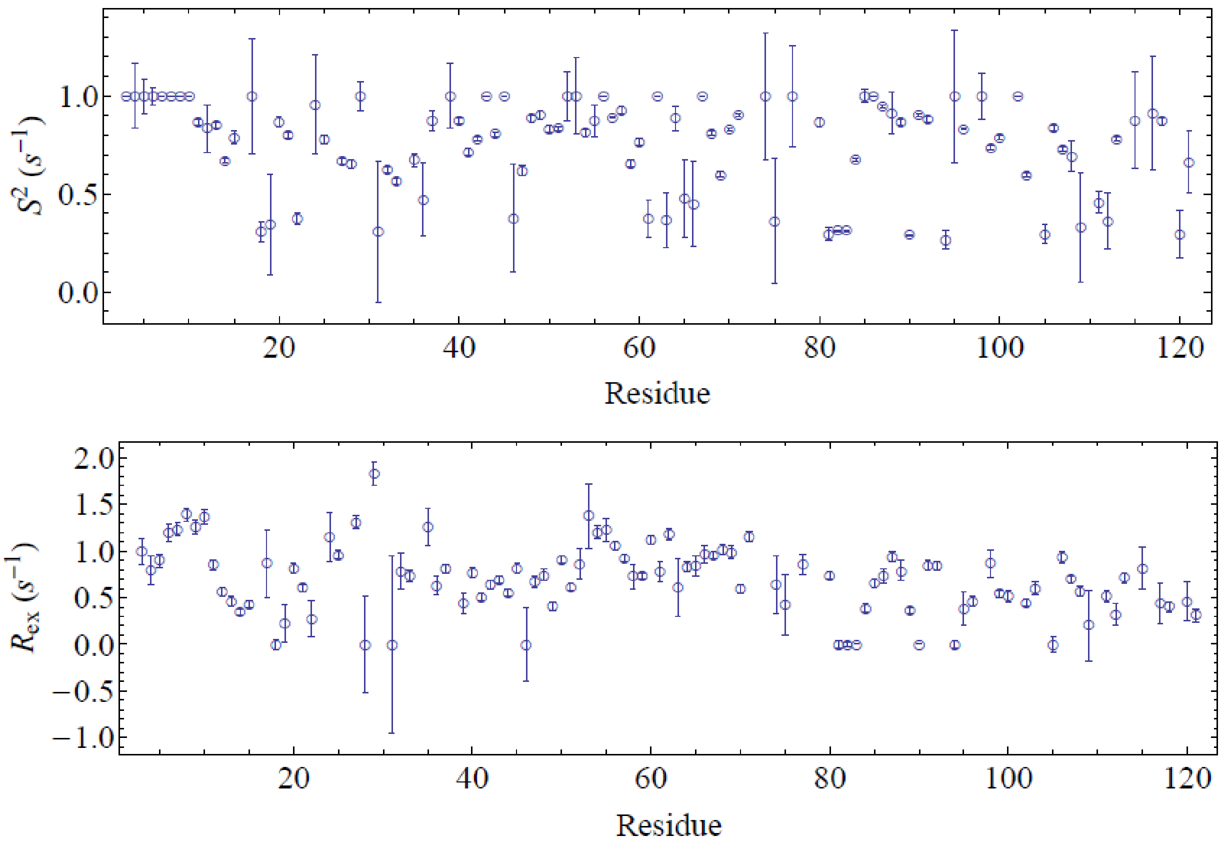


Figure 5.20. Order parameters and exchange rates estimated using model-free analysis on relaxation data at 2 fields 14.1 T and 18.8 T.

5.5.2. μs – ms timescale.

Relaxation dispersion is a powerful method for characterizing slow motions and conformational exchange on the millisecond timescale. Detection of this type of motion has been very helpful in defining the mechanisms by which IDPs interact with their partners[191, 406] IDPs that undergo slow conformational exchange are sampling different conformations in solution, allowing for the possibility of folding by a conformation selection mechanism. However, when fully reduced and at pH 4.0, Tat_{101} shows flat dispersion profiles for all residues measured at both 14.1 and 18.8 T (Figure 5.21), suggesting that the backbone of Tat_{101} has no slow motion under these conditions. Relaxation dispersion experiments were also conducted on Tat$_{101}$ at pH 7.0 for the observable peaks even though these peaks were mostly not assigned (data not shown). The profiles for all peaks are flat, indicating no slow motion at these sites. This result does not rule out the possibility that slow motion may exist in other protein sites that are unobservable because of lost intensity at pH 7. The origin of the peak loss, whether due to slow exchange broadening or fast hydrogen exchange, was probed by hydrogen exchange measurements (see Section 5.6 below).

Since these experiments on HIV-1 Tat$_{101}$ were the first measurements of relaxation dispersion on our spectrometer we made measurements on another protein suspected to have regions undergoing conformational exchange to ensure that our experimental protocols and the spectrometer were working properly. Figure 5.22 shows several relaxation dispersion profiles for the 18 kDa Ovarian Tumour (OTU) Domain Protease provided by S. Saran. Several resonances in this enzyme exhibited broadened resonances that also showed slow conformational exchange by relaxation dispersion, indicating that our spectrometer and pulse programs were working properly.

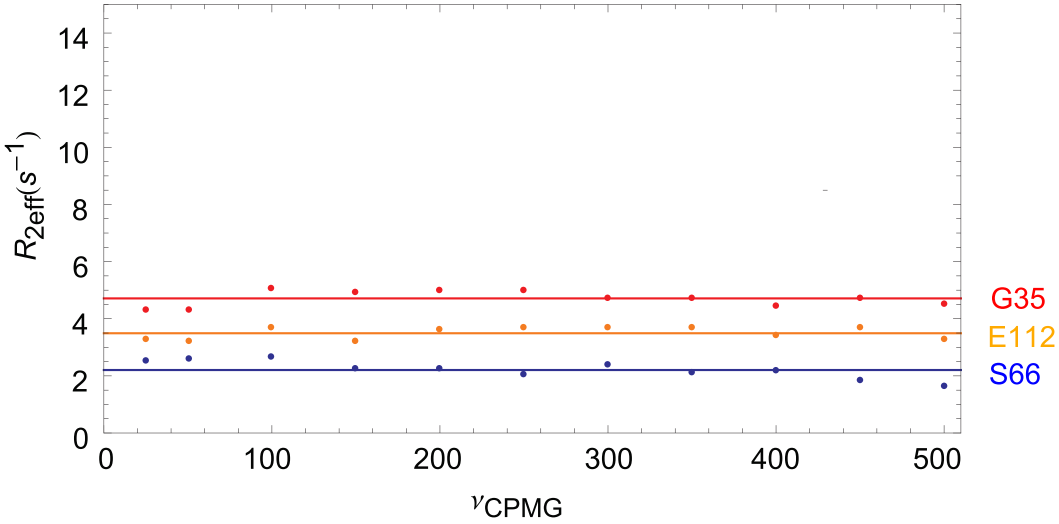
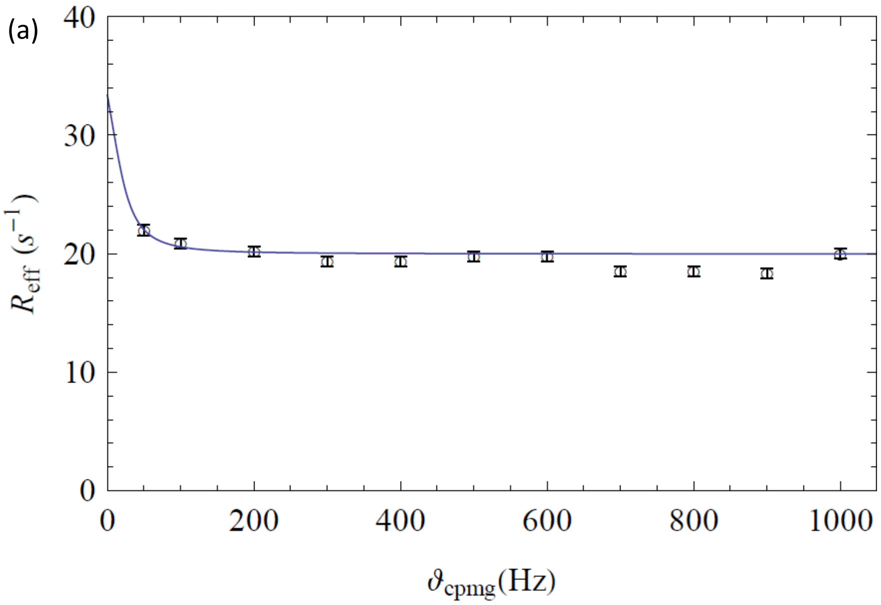


Figure 5.21. Relaxation dispersion profiles for residue Gly35 (red), Ser66 (blue) and Glu112 (orange) of full-length Tat$_{101}$ protein. The experiments were done at 14.1 T and 18.8 T. Data were fit using the NESSY program[380].



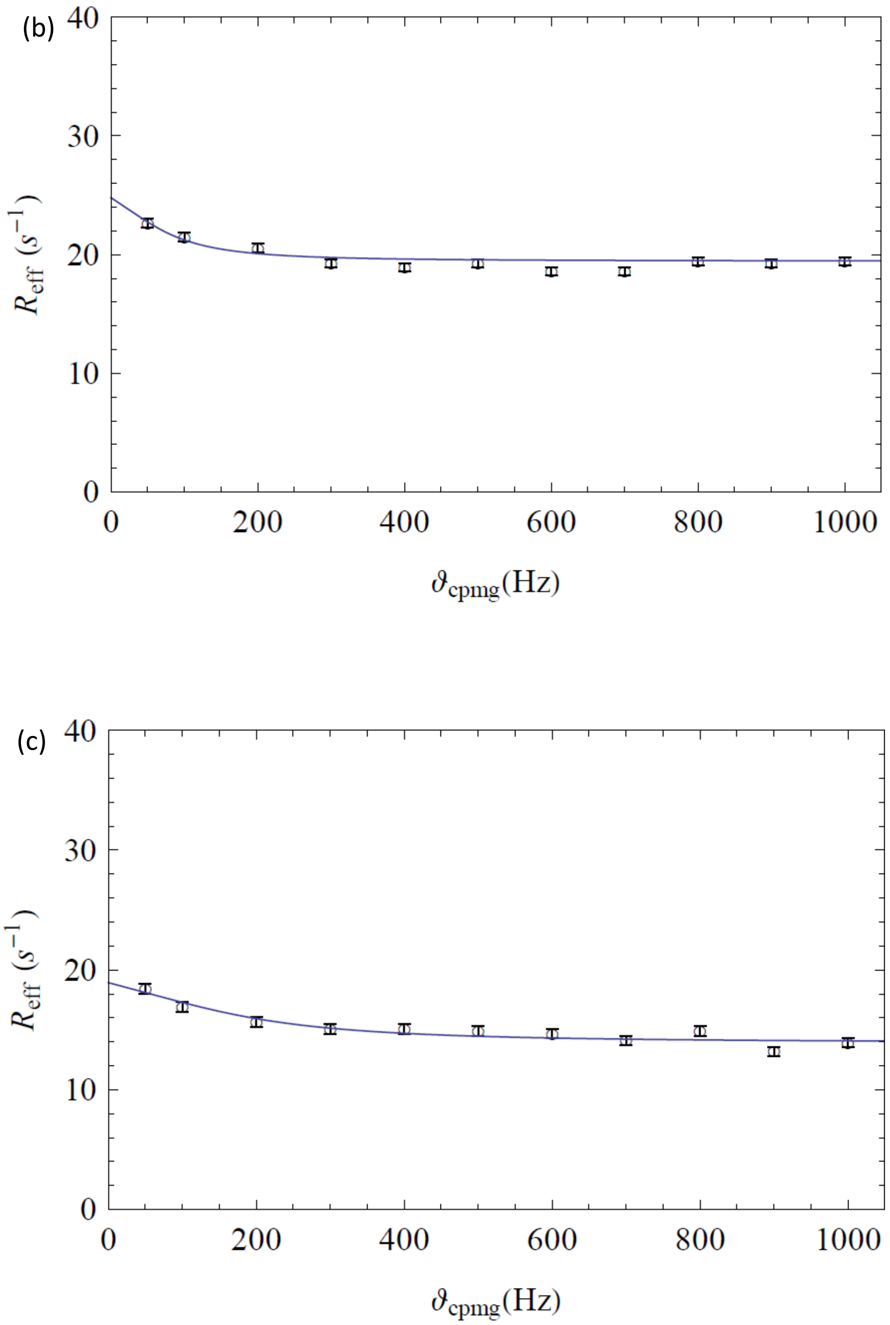


Figure 5.22. Relaxation dispersion profiles for residue (a) Ile13, (b) Ser95, (c) Thr54 of OTU protease. The experiment was done on the INOVA 600 MHz NMR spectrometer at University of Manitoba. Data were fit using the NESSY program[385].

5.6. Hydrogen exchange.

Hydrogen exchange rates can be a powerful high-resolution probe of protein folding and dynamics on the millisecond timescale and longer[407]. Exchange rates and protection factors for full-length Tat_{101} were extracted from CLEANEX-PM[329, 331] and SOLEXSY[330] experiments that probe hydrogen exchange on different timescales. Figure 5.23 shows sample spectra at one mixing time of the N^H and N^D cross-peaks in a SOLEXSY experiment of Tat_{101} dissolved in 50% D_2O. In the figure, the N^H peaks represent the amides that were initially protonated and remained protonated after the exchange period. The N^D peaks, represents the amides that were initially deuterated, and became protonated after the exchange period. Figure 5.24 shows a sample spectrum at 75 ms mixing time for the hydrogen exchange measurement of Tat_{101} using the CLEANEX-PM pulse program.

Even though CLEANEX-PM is the benchmark experiment for hydrogen exchange measurements in folded proteins, results may be confounded by errors introduced by water manipulation and, in the case of IDPs, from unwanted NOE-mediated magnetization transfer mechanisms[330]. To verify the CLEANEX-PM hydrogen exchange results, we measured hydrogen–deuterium exchange rates using the SOLEXSY experiment. SOLEXSY is considered accurate if the exchange rates are faster than approximately 0.2–0.5 s^{-1} and slower than approximately 2 s^{-1} [330]. Figure 5.25 compares hydrogen exchange rates in Panel (a) and protection factors in Panel (b) measured by CLEANEX-PM (empty bars) and SOLEXSY (filled bars). In general the SOLEXSY results are in good agreement with the CLEANEX-PM measurements throughout the protein. Only the residues in the Hig tag (residue 1–21) fall in the valid SOLEXSY measuring range, and the SOLEXSY-measured exchange rates there agree very well with the CLEANEX-PM measured rates. Even in the rest of the protein where the SOLEXSY-measured rates are outside the valid region, there is good agreement between the two methods, although, as expected, the SOLEXSY-measured rates are slightly lower than those measured by CLEANEX-PM. However, the protection factors measured by both methods are also in good agreement, suggesting that any systematic error in the CLEANEX-PM results appears to be minimal.

In general, the high rates of hydrogen exchange and low protection factors throughout the protein support the conclusion that the entire protein, including the second exon product, is disordered. The rates of hydrogen exchange are much faster in the His-tag region (residues 1–20) compared to the rest of the protein (Figure 5.25), giving rise to the lowest protection factors in the protein. However, the protection factors of all residues are well below 1, paradoxically suggesting faster exchange than in a completely disordered protein. A likely explanation for this is that because Tat_{101} has a high fraction of basic residues it is highly positively charged (+29.2 at pH 4 and +14.1 at pH 7, estimated by Protein Calculator 3.4) and condenses hydroxyl ions around it, leading to a local pH higher than that measured in the bulk solvent. The higher local pH leads to an increased rate of base-catalyzed hydrogen exchange. The highest positive charge density in the protein can be found in the His tag region and in the basic segment, supporting this conclusion. In both experiments, the protection factors in the cysteine-rich region are noticeably higher, suggesting the possibility of conformational restraint in this region compared to the rest of protein. Above, it was noted that the cysteine-rich region also shows a propensity to form α-helix. The CLEANEX-PM results do not indicate the presence of any folding nuclei or folding propensity in any part of the protein.

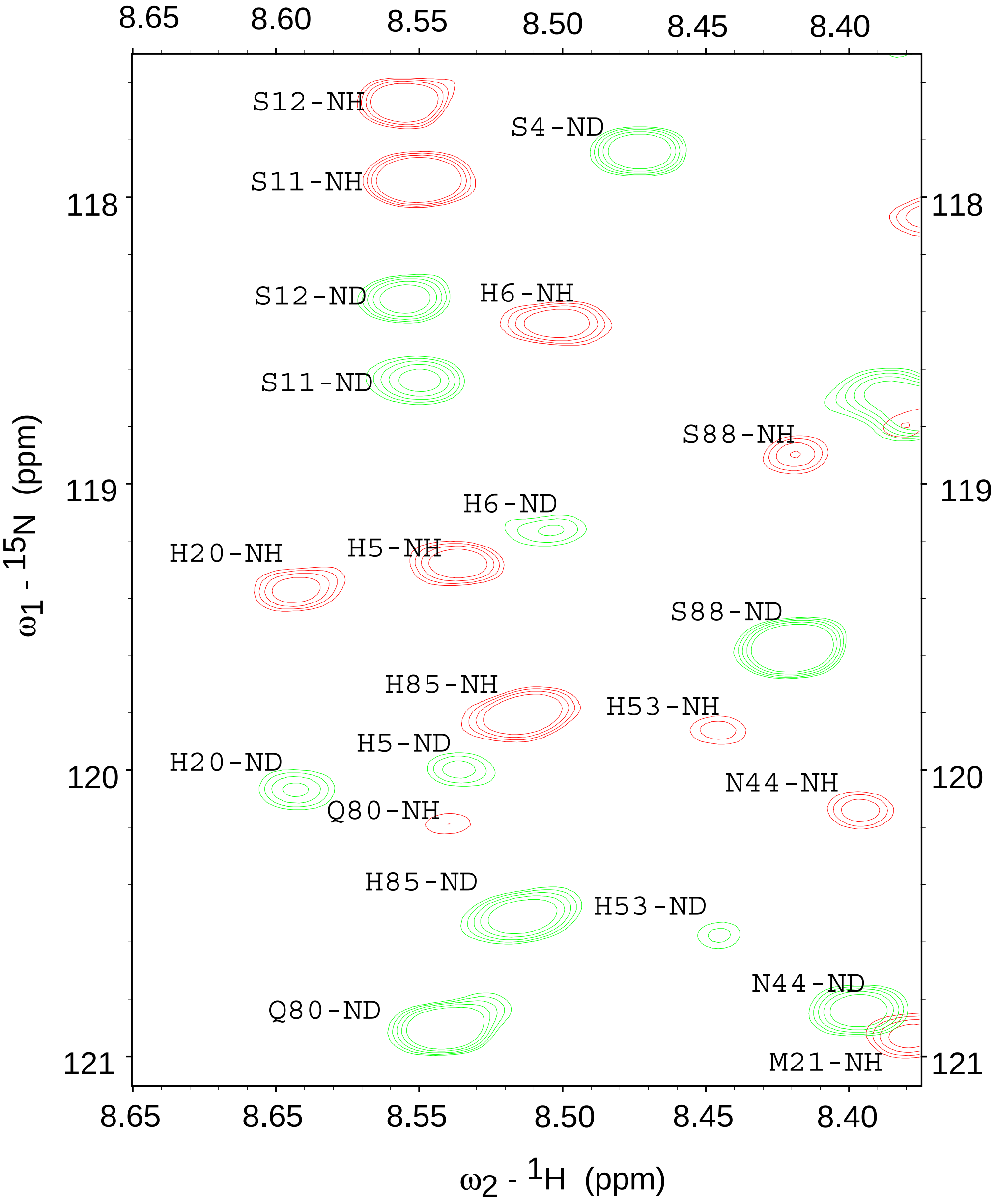


Figure 5.23. Overlay of a region of two SOLEXSY[329] frequency-labeled data points, N^H (red) and N^D (green), collected at a mixing time (t_{mix}) of 1000 ms. Peaks of the same residue are resolved by the deuterium shift of 0.7 ppm.

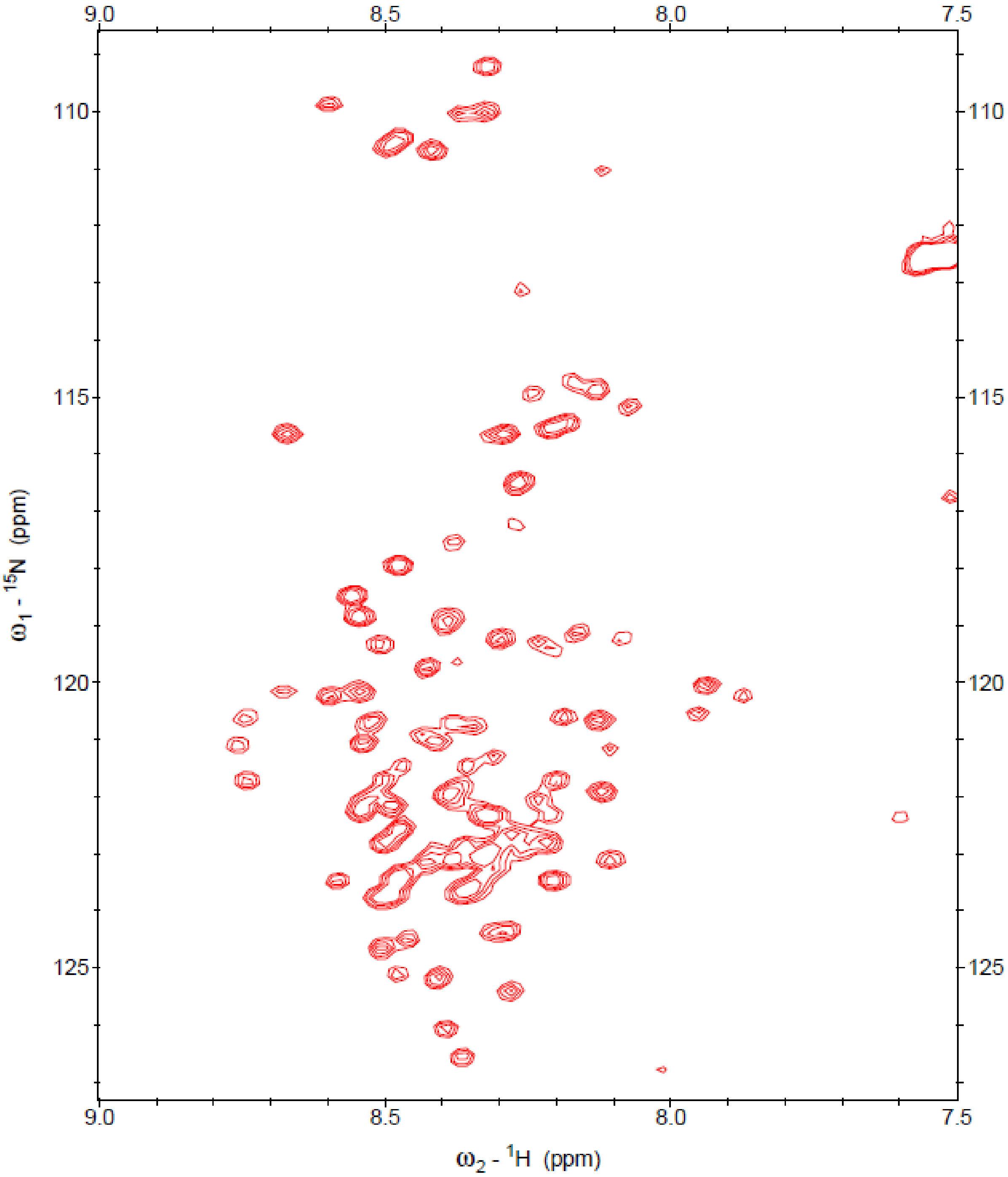


Figure 5.24. Sample spectrum (at 75 ms mixing time) for the hydrogen exchange measurement of His-tagged Tat$_{101}$ at pH 4.0 and 293 K using the CLEANEX-PM pulse program[329, 331]*. The spectrum was acquired on a Varian INOVA 600 MHz spectrometer.*

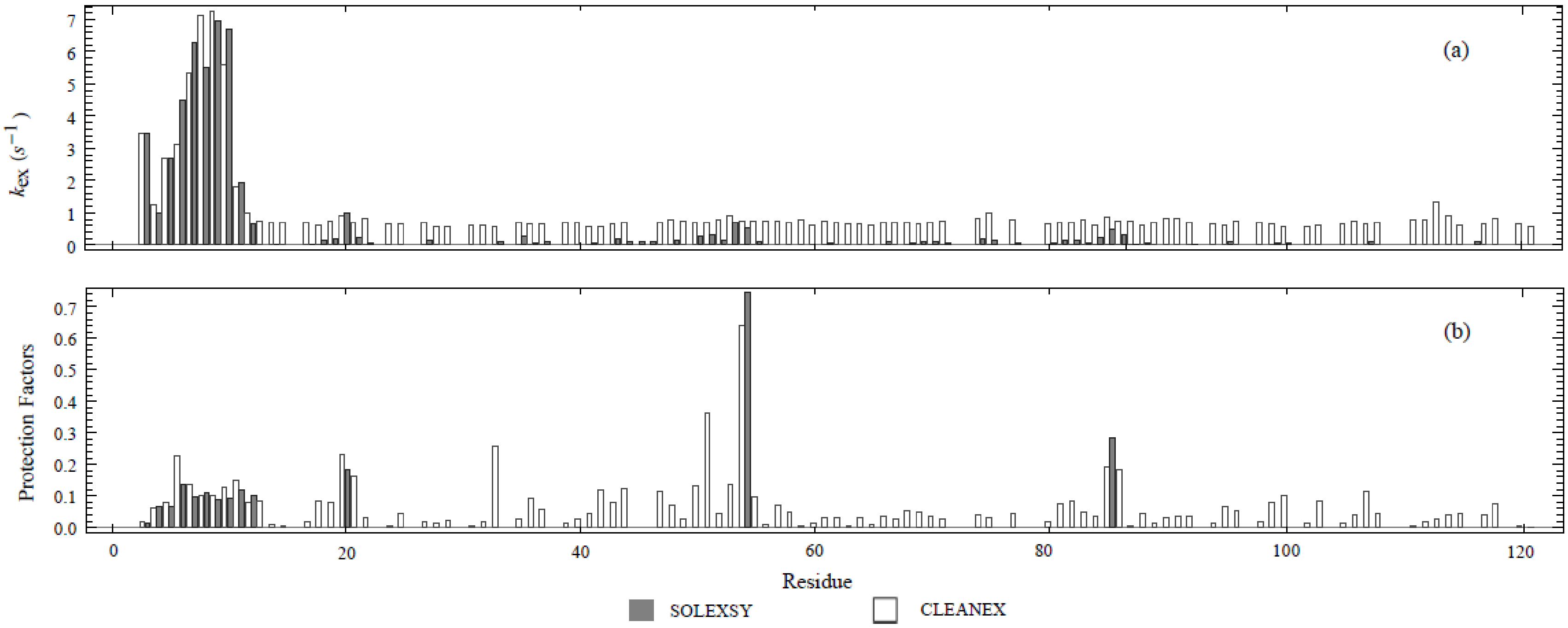


Figure 5.25. Exchange rate (a) and protection factor (b) of full-length Tat protein from HIV-1, determined using CLEANEX-PM (empty bars) and SOLEXSY (filled bars) methods.

5.7. Oxidation of Tat_{101} at pH 4.0 and pH effect.

During the course of these experiments, it was observed that the resonances of residues in the cysteine-rich region shifted and weakened in intensity over the course of several days (see Figure 5.26). By addition of TCEP and degassing of the protein, most of the resonances could be restored, although a few peaks in the cysteine-rich region never fully recovered their initial intensity.

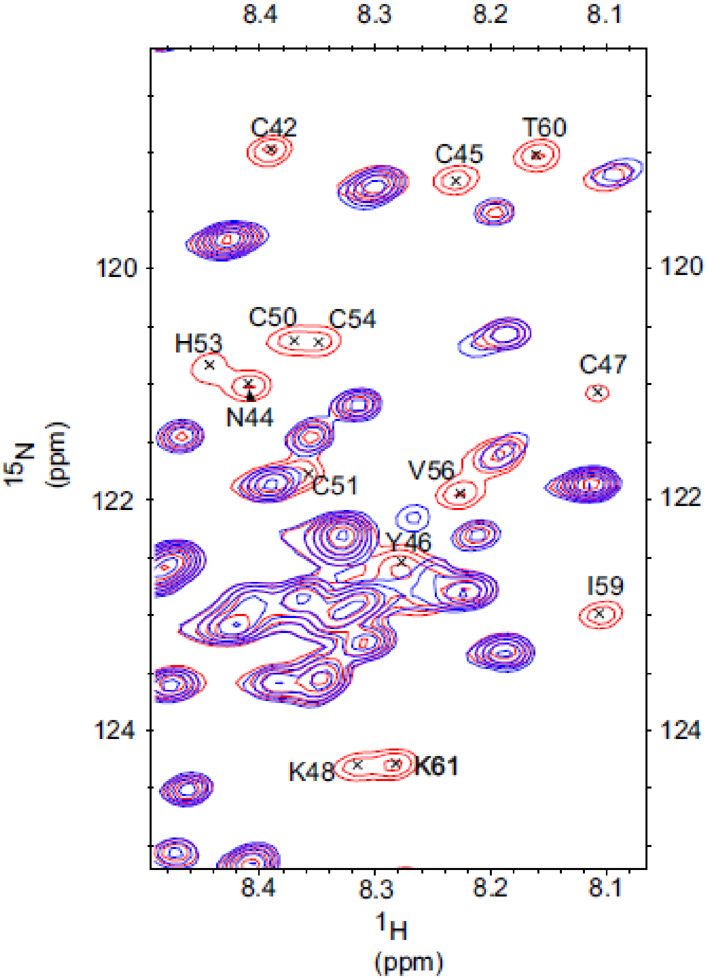


Figure 5.26. Cys-regions of ^{1}H-^{15}N HSQC spectra of full-length Tat_{101} protein in reduced (red) and oxidized (blue) states. Without reducing reagent, Tat_{101} is oxidized in the course of a few days, leading to the loss of peaks in the region.

This suggested that, even at pH 4.0 where all the cysteines should be fully reduced, Cys oxidation is still possible. This may be explained by the presence of basic residues present in the Cys-rich region that likely lower the thiol pK_a of one or more Cys residues, elevating their reactivity at low pH[408]; the overall net positive charge on the protein also likely contributes. With careful degassing and sealing of the NMR tubes, the protein could be kept in a fully reduced state for ≥7 days. Broadening of peaks and loss of signal intensity for residues in the cysteine region were noticeable even in sealed tubes after 15 days.

In an attempt to study the protein at pH values closer to physiological conditions, a pH titration was conducted. At close to neutral pH, fully reduced Tat_{101} shows many fewer resonances and broadened resonances (Figure 5.27). Several explanations are possible. The loss of peaks may be explained by protein aggregation, an increase in the rate of hydrogen exchange or by the effects of conformational exchange. The flat relaxation dispersion curves at all pH's (Section 5.5) rule out conformational exchange on the millisecond timescale as an explanation for the observed line-broadening. They do not rule out conformational exchange on the microsecond timescale. However, the fast hydrogen exchange rates that were measured (Section 5.6) suggest that the loss of peak intensity is best explained by rapid hydrogen exchange rather than the onset of slow conformational exchange.

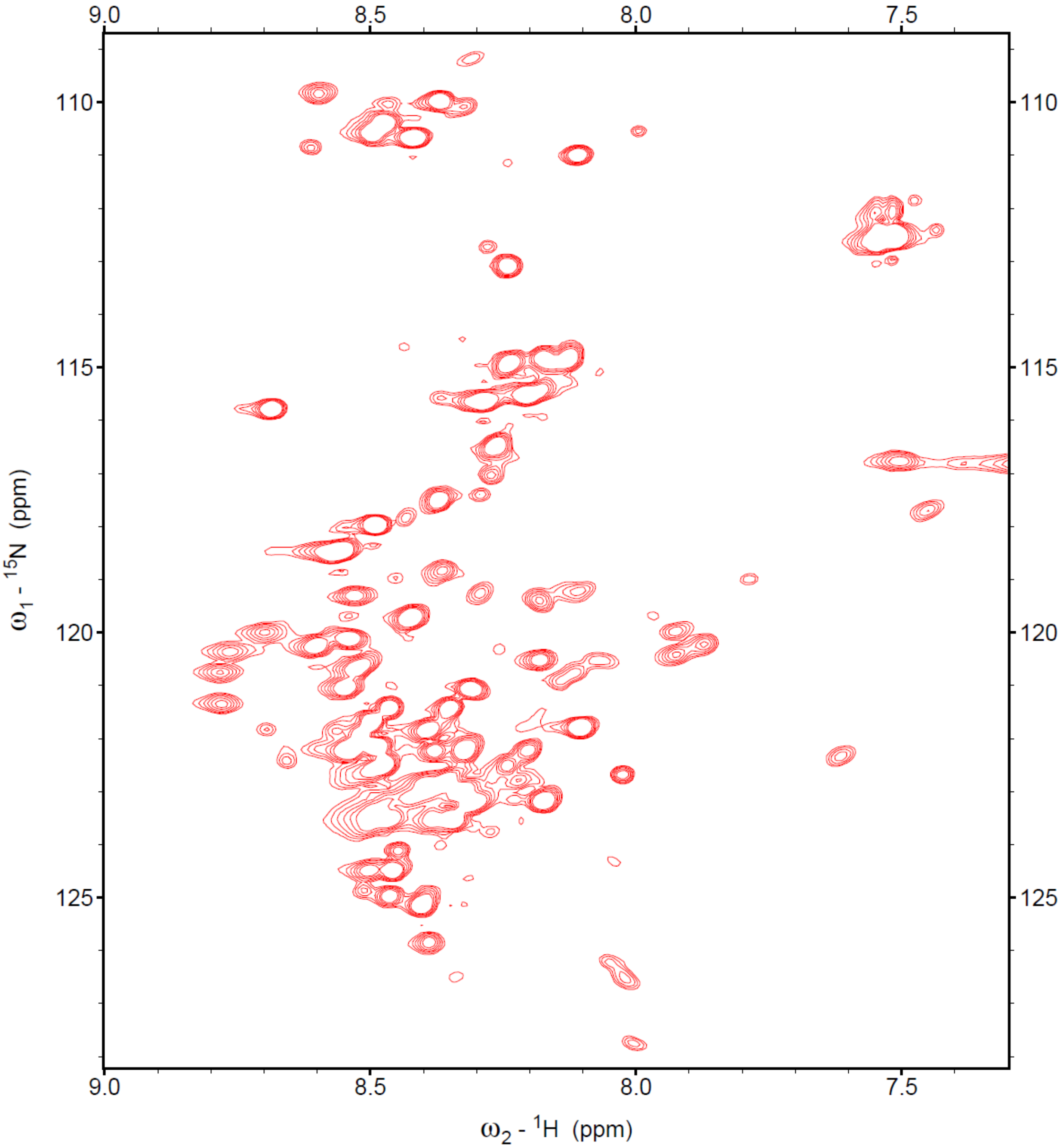


Figure 5.27a. ^{1}H-^{15}N HSQC spectrum of Tat$_{101}$ at pH 4.0 and 298 K, measured at 14.1 T

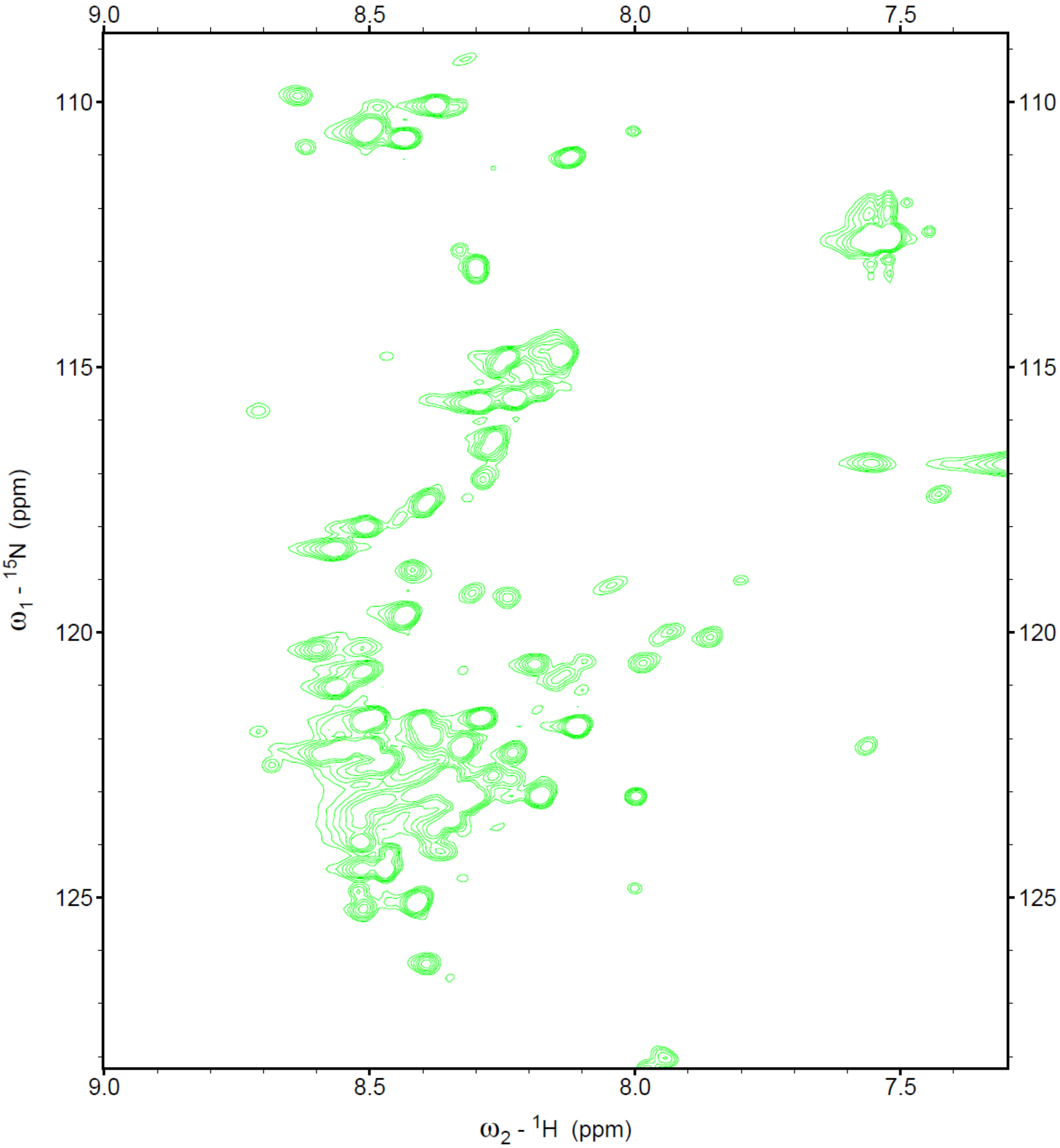


Figure 5.27b. ^{1}*H-*^{15}*N HSQC spectrum of Tat*$_{101}$ *at pH 5.0 and 298 K, measured at 14.1 T*

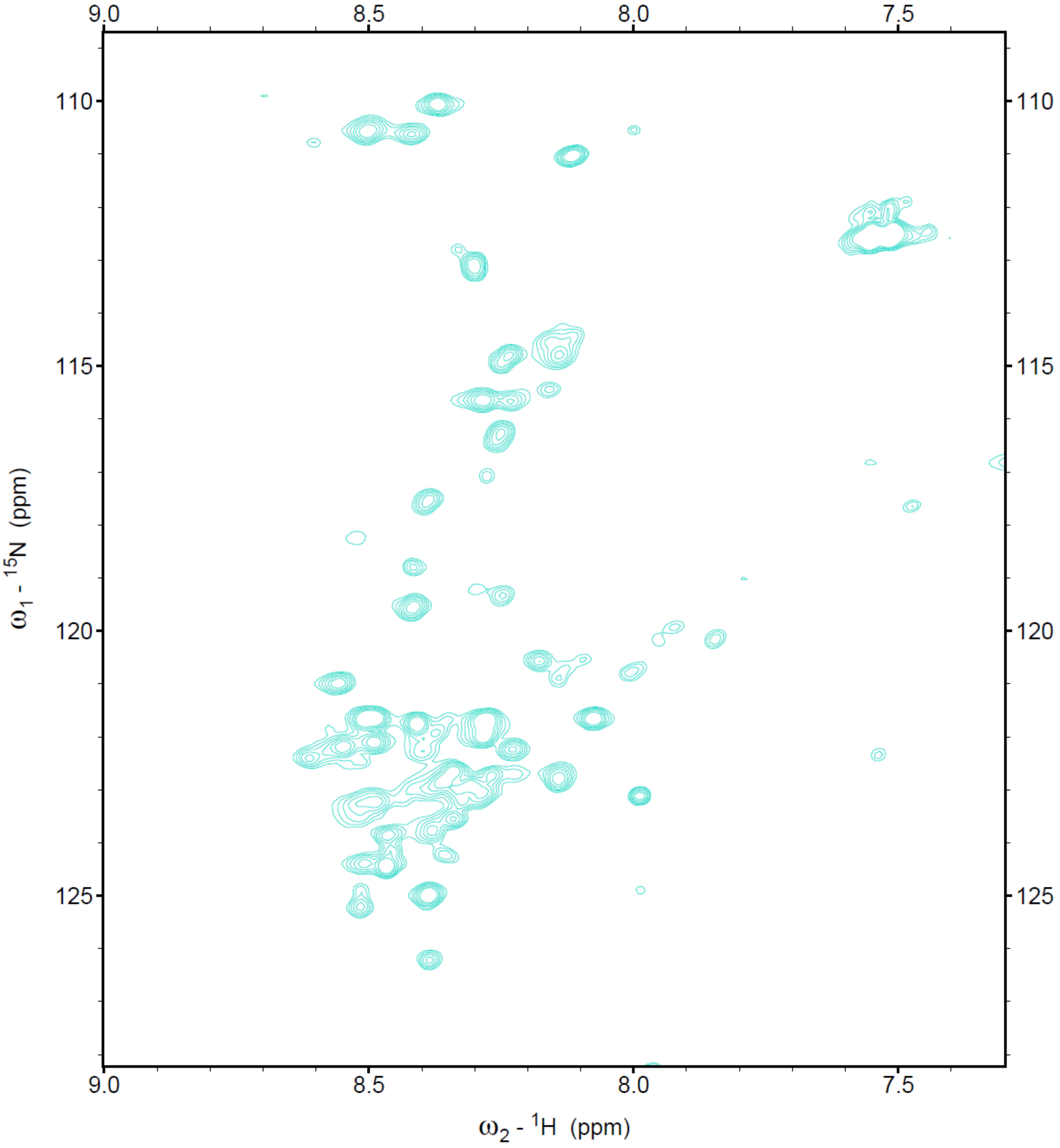


Figure 5.27c. ^{1}H-^{15}N HSQC spectrum of Tat$_{101}$ at pH 6.0 and 298 K, measured at 14.1 T

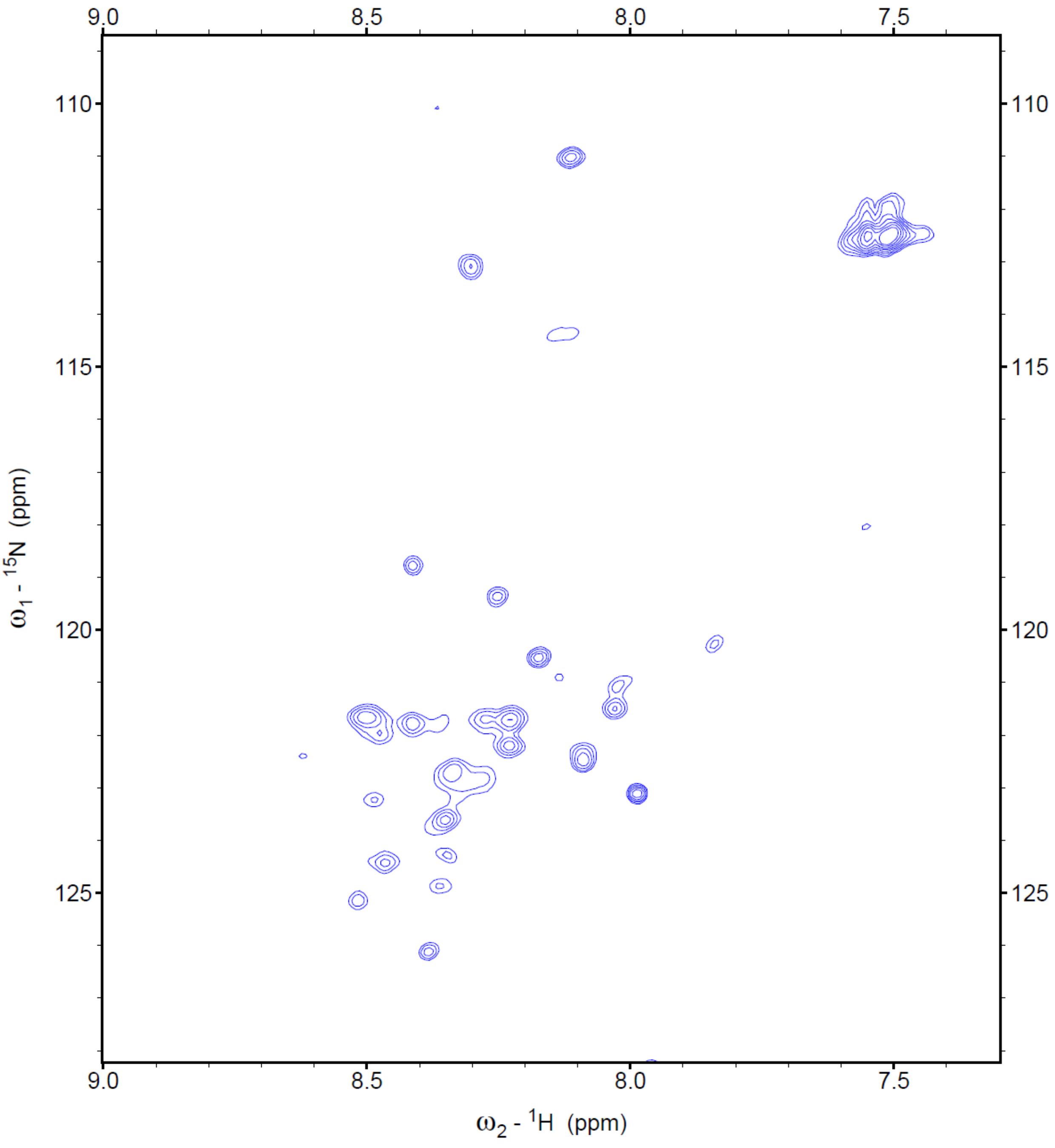


Figure 5.27d. ^{1}*H*-^{15}*N HSQC spectrum of Tat*$_{101}$ *at pH 7.0 and 298 K, measured at 14.1 T*

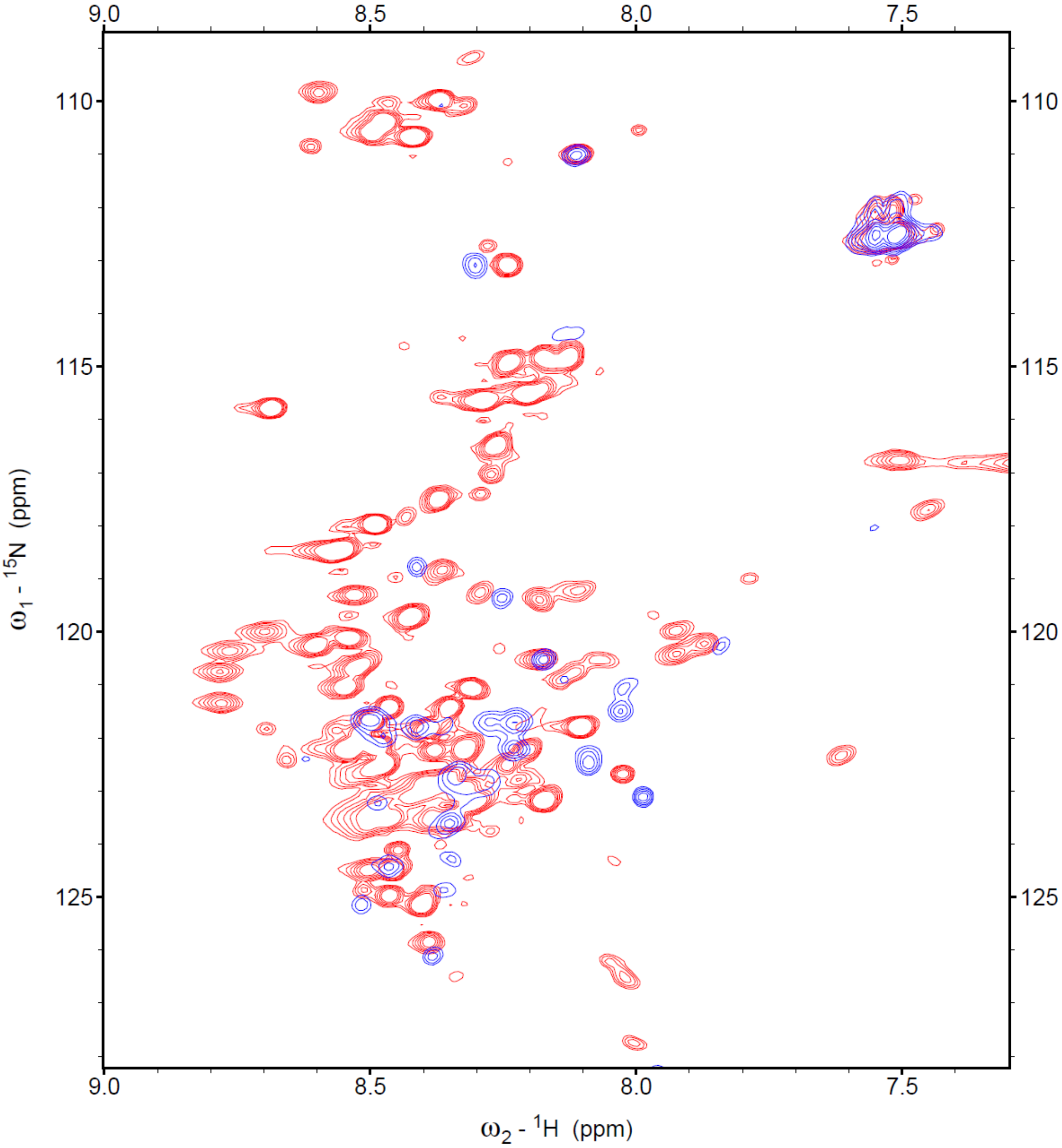


Figure 5.27e. Overlay of ^{1}H-^{15}N HSQC spectra of Tat$_{101}$ at pH 4.0 (red) and pH 7.0 (blue), 298 K, measured at 14.1 T

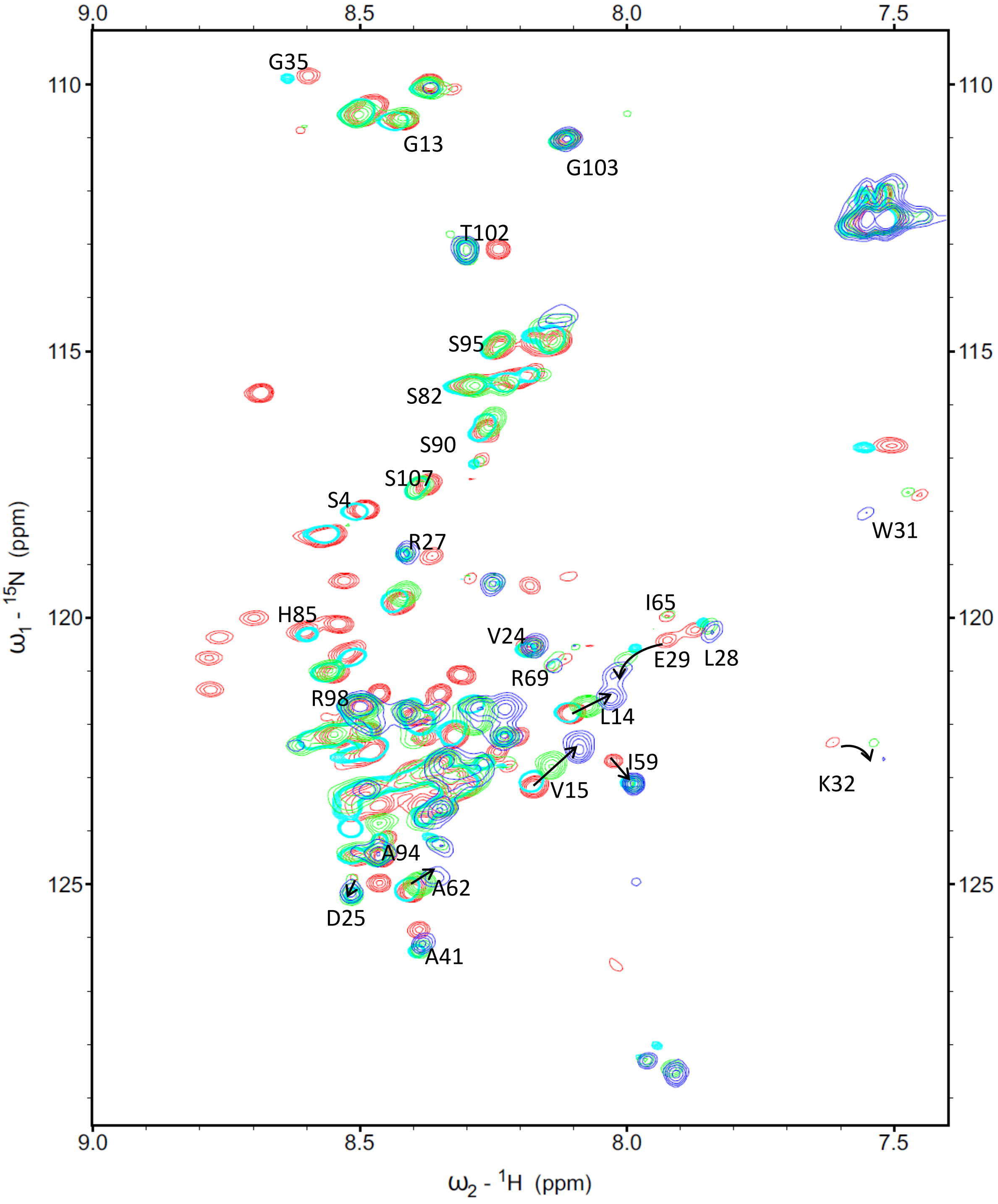


Figure 5.27f. Overlay of ^{1}H-^{15}N HSQC spectra of Tat$_{101}$ *at pH 4.0 (red), pH 5.0 (green), pH 6.0 (cyan) and pH 7.0 (blue), measured at 14.1 T and 298 K. Several resonances can be followed from pH 4.0 to pH 7.0, yielding a partial assignment for the higher pH spectra.*

5.8. Attempts to form a cyclin T1 – Tat_{101} complex in solution.

In the crystal structure of the Tat - pTEFb complex reported by Tahirov *et al.*[165], Tat was shown to directly bind to cyclin T1. However, the Tat construct was 86 amino – acids in length, and only the first 49 residues of Tat were visible in the electron density map, suggesting that only the visible parts of the protein adopt a fixed conformation in the complex. I sought to measure the changes in the structure and dynamics of full-length Tat_{101} upon interaction with cyclin T1 in solution by NMR spectroscopy.

In order to make the complex for crystallization, Tahirov *et al.* co-expressed three proteins of the complex, human Cdk9 (1–345), human cyclin T1 (1–266) and HIV-1 Tat (1–86) in insect cells. However, this method does not allow selective isotopic labeling of Tat protein for NMR study. Instead, I planned to mix ^{15}N-labeled Tat_{101} protein with non-labeled cyclin T1, both of which were separately purified, so that I could observe the resonances of ^{15}N- Tat_{101} without the interference from the cyclin T1. Both proteins were found to be highly soluble in HEPES buffer, and required TCEP for stability. Circular dichroism spectra (Figure 5.28) showed that cyclin T1 denatures at low pH, requiring that the pH for complex formation is higher than 6.2. As described above in Section 5.7, a pH titration of Tat_{101} showed that the NMR signals of the Tat_{101} protein are significantly reduced as the pH approaches neutrality. Therefore, a pH of 6.35 was determined to be a reasonable compromise for complex formation. Salt concentration also has a significant effect on both the solubility of Tat_{101} and the stability of cyclin T1. High salt concentration significantly reduces the solubility of Tat_{101} but improves the stability of cyclin T1. Therefore, the two proteins were prepared in 25 mM HEPES buffer, 2 mM TCEP, 100 mM NaCl and pH 6.35 before adding them together with the aim of seeing as many resonances of Tat_{101} as possible in conditions in which cyclin T1 would remain stable. I also speculated that cyclin T1 might be more stable when bound to Tat.

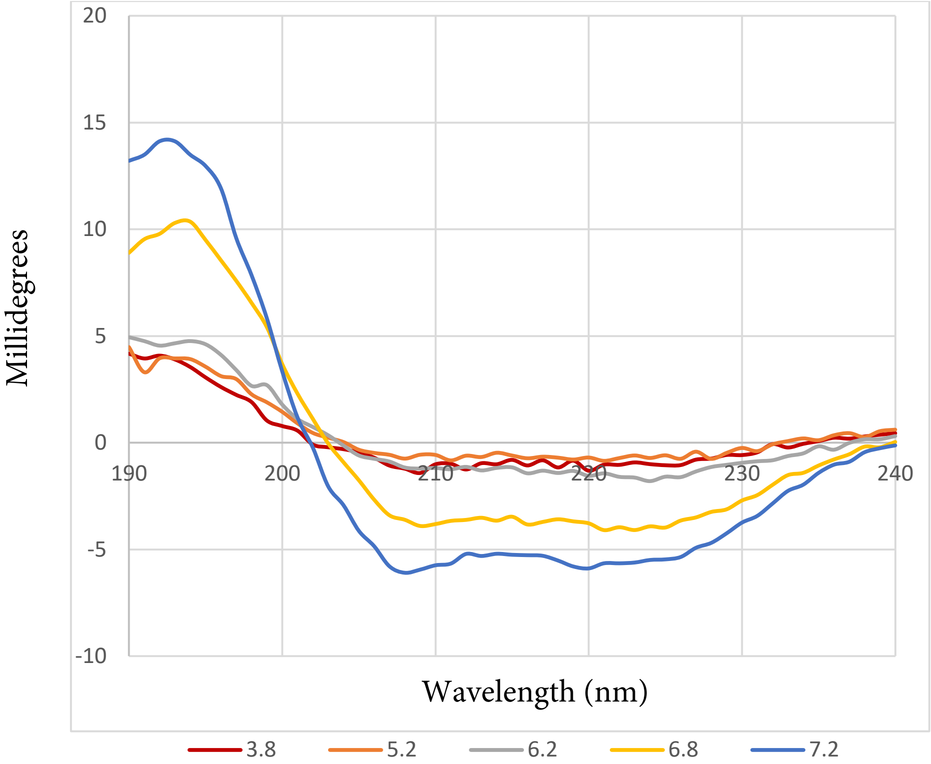


Figure 5.28. Circular dichroism of cyclin T1 at different pH.

In the first attempt, a small volume of concentrated cyclin T1 (2 µL) was added to 500 µL of concentrated Tat_{101} but unfortunately aggregates were observed to form immediately. In a second attempt, both proteins were diluted to 5 µM, and 25 mL of each protein solution were added together. The mixture formed no visible precipitate, and was subjected to spin concentration to a final volume of 3 mL. An NMR spectrum of the concentrated mixture (blue) was overlaid onto a reference spectrum of Tat_{101} protein measured under identical conditions but in the absence of cyclin T1 (red) and is shown in Figure 5.29. The spectrum of Tat_{101} in the presence of cyclin T1 showed many changes, including the appearance of new resonances, the diminishing of some resonance intensities and the shifting of several resonances, as indicated in the figure. These observations suggest that the Tat_{101} conformation was changed significantly upon interaction with cyclin T1. However, because the spectrum of Tat_{101} assigned at pH 4.0 was so different from that observed at pH 6.35 in the presence of cyclin T1, it was not possible to transfer any assignments to the new conditions. No further attempts were made to assign this spectrum however, in the future site-specific labeling of multiple amino acids or NMR experiments using residue – selective pulse sequences might be able to assign some of the peaks.

Another notable feature of the spectrum of Tat_{101} interacting with cyclin T1 is that the chemical shift dispersion of the peaks is small suggesting that although the Tat_{101} conformational ensemble may have changed significantly in some parts of the protein upon interaction with cyclin T1, it still appears to be substantially disordered.

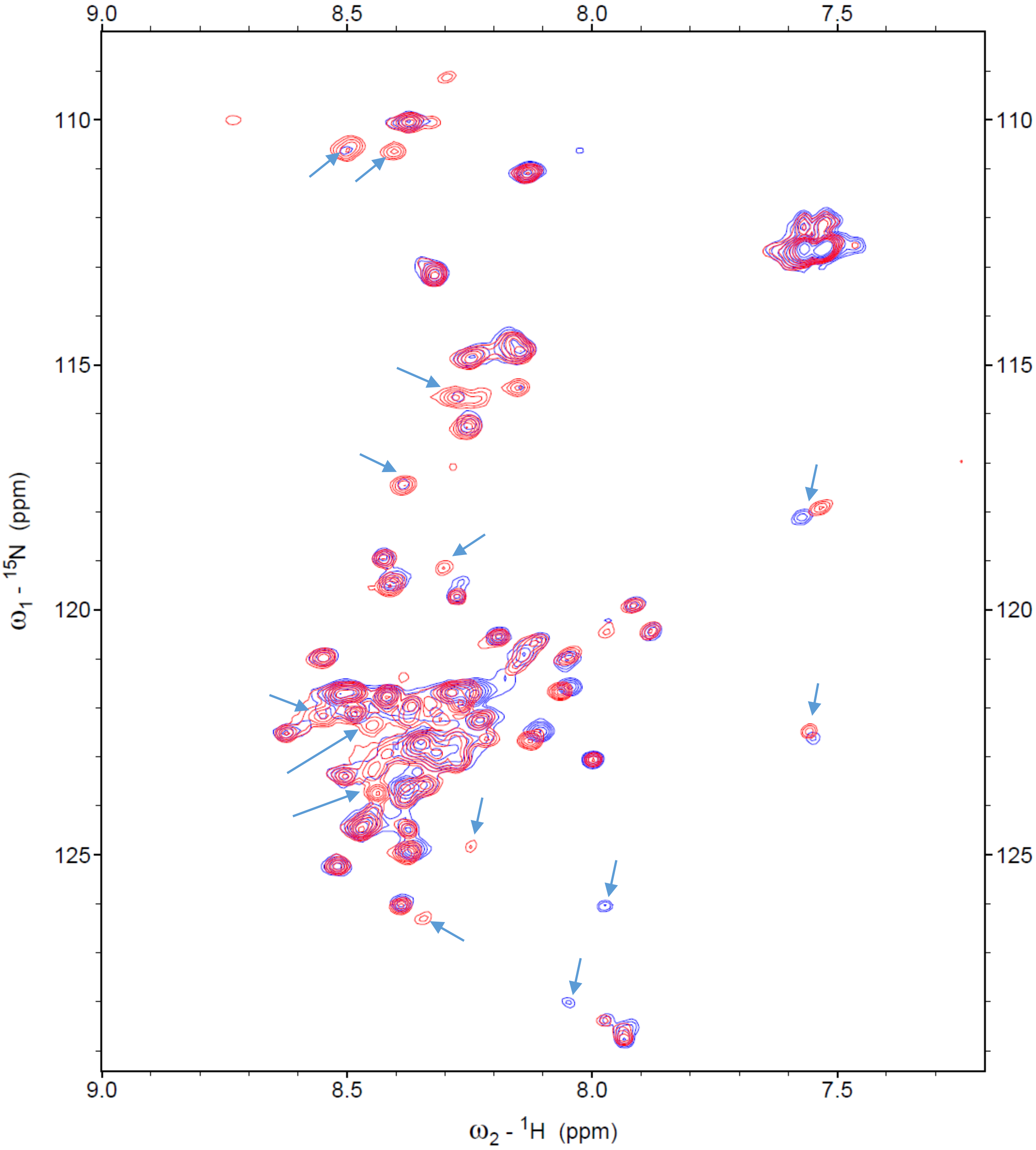


Figure 5.29. ^{1}H-^{15}N HSQC spectrum of Tat$_{101}$ - cyclin T1 complex (blue) overlaid onto ^{1}H-^{15}N HSQC spectrum of Tat$_{101}$ (reference spectrum) at 298K and pH 6.35 (red). The arrows indicate peaks that shifted, or changed in intensity compared with the reference. Spectra were acquired on the INOVA 600 MHz NMR spectrometer at University of Manitoba.

5.9. Molecular dynamics simulations.

Recently, advances in hardware and software have permitted long time-course classical molecular dynamics simulations of proteins including IDP's[409]. In order to explore the conformational preferences of HIV-1 Tat, I computed two 100 ns simulations of the protein solvated in water using GROMACS[343]. An IDP is an ensemble of many different conformations. Therefore, simulation of an IDP requires sampling a large potential energy surface (PES) with multiple minima corresponding to the possible conformations. By starting with a linear, fully-extended conformation, the bias toward any conformation or class of conformations can be avoided since the protein can randomly adopt any conformation during the simulation time. However, in comparison with a condensed, globular conformation a completely extended protein structure requires a much larger boundary box and a much higher number of water molecules to solvate the protein. And the longer the protein sequence, the more computationally demanding the simulation becomes. Therefore, instead of using full-length hexahistidine tagged protein, only the first exon of Tat (Tat_{72}) was simulated. A cubic solvent box required 1126485 mTIP3P water molecules to solvate Tat_{72}. This could be reduced significantly by using a rhombic dodecahedron where the number of water molecules was reduced to 442085 mTIP3P water molecules.

Two 100 ns simulations of Tat_{72} were conducted. Each simulation took 15 days to complete on a computer that had 36 physical cores (72 logical cores) and 2 GTX1080 Pascal GPUs, and generated 50 million structures for each of the 2 resulting trajectories. The resulting trajectories (A and B) showed different conformational evolution pathways, suggesting that the PES of Tat_{72} was not fully explored. In order to fully sample the PES, a multicanonical simulation method, such as replica exchange, would be needed[410]. However, the simulations still provide instructive information about the dynamics of Tat_{72}.

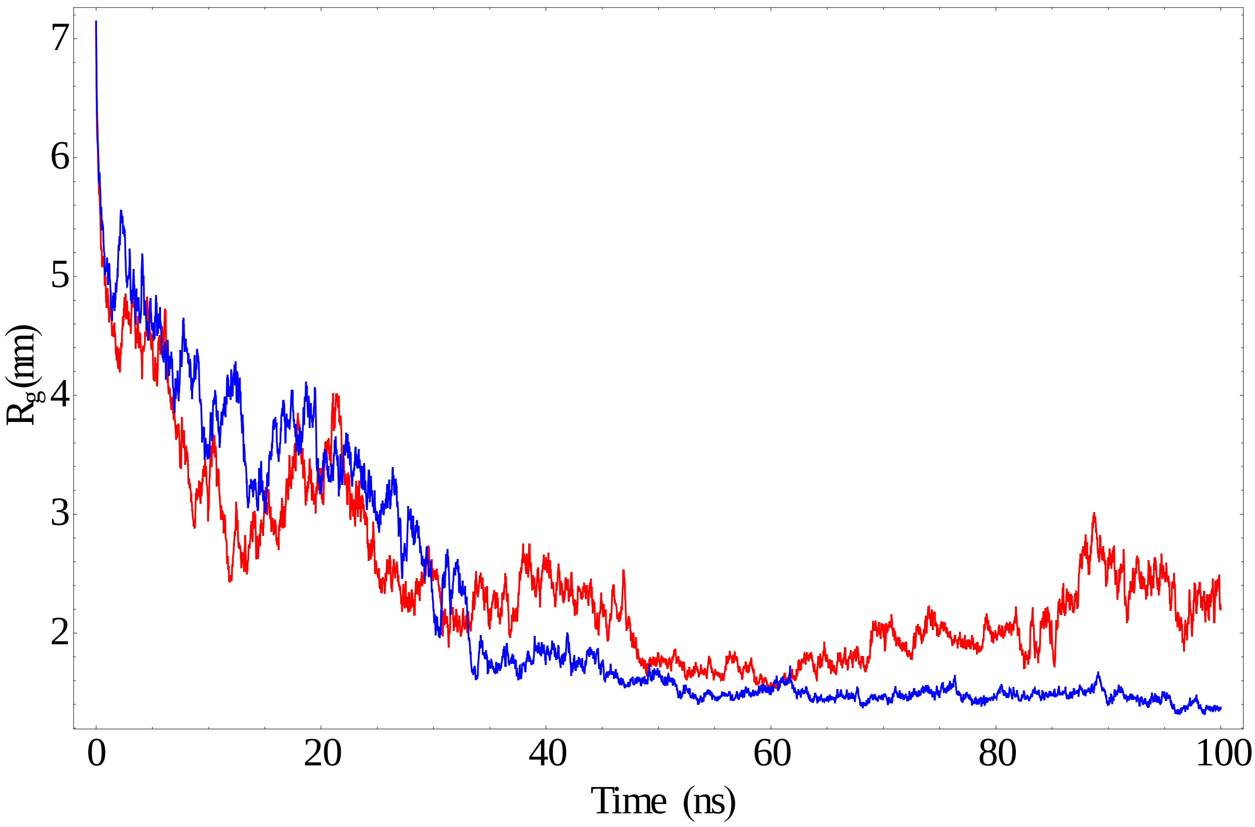


Figure 5.30. Radius of gyration of the two 100 ns simulations of Tat$_{72}$. The simulations were conducted on a computer with 72 logical cores and 2 GPUs, using a linear starting conformation. After the first 35 ns, both resulting trajectories (A in red and B in blue) adopt lower and fairly stable compactness.

The radius of gyration (R_g) represents the compactness of a protein. Formally, it is the root-mean-square distance of the atoms in the protein from its center of mass. Structural elements such as α-helix and β-sheet greatly increase the compactness of protein, resulting in a lower Rg. Folded proteins are much more compact compared with IDPs, and therefore have lower Rg values. Radius of gyration profiles for the two simulated trajectories are shown in Figure 5.30. At the beginning of the simulations, Rg values in both trajectories are high due to the highly extended starting structures of Tat$_{72}$. After 35 ns, the protein in both trajectories adopts lower and fairly stable compactness, reflected in the low standard deviations of 0.31 nm for trajectory A (red) and 0.13

nm for trajectory B (blue), for the remaining 65 ns in both simulations. The protein in trajectory B (blue) adopts a very stable R_g between 50-100 ns (1.2 - 1.4 nm) whereas the protein in trajectory A (red) is less compact (2.0 - 2.6 nm). The radius of gyration of trajectory A (red) has a small reduction after 10 ns (from 35 ns to 45 ns), and the protein becomes more compact for 40 ns (from 45 ns to 85 ns) before the compactness is increased again. This suggests a structural fluctuation during this time frame, and it happens on the nanosecond timescale. The radius of gyration of trajectory B remains stable over the final 60 ns of the simulation. The radius of gyration of a folded protein of similar size to Tat_{72} (8.345 kDa) is estimated to be 1.2 nm using the method proposed by Smilgies *et al.*[411] In both simulations, Tat_{72} shows a lower compactness, reflected in a higher value of radius of gyration, compared to a folded protein.

A timeline analysis was done using a program integrated in the GROMACS package, called DSSP (Define Secondary Structure of Proteins)[412, 413], showing the structure of the protein at each time step of the simulation (called a trajectory frame). The DSSP algorithm defines cooperative secondary structure as repeats of the elementary hydrogen-bonding patterns “turn” and “bridge”. Repeating turns are “helices,” repeating bridges are “ladders,” connected ladders are “sheets”, and curved segments are defined as “bends.” Results for trajectory A and B are shown in Figure 5.31 and 3.32 respectively.

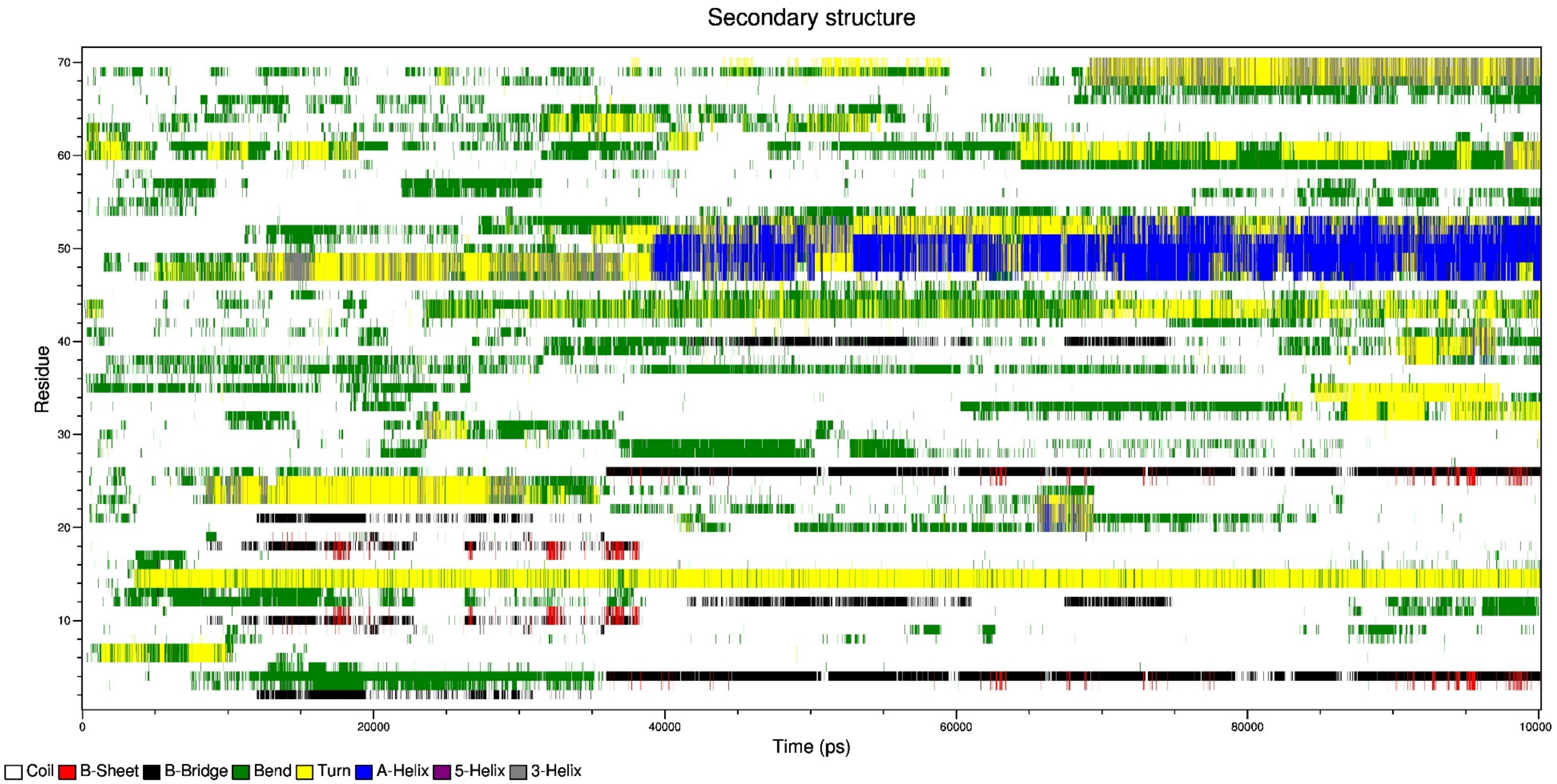


Figure 5.31. Secondary structure timeline analysis of trajectory A from 100 ns simulation of Tat$_{72}$

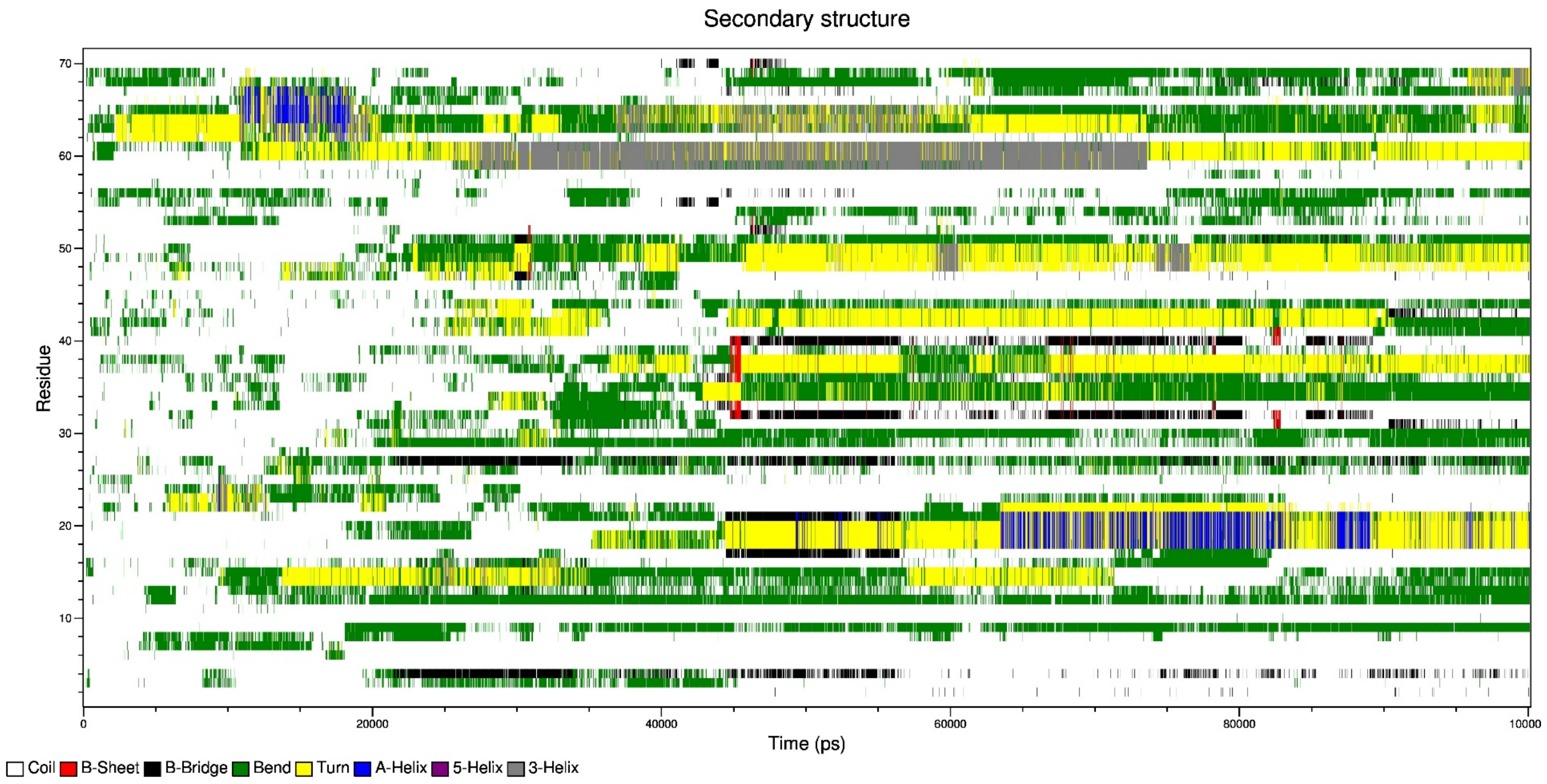


Figure 5.32. Secondary structure timeline analysis of trajectory B from 100 ns simulation of Tat$_{72}$

Both trajectory timelines, are dominated by non-structural elements such as coils-white, turns-yellow, and bends-green. This is expected from the NMR data reported above. In addition however, the trajectories show the formation of several transient structures. On trajectory A, the most obvious is the segment from Ser46 (Ser66 in hexahistidine tagged Tat_{101}) to Gln54 (Gln74 in hexahistidine tagged Tat_{101}) that forms a fairly stable α-helix after about 40 ns (blue in Figure 5.31). In trajectory B (Figure 5.32), this helix does not exist, instead, the segment contains a turn that transiently converts into a 3_{10} helix with a life time of around 2 ns. Therefore the existence of the helical segment is evidence of conformational fluctuation of Tat protein. This fluctuation may happen on a timescale longer than ps-ns as it is stable over the course of more than 60 ns on trajectory A. However, it is undetectable by NMR relaxation dispersion and likely does not remain stable on the µs – ms timescale. It also seems to be biologically relevant because the crystal structure of EIAV Tat and TAR in complex[394] also shows the formation of helical structure in the arginine-rich region (Leu48 - Ile59 of EIAV Tat) that covers this segment (Figure 5.33), in agreement with the simulation.

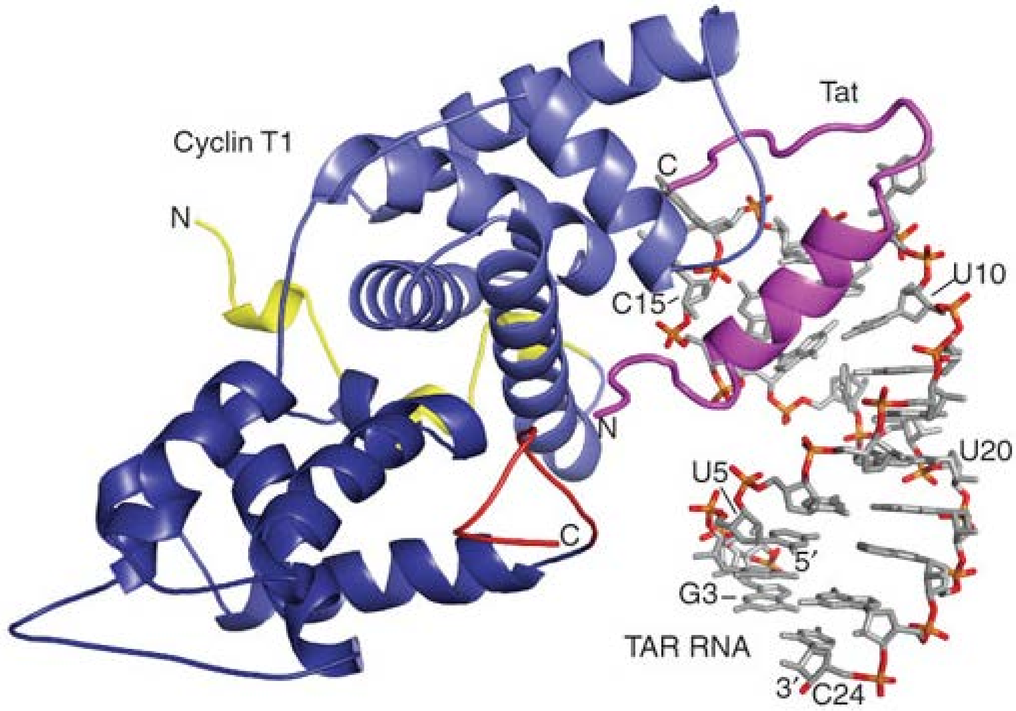
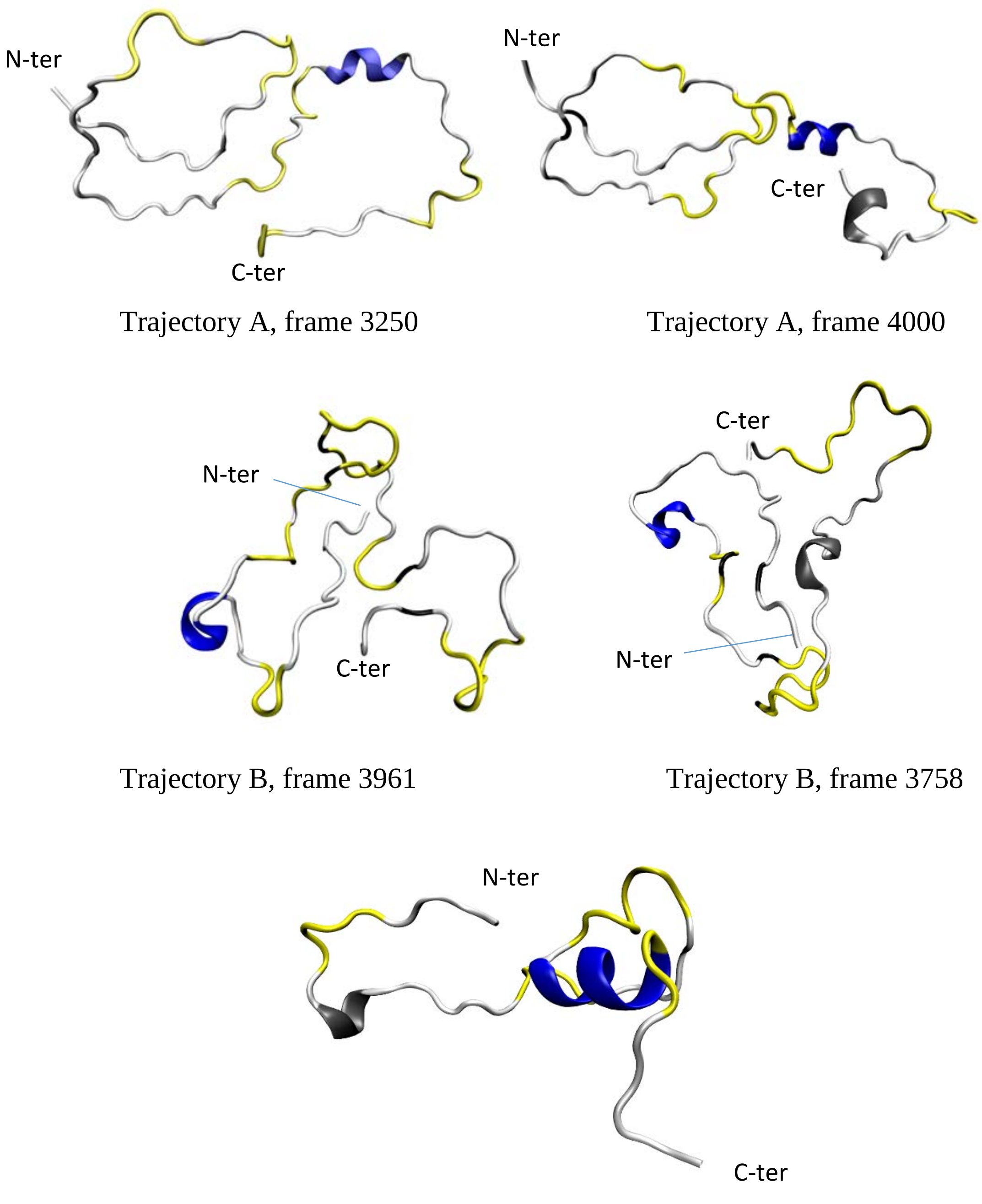


Figure 5.33. Structure of the cyclin T1 - EIAV Tat - TAR complex. The C-terminal region of EIAV Tat (residues 41–69, magenta) binds to the first cyclin box repeat of CycT1 (residues 5–267, blue) and also interacts with the major groove and loop region of TAR (nucleotides 3–24, gray). The cyclin-specific N- and C-terminal helices of CDK9 are colored yellow and red, respectively. Nucleotides A3 and U24 were mutated to G3-C24 to stabilize the stem formation at the 5' and 3' ends of TAR. Reprinted with permission from reference [394]*. Copyright © 2008 Nature Publishing Group.*

Other short transient structures that consist of 3 or 4 residues are also observed such as the α-helix from Pro17 to Ala21 (Pro37 - Ala41 in Tat_{101}) with a life time of around 15 ns, and the 3_{10} helix from Pro59 to Gly61 (Pro79 - Gly81 in Tat_{101}) with life a time of about 40 ns.

Figure 5.34 shows some conformations of Tat_{72} that are sampled during the simulation. Besides the structural elements (helices and sheets) recognized by the DSSP algorithm, a hairpin is formed at the N-terminus in both trajectories. Interestingly, it is also found in the crystal structure of Tat – pTEFb[165] (see Figure 5.34). This structure is stabilized in the complex by wrapping around the cyclin T1 of pTEFb. Overlaying structures from the simulation results onto the experimental structure does not yield a perfect fit due to the constant local fluctuation in the hairpin region in the free protein.



Trajectory A, frame 3250

Trajectory A, frame 4000

Trajectory B, frame 3961

Trajectory B, frame 3758

Tat in complex with pTEFb (not shown)

Figure 5.34. Sampled conformations during simulations of Tat$_{72}$ and the structure of the visible segment, (residues 1 – 49), of Tat$_{86}$ in the crystal structure of Tat-pTEFb complex[165] *(PDB entry 3mi9). The structures were generated using VMD and the structural elements are color encoded as follows: coil – white, turn – yellow, α-helix – blue, 3$_{10}$ - helix – gray.*

Figure 5.35 shows the order parameters for the N-H bond vectors of Tat$_{72}$ that were calculated using the last 60 ns of each simulation where the protein adopts a lower and fairly stable compactness as shown on the radius of gyration analysis (Figure 5.30). The values of S^2 below 0.8 throughout the protein in both simulations suggest the high flexibility of the Tat$_{72}$ backbone. In both simulations, the C-terminus of Tat$_{72}$ is more flexible than the rest of the protein, reflected in the lower order parameter value of the last 10 residues in this region. These observations are in agreement with the previously reported results on this protein segment[160], and also in agreement with the NMR results on the full-length Tat$_{101}$ protein described in previous sections. During the course of the simulations, the sampled conformations of Tat$_{72}$ in trajectory A (in red, top panel) are more flexible than in trajectory B, reflected in the significantly lower order parameters. Several residues have order parameter values close to zero with excessively high errors due to the insufficient simulation time. The discrepancy between the simulated and measured order parameters is further evidence that the entire PES was not sampled during each of the simulations.

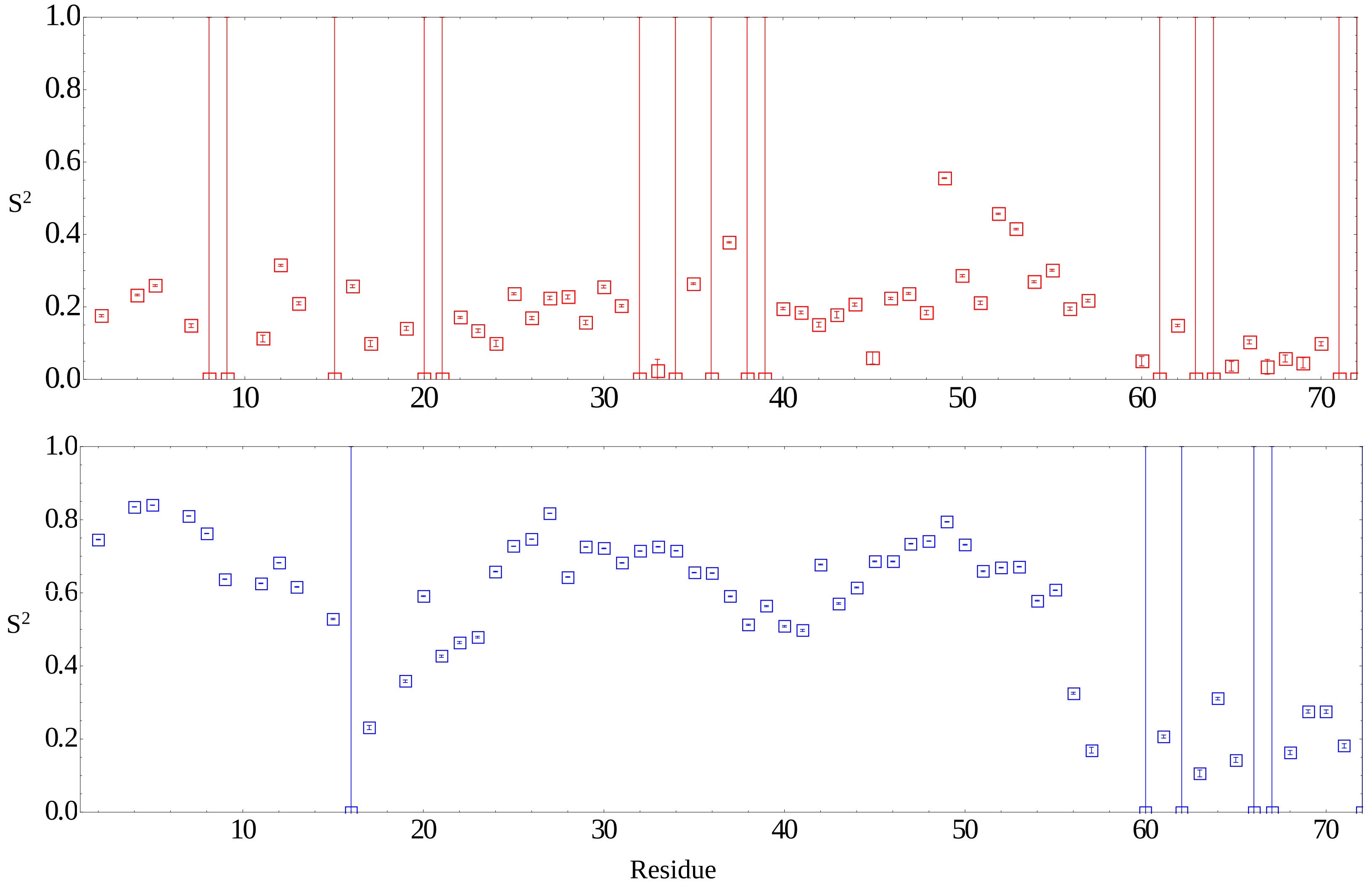


Figure 5.35. Order parameters calculated from the two simulations of Tat$_{72}$ protein, trajectory A in red in top panel and trajectory B in blue in bottom panel.

Chapter 6

Conclusions

Full-length Tat_{101} protein was successfully expressed and purified in high yield by using a bacterial expression system and metal affinity chromatography. The protein was isotopically labeled using different labeling strategies including uniform labeling, site-specific labeling and unlabeling, and the product was suitable for NMR study. During the purification, I found some evidence suggesting that the second exon of Tat increases the solubility of the protein.

The backbone resonances of non-proline residues of Tat_{101} protein were nearly completely assigned. The sequential assignment process was facilitated by uniformly ^{13}C- and ^{15}N-labeled protein. The site-specific labeling and unlabeling methods greatly reduced the spectral crowding and provided keystones for the assignment of full-length Tat. The unlabeling method used much cheaper chemicals (^{15}N-NH_4Cl and non-labeled amino-acids) compared to the site-specific labeling method which requires expensive isotopically labeled amino acids, and still provides a similar amount of information to support the sequential assignment process.

Relaxation and hydrogen exchange measurements on reduced Tat_{101} at pH 4.0 confirmed the disordered nature of the protein and the second exon product. The protection factors in backbone hydrogen exchange experiments were less than one and this suggested that the highly basic Tat protein concentrates hydroxide ions. To my knowledge this is the first time this has been observed for a protein. Spectral density mapping showed a high flexibility in the termini, and a slightly more restricted region around the only tryptophan residue suggesting the possibility of a propensity to fold there. Secondary structure propensity was mapped along the sequence, showing the protein tendency to adopt different conformations. Molecular dynamics simulations on the first exon of Tat showed transient structural elements formed during the simulation that is dominated by disordered conformations.

Comparisons of my results to published X-ray diffraction structures of Tat complexes strongly suggest that different segments of Tat function by different mechanisms. The acidic/proline-rich region has a propensity to form α-helix, but in the Tat-P-TEFb structure, the bound conformation contains β-turns and extended regions[165] (see Figure 1.16). Because the fraction of β-structure is undetectable, conformational selection seems unlikely unless the binding affinity of the bound conformer is sufficiently high to compensate for the very small fraction of protein in the β-conformation. It is more likely that this region of Tat forms the bound conformation through interactions with cyclin T1 in an induced-fit mechanism.

The TAR-binding region (Gly48 – Arg57 in Tat_{101} without the His-tag) shows a similar behavior. Free in solution, it has a propensity to form a β-conformation, which is likely to be a consequence of the high positive charge density in this segment. In contrast, in the EIAV-Tat-TAR-cyclin T1 complex, the structure is helical[394] (Figure 5.33) suggesting that binding to TAR induces the helical conformation. Another possibility is that the interaction between Tat and TAR is initiated by conformational selection by TAR for the transient α-helix on the N-terminus of Tat's arginine-rich region (Ser46 – Gln54) which is observed in MD simulation A (see Figure 5.31). After the initial contact, the anionic RNA would help direct the folding of the basic segment of Tat (Gly48 – Arg57) through charge compensation to form the helical structure. It has been argued previously[414] that despite the apparent entropy penalty in flexibility, preformed secondary structure may be advantageous during IDP binding interactions, suggesting a mechanism that combines elements of conformational selection and induced fit. This argument rests partly on experiments that show weaker binding of IDPs rigidified in the apparent bound conformation[415]. However, the reduced affinities of rigidified structures could be the result of an imperfect match with the bound conformation rather than a requirement for flexibility during the binding process. They also do not take into account the likelihood that bound structures are flexible. TAR, for example, is a highly flexible molecule, and flexibility in the bound conformation of Tat may lower the entropy penalty of rigidification of TAR.

The cysteine-rich region (C22 – C37 in Tat_{101} without the His-tag) was found to be prone to oxidation even in reduced conditions and low pH. It is predicted to form an α-helix region in most structure prediction programs (Figure 5.12). Experimental data shows a measurable propensity to form the helical structure that is found in the Tat-P-TEFb structure. However, in the two 100 ns simulation trajectories, this conformation is not observed, suggesting that the formation of the helix may happen on a longer timescale that requires a longer simulation time to sample. It is also possible that this region requires zinc ions to chaperone the folding. Indeed, it’s very likely that the folding of the cysteine-rich region is strongly influenced by zinc ions *in vivo* and their absence both from NMR experiments and in the MD simulations is a weakness of these studies. The addition of zinc would significantly improve the validity of the structural and dynamics information obtained for this region of Tat.

IDPs are often described as having low complexity because they are rich in some amino acids and sometimes exhibit residue repeats[70]. Especially in the case of multifunctional hub proteins such as Tat, the low sequence complexity hides an exquisite design that has been fine-tuned by evolution. HIV-1 Tat_{101} contains multiple segments that can bind to multiple proteins and/or nucleic acids with different affinities while at the same time avoiding misfolding and aggregation in the unbound state. Random polymers, while highly flexible, tend to form glassy structures in which single mutations lead to local free energy minima and kinetic traps preventing folding and the biological activity upon which it depends[201]. In contrast, the Tat_{101} sequence is highly conserved to allow it to exist in an unfolded conformation that avoids kinetic traps but is poised to fold in the presence of partners. Understanding the relationships among IDP sequence, structure, and dynamics will not only deepen our understanding of IDP behavior but also likely lead to new inhibitors in a variety of diseases, including viral infections[416, 417].

Given the results of this research, there are many directions in which to take future studies. Of key importance is to study Tat_{101} in the process of binding with partners, which will require development of protocols to form Tat_{101} complexes *in vitro*. In the crystal structures of Tat-pTEFb[165] and Tat - pTEFb - AFF4[418, 419], only the first exon product of Tat (Tat_{86}) was used. In the most recent study that reports the crystal structure of Tat - pTEFb - AFF4 - TAR[178], the Tat construct used was even shorter, having only 57 amino acids in length. In both cases, only the first 49 residues of Tat were visible. Therefore, NMR or molecular dynamics simulations of the complexes using the full-length Tat_{101} protein will provide a lot more useful information on the remaining regions of the protein.

FUTURE DIRECTIONS:

In my research, I attempted to make soluble complexes of full-length Tat with TAR RNA as well as with cyclin T1 in order to study the changes in the structure and dynamics of Tat induced by its binding partners. All of the complexes I made were insoluble at the relatively high concentrations needed for analysis by NMR spectroscopy. In contrast, as mentioned above, several soluble complexes of Tat fragments have been achieved for the purposes of crystallization for X-ray diffraction analysis[165, 177, 419]. It's possible that the methods used in those studies might permit the formation of soluble Tat complexes for NMR purposes. One approach is to express Tat_{101} as a fusion protein with a binding partner. Another approach is to co-express Tat_{101} with a binding partner and isolate the soluble complex from the expression host. However, neither of these approaches has been shown to work with full-length Tat_{101} and it's possible that the large amount of disordered unbound protein may result in insolubility of complexes, at least at high concentrations. So, these approaches may work only for short Tat fragments. Also, it's important to point out the solubility requirements of NMR and X-ray diffraction are different. For X-ray diffraction, concentrated samples must be temporarily soluble and crystallizable whereas for NMR, samples need to be stable and soluble at high concentrations for about 1 week. Finally, it's worth mentioning that in one of these reports among dozens of trials complexes prepared under identical conditions only one yielded useful data illustrating the difficulties and unpredictability in the preparation of these complexes.

Following the X-ray crystallographers, one might also consider studying fragments of Tat that might be much more soluble in complexes than the full-length protein. If each of the binding regions of Tat_{101} interacts with a binding partner independently of the rest of the protein regions then this could be a powerful approach to understanding Tat_{101} structure, function and dynamics both by NMR spectroscopy and MD simulations (see below). Indeed, we have shown here that the entire protein is intrinsically disordered and this suggests that each region might fold and function independently of the rest of the protein. We were reluctant to take this approach in the present study as it presupposes the independence of each part of the protein and it is certainly possible that the folding of one region of Tat might influence the folding of adjacent regions. Furthermore, *in vivo*, the functioning protein is 101 residues and we wanted our results to be as directly applicable to a living cell as possible.

Along these lines the in-cell NMR technique[420] might also be used in the future to further study the full-length Tat_{101} protein. Both prokaryotic and eukaryotic cells can be used as the host system. For prokaryotic host cells, the cell culture is subjected to cell growth and induction protocols so that the recombinant protein will be overexpressed to a sufficient level to be detected by NMR above the other cellular components. For eukaryotic host cells, the labeled protein is usually purified first, and then incorporated into cells by microinjection, or through re-sealable pores in the membrane. High-resolution heteronuclear multidimensional NMR spectra of proteins in living cells can be observed. As opposed to "*in vitro* NMR", the technique is capable of providing highly important information about the behavior of the protein in its functioning environment, as well as indicating effects such as crowding in the cellular environment on the protein.

In my research, I used NMR-based chemical shift and scalar coupling constant measurements to characterize the conformational landscape of unbound Tat_{101}. These measurements were supplemented with preliminary MD simulations. More detail about the conformations of Tat_{101} may be measurable by several different NMR approaches. For example, residual dipolar coupling (RDC) measurements[421], in which the protein is dissolved in an anisotropic medium to restrict the overall reorientation, giving rise to non-zero RDCs, may provide additional information about specific structural properties such as transient secondary and tertiary structures. Another experiment may be used is the paramagnetic relaxation enhancement (PRE) measurement[421] which can provide the transient long-range contacts of the protein. Both of these approaches have been used successfully when applied to IDP's and yielded valuable information about the conformational landscape of the proteins.

The molecular dynamics simulations in this project generated two 0.1 μs trajectories, which likely insufficiently sampled the PES of the Tat protein. A possible improvement is to use multicanonical molecular dynamics simulations to enhance sampling of the system and explore the full conformational ensemble. Also, by extending the simulation time to the millisecond range, the dynamics on the microsecond and millisecond timescales can be explored, providing close to the full dynamic landscape of the Tat protein. In the long-term, when computational speed has increased sufficiently it should be possible to simulate the full-length Tat protein in interactions with each and all of its binding partners. Such studies will help to determine the mechanism(s) by which Tat interacts with its partners and enhances transcription activation. This knowledge might lead to the design of new anti-HIV-1 therapies because Tat is essential to viral replication. Unfortunately, we may have to wait perhaps a decade for this approach to become a possibility.

One of the difficulties encountered in the MD simulations of Tat is the large number of water molecules that must be simulated because the simulation starts with a fully extended polypeptide. The number of water molecules could be greatly decreased by simulating the dynamics of each of the individual binding segments as discussed above. It might even be possible to run long simulations of Tat fragments with small binding partners such as TAR and cyclin T1 using current technology. Another possibility is the simulation of full-length Tat_{101} or of its arginine-rich fragment in a lipid bilayer that might provide insight on the mechanism through which the protein crosses membranes and is expelled from HIV-1 infected cells.

References

[1] Fischer, E. (1984) Einfluss der Configuration auf die Wirkung der Enzyme, *Berichte der deutschen chemischen Gesellschaft 27*, 2985-2993.

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

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

[4] Anfinsen, C. B., Redfield, R. R., Choate, W. L., Page, J., and Carroll, W. R. (1953) Studies on the gross structure, cross-linkages, and terminal sequences in ribonuclease, *Journal of Biological Chemistry*, 201-210.

[5] Blake, C. C. F., Koenig, D. F., Mair, G. A., North, A. C. T., Phillips, D. C., and Sarma, V. R. (1965) Structure of hen egg-white lysozyme, *Nature 206*, 757-761.

[6] Wyckoff, H. W., Hardman, K. D., Allewell, N. M., Inagami, T., Tsernoglou, D., Johnson, L. N., and Richards, F. M. (1967) The structure of Ribonuclease-S at 6 A resolution, *Journal of Biological Chemistry 242*, 3749-3753.

[7] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne., P. E. (2000) The Protein Data Bank, *Nucleic Acids Res. 28*, 235-242.

[8] McMeekin, T. L. (1952) Milk protein, *J. Food Prot. 15*, 57-63.

[9] Jirgensons, B. (1958) Optical rotation and viscosity of native and denatured proteins. X. Further studies on optical rotatory dispersion, *Arch Biochem. Biophys. 74*, 57-69.

[10] Jirgensons, B. (1965) Classification of Proteins According to Conformation, *Die Makromolekulare Chemie 91*, 74-86.

[11] Grizzuti, K., and Perlmann, G. E. (1970) Conformation of the phosphoprotein, phosvitin, *J. Biol. Chem. 245*, 2573-2578.

[12] Vogel, H. J. (1983) Structure of hen phosvitin: A 31P NMR, 1H NMR, and laser photochemically induced dynamic nuclear polarization 1H NMR study, *Biochemistry 22*, 668-674.

[13] Arnone, A., Bier, C. J., Cotton, F. A., and al., e. (1971) A high resolution struction of an inhibitor complex of the extracellular nuclease of Staphylococcus aureus. I. Experimental procedures and chain tracing., *J. Biol. Chem. 246*, 2302-2316.

[14] Holt, C., and Sawyer, L. (1993) Caseins as rheomorphic proteins: interpretation of primary and secondary structures of the αS1-, β- and κ-caseins, *J. Chem. Soc., Faraday Trans., 89*, 2683-2692.

[15] Pullen, R. A., Jenkins, J. A., Tickle, I. J., Wood, S. P., and Blundell, T. L. (1975) 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*, 5-20.

[16] Cary, P. D., Moss, T., and Bradbury, E. M. (1978) High-resolution proton-magnetic-resonance studies of chromatin core particles., *Eur. J. Biochem. 89*, 475-482.

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

[18] Weinreb, P. H., Zhen, W., Poon, A. W., Conway, K. A., and Lansbury, P. T. J. (1996) NACP, a protein implicated in Alzheimer's disease and learning, is natively unfolded, *Biochemistry 35*, 13709-13724.

[19] Wright, P. E., and Dyson, H. J. (1999) Intrinsically unstructured proteins: re-assessing the protein structure - function paradigm, *J. Mol. Biol. 293*, 321-352.

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

[21] Chen, J., Liang, H., and Fernandez, A. (2008) Protein structure protection commits gene expression patterns, *Genome Biol. 9*.

[22] Uversky, V. N. (2003) A protein - chameleon: conformational plasticity of alpha-synuclein, a disordered protein involved in neurodegenerative disorders, *J. Biomol. Struct. Dyn. 21*, 211-245.

[23] Dunker, A. K., and Uversky, V. N. (2010) Drugs for 'protein clouds': targeting intrinsically disordered transcription factors, *Curr. Opin. Pharmacol. 10*, 782-790.

[24] Livesay, D. R. (2010) Protein dynamics: dancing on an ever-changing free energy stage, *Curr. Opin. Pharmacol. 10*, 706-714.

[25] Habchi, J., Tompa, P., Longhi, S., and Uversky, V. N. (2014) Introducing Protein Intrinsic Disorder, *Chem. Rev. 114*, 6561-6588.

[26] Piovesan, D., Tabaro, F., Micetic, I., Necci, M., Quaglia, F., Oldfield, C. J., Aspromonte, M. C., Davey, N. E., Davidovic, R., Dosztanyi, Z., Elofsson, A., Gasparini, A., Hatos, A., Kajava, A.

V., Kalmar, L., Leonardi, E., Lazar, T., Macedo-Ribeiro, S., Macossay-Castillo, M., Meszaros, A., Minervini, G., Murvai, N., Pujols, J., Roche, D. B., Salladini, E., Schad, E., Schramm, A., Szabo, B., Tantos, A., Tonello, F., Tsirigos, K. D., Veljkovic, N., Ventura, S., Vranken, W., Warholm, P., Uversky, V. N., Dunker, A. K., Longhi, S., Tompa, P., and Tosatto, S. C. E. (2016) DisProt 7.0: a major update of the database of disordered proteins *Nucleic Acids Res.*

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

[28] Uversky, V. N., Gillespie, J. R., and Fink, A. L. (2000) Why are "natively unfolded" proteins unstructured under physiologic conditions?, *Proteins: Structure, Function, and Bioinformatics 41*, 415-427.

[29] Lise, S., and Jones, D. T. (2005) Sequence patterns associated with disordered regions in proteins, *Proteins 58*, 144-150.

[30] Prilusky, J., Felder, C. E., Zeev-Ben-Mordehai, T., Rydberg, E., Man, O., Beckmann, J. S., Silman, I., and Sussman, J. L. (2005) FoldIndex: s simple tool to predict whether a given protein sequence is intrinsically unfolded, *Bioinformatics*.

[31] Quevillon-Cheruel, S., Leulliot, N., Gentils, L., van Tilbeurgh, H., and Poupon, A. (2007) Production and crystallization of protein domains: how useful are disorder predictions?, *Curr. Protein Pept. Sci. 8*, 151-160.

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

[33] Fukuchi, S., Amemiya, T., Sakamoto, S., Nobe, Y., Hosoda, K., Kado, Y., Murakami, S. D., Koike, R., Hiroaki, H., and Ota, M. (2014) IDEAL in 2014 illustrates interaction networks composed of intrinsically disordered proteins and their binding partners, *Nucleic Acids Res. 42*, D320-D325.

[34] Potenza, E., Domenico, T. D., Walsh, I., and Tosatto, S. C. E. (2014) MobiDB 2.0: an improved database of intrinsically disordered and mobile proteins, *Nucleic Acids Res.*

[35] Romero, P., Obradovic, Z., Li, X., Garner, E., Brown, C. J., and Dunker, A. K. (2001) Sequence complexity of disordered protein, *Proteins 42*, 38-48.

[36] Linding, R., Jensen, L. J., Diella, F., Bork, P., Gibson, T. J., and Russell, R. B. (2003) Protein disorder prediction: implications for structural proteomics, *Structure 11*.

[37] Ward, J. J., Sodhi, J. S., McGuffin, L. J., Buxton, B. F., and Jones, D. T. (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life, *J. Mol. Biol. 337*, 635-645.

[38] Wang, S., Ma, J., and Xu, J. (2016) AUCpreD: proteome-level protein disorder prediction by AUC-maximized deep convolutional neural fields, *Bioinformatics 32*, i672-i679.

[39] Galzitskaya, O. V., Garbuzynskiy, S. O., and Lobanov, M. Y. (2006) FoldUnfold: web server for the prediction of disordered regions in protein chain, *Bioinformatics 22*, 2948-2957.

[40] Dosztányi, Z., Csizmók, V., Tompa, P., and Simon, I. (2005) IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content, *Bioinformatics 21*.

[41] Atkins, J. D., Boateng, S. Y., Sorensen, T., and McGuffin, L. J. (2015) Disorder Prediction Methods, Their Applicability to Different Protein Targets and Their Usefulness for Guiding Experimental Studies, *Int. J. Mol. Sci. 16*, 19040-19054.

[42] Oates, M. E., Romero, P., Ishida, T., Ghalwash, M., Mizianty, M. J., Xue, B., Dosztanyi, Z., Uversky, V. N., Obradovic, Z., Kurgan, L., Dunker, A. K., and Gough, J. (2013) D2P2: Database of Disordered Protein Predictions, *Nucleic Acids Res. 41*, D508-D516.

[43] Ishida, T., and Kinoshita, K. (2008) Prediction of disordered regions in proteins based on the meta approach., *Bioinformatics 24*, 1344-1348.

[44] Mizianty, M. J., Stach, W., Chen, K., Kedarisetti, K. D., Disfani, F. M., and Kurgan, L. (2010) Improved sequence-based prediction of disordered regions with multilayer fusion of multiple information sources., *Bioinformatics 26*, i489-i496.

[45] Dunker, A. K., Brown, C. J., Lawson, J. D., Iakoucheva, L. M., and Obradovic, Z. (2002) Intrinsic disorder and protein function, *Biochemistry 41*, 6573-6582.

[46] Dyson, H. J., and Wright, P. E. (2015) Intrinsically disordered proteins in cellular signalling and regulation, *Nature 16*.

[47] Mukhopadhyay, R., and Hoh, J. H. (2001) AFM force measurements on microtubule - associated proteins: the projection domain exerts a long-range repulsive force, *FEBS Lett. 505*, 374-378.

[48] Hernandez, M. A., Avila, J., and Andreu, J. M. (1986) Physicochemical characterization of the heat - stable microtubule - associated protein MAP2, *Eur. J. Biochem. 154*, 41-48.

[49] Woody, R. W., Roberts, G. C. K., Clark, D. C., and Bayley, P. M. (1982) 1H NMR evidence for flexibility in microtubule - associated proteins and microtubule protein oligomers, *FEBS Lett. 141*, 181-184.

[50] Ringel, I., and Sternlicht, H. (1984) Carbon-13 Nuclear Magnetic Resonance Study of Microtubule Protein: Evidence for a Second Colchichine Site in the Inhibition of Microtubule Assembly, *Biochemistry 23*, 5644-5653.

[51] Hoshi, T., Zagotta, W. N., and Aldrich, R. W. (1990) Biophysical and molecular mechanisms of Shaker potassium channel inactivation, *Science 250*, 533-538.

[52] Uversky, V. N., and Dunker, A. K. (2010) Understanding protein non-folding, *Biochim. Biophys. Acta. - Proteins and Proteomics 1804*, 1231-1264.

[53] Kriwacki, R. W., Hengst, L., Tennant, L., Reed, S. I., and Wright, P. E. (1996) Structural studies of p21 Waf1/Cip1/Sdi1 in the free and CdK2 - bound state: Conformational disorder mediates binding diversity., *Proc. Natl. Acad. Sci. USA 93*, 11504-11509.

[54] Russo, A. A., Jeffrey, P. D., Patten, A. K., Massague, J., and Pavletich, N. P. (1996) Crystal structure of the p27Kip1 cyclin-dependent-kinase inhibitor bound to the cyclin A - CdK2 complex, *Nature 382*, 325-331.

[55] Daughdrill, G. W., Chadsey, M. S., Karlinsey, J. E., Hughes, K. T., and Dahlquist, F. W. (1997) The C-terminal half of the anti-sigma factor, FlgM, becomes structured when bound to its target, sigma28., *Nature Structural Biology 4*, 285-291.

[56] Fitzgerald, D. M., Bonocora, R. P., and Wade, J. T. (2014) Comprehensive Mapping of the Escherichia coli Flagellar Regulatory Network, *PLOS Genetics 10*.

[57] Lu, Y., and Bennick, A. (1998) Interaction of tannin with human salivary proline - rich proteins, *Arch Oral Biol. 43*, 717-728.

[58] Lisse, T., Bartels, D., Kalbitzer, H. R., and Jaaenicke, R. (1996) The recombinant dehydrin - like desiccation stress protein from the resurrection plant Craterostigma plantagineum displays no defined three-dimensional structure in its native state., *Biol. Chem. 377*, 555-561.

[59] Pelka, P., Ablack, J. N. G., Fonseca, G. J., Yousef, A. F., and Mymryk, J. S. (2008) Intrinsic Structural Disorder in Adenovirus E1A: a Viral Molecular Hub Linking Multiple Diverse Processes, *J. Virol. 82*, 7252-7263.

[60] Ferreon, J. C., Martinez-Yamout, M., Dyson, H. J., and Wright, P. E. (2009) Structural basis for subversion of cellular control mechanisms by the adenoviral E1A oncoprotein, *Proc. Natl. Acad. Sci. USA 106*, 13260-13265.

[61] Xue, B., Romero, P. R., Noutsou, M., Maurice, M. M., Rüdiger, S. G. D., Jr., A. M. W., Marcin J. Mizianty, Kurgan, L., Uversky, V. N., and Dunker, A. K. (2013) Stochastic machines as a colocalization mechanism for scaffold protein function, *FEBS Lett. 587*, 1587-1591.

[62] Kim, T. D., Paik, S. R., and Yang, C. H. (2002) Structural and functional implication of C-terminal regions of alpha-synuclein, *Biochemistry 41*, 13782-13790.

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

[64] Chakrabortee, S., Boschetti, C., Walton, L. J., Sarkar, S., Rubinsztein, D. C., and Tunnacliffe, A. (2007) Hydrophilic protein associated with desiccation tolerance exhibits broad protein stabilization function, *Proc. Natl. Acad. Sci. USA 104*, 18073-18078.

[65] Kovacs, D., Kalmar, E., Torok, Z., and Tompa, P. (2008) Chaperone activity of ERD10 and ERD14, two disordered stress-related plant proteins, *Plant Physiol. 147*, 381-390.

[66] Pasta, S. Y., Raman, B., Ramakrishna, T., and Rao, C. (2002) Role of the C-terminal extensions of alpha-crystallins. Swapping the C-terminal extension of alpha-crystallin to alphaB-crystallin results in enhanced chaperone activity, *J. Biol. Chem. 277*, 45821-45828.

[67] Andley, U. P., Mathur, S., Griest, T. A., and Petrash, J. M. (1996) Cloning, expression, and chaperone-like activity of human alphaA-crystallin, *J. Biol. Chem. 271*, 31973-31980.

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

[69] Ivanyi-Nagy, R., Davidovic, L., Khandjian, E. W., and Darlix, J. L. (2005) Disordered RNA chaperone proteins: from functions to disease., *Cell Mol. Life Sci. 62*, 1409-1417.

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

[71] Iakoucheva, L. M., Radivojac, P., Brown, C. J., O'Connor, T. R., Sikes, J. G., Obradovic, Z., and Dunker, A. K. (2004) The importance of intrinsic disorder for protein phosphorylation, *Nucleic Acids Res. 32*, 1037-1049.

[72] Radivojac, P., Vacic, V., Haynes, C., Cocklin, R. R., Mohan, A., Heyen, J. W., Goebl, M. G., and Iakoucheva, L. M. (2010) Identification, analysis, and prediction of protein ubiquitination sites, *Proteins 78*, 365-380.

[73] Forman-Kay, J. D., and Mittag, T. (2013) From Sequence and Forces to Structure, Function, and Evolution of Intrinsically Disordered Proteins, *Cell*, 1492-1499.

[74] Meek, D. W., and Milne, D. M. (2000) Analysis of multisite phosphorylation of the p53 tumor-suppressor protein by tryptic phosphopeptide mapping., *Methods Mol. Biol. 99*, 447-463.

[75] Ferreon, J. C., Lee, C. W., Arai, M., Martinez-Yamout, M., Dyson, H. J., and Wright, P. E. (2009) Cooperative regulation of p53 by modulation of ternary complex formation with CBP/p300 and HDM2., *Proc. Natl. Acad. Sci. USA 106*.

[76] Lee, C. W., Ferreon, J. C., Ferreon, A. C. F., Arai, M., and Wright, P. E. (2010) Graded enhancement of p53 binding to CREB-binding protein (CBP) by multisite phosphorylation., *Proc. Natl. Acad. Sci. USA 107*, 19290-19295.

[77] Binolfi, A., Theilliet, F. X., and Selenko, P. (2012) Bacterial in-cell NMR of human alpha - synuclein: a disordered monomer by nature?, *Biochem. Soc. Trans. 40*, 950-954.

[78] Smith, A. E., Zhou, L. Z., and Pielak, G. J. (2015) Hydrogen exchange of disordered protein in Escherichia coli. , *Protein Sci. 24*, 706-713.

[79] Theillet, F. X., Binolfi, A., Frembgen-Kesner, T., Hingorani, K., Sarkar, M., Kyne, C., Li, C., Crowley, P. B., Gierasch, L., Pielak, G. J., Elcock, A. H., Gershenson, A., and Selenko, P. (2014) Physicochemical properties of cells and their effects on intrinsically disordered proteins (IDPs), *Chem. Rev. 114*, 6661-6714.

[80] Wright, P. E., and Dyson, H. J. (2009) Linking folding and binding, *Current Opinion in Structural Biology 19*, 8.

[81] Mohan, A., Oldfield, C. J., Radivojac, P., Vacic, V., Cortese, M. S., Dunker, A. K., and Uversky, V. N. (2006) Analysis of molecular recognition features (MoRFs), *J. Mol. Biol. 362*, 1043-1059.

[82] Tompa, P., Fuxreiter, M., Oldfield, C. J., Simon, I., Dunker, A. K., and Uversky, V. N. (2009) Close encounters of the third kind: disordered domains and the interactions of proteins, *Bioessays 31*, 328-335.

[83] Chen, J. W., Romero, P., Uversky, V. N., and Dunker, A. K. (2006) Conservation of intrinsic disorder in protein domains and families: II. functions of consered disorder., *J. Proteome Res. 5*, 888-898.

[84] Chen, J., Romero, P., Uversky, V. N., and Dunker, A. K. (2006) Conservation of intrinsic disorder in protein domains and families: I. A database of conserved predicted disordered regions., *J. Proteome Res. 5*, 879-887.

[85] Hsu, W. J., Oldfield, C. J., Xue, B., Meng, J., Huang, F., Romero, P., Uversky, V. N., and Dunker, A. K. (2013) Exploring the binding diversity of intrinsically disordered proteins involved in one - to - many binding., *Protein Sci. 22*, 258-273.

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

[87] Oldfield, C. J., Meng, J., Yang, J. Y., Yang, M. Q., Uversky, V. N., and Dunker, A. K. (2008) Flexible nets: disorder and induced fit in the associations of p53 and 14-3-3 with their partners, *BMC Genomics 9*, Suppl 1:S1.

[88] Oldfield, C. J., and Dunker, A. K. (2014) Intrinsically disordered proteins and intrinsically disordered protein regions, *Annu. Rev. Biochem. 84*, 553-584.

[89] Dosnon, M., Bonetti, D., Morrone, A., Erales, J., Silvio, E. d., Longhi, S., and Gianni, S. (2015) Demonstration of a Folding after Binding Mechanism in the Recognition between the Measles Virus N_{TAIL} and X Domains, *ACS Chem. Biol. 10*, 795-802.

[90] Humphrey, W., Dalke, A., and Schulten, K. (1996) VMD - Visual Molecular Dynamics, *J. Molec. Graphics 14*, 33-38.

[91] Arai, M., Sugase, K., Dyson, H. J., and Wright, P. E. (2015) Conformational propensities of intrinsically disordered proteins influence the mechanism of binding and folding., *PNAS 112*, 9614-9619.

[92] Zor, T., De Guzman, R. N., Dyson, H. J., and Wright, P. E. (2004) Solution structure of the KIX domain of CBP bound to the transactivation domain of c-Myb, *J. Mol. Biol. 337*, 521-534.

[93] Berlow, R. B., Dyson, H. J., and Wright, P. E. (2015) Functional advantages of dynamic protein disorder, *EBS Lett.*

[94] Espinoza-Fonseca, L. M. (2009) Reconciling binding mechanisms of intrinsically disordered proteins, *Biochem. Biophys. Res. Commun. 382*, 479-482.

[95] Shoemaker, B. A., Portman, J. J., and Wolynes, P. G. (2000) Speeding molecular recognition by using the folding funnel: The fly-casting mechanism, *Proc. Natl. Acad. Sci. USA 97*, 8868-8873.

[96] Yongqi Huang, and Liu, Z. (2009) Kinetic advantage of intrinsically disordered proteins in coupled folding - binding process: a critical assessment of the "fly-casting" mechanism, *J. Mol. Biol. 393*, 1143-1159.

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

[98] Sharma, R., Raduly, Z., Kiskei, M., and Fuxreiter, M. (2015) Fuzzy complexes: Specific binding without complete folding, *FEBS Lett. 589*, 2533-2542.

[99] Uversky, V. N., Li, J., Souillac, p., Millet, I. S., Doniach, S., Jakes, R., Goedert, M., and Fink, A. L. (2002) Biophysical properties of the synucleins and their propensities to fibrillate: inhibition of alpha-synuclein assembly by beta- and gamma-synucleins., *J. Biol. Chem. 2002*, 11970-11978.

[100] Dunker, A. K., and Oldfield, C. J. (2014) Chapter 1. Back to the Future: Nuclear Magnetic Resonance and Bioinformatics Studies on Intrinsically Disordered Proteins, In *Intrinsically Disordered Proteins Studied by NMR Spectroscopy* (Felli, I. C., and Pierattelli, R., Eds.), pp 1-34, Springer.

[101] Dodero, V. I., Quirolo, Z. B., and Sequeira, M. A. (2011) Biomolecular studies by circular dichroism., *Frontiers in Bioscience 16*, 61-73.

[102] Shojania, S. (2007) Nuclear Magnetic Resonance and Dynamic Characterization of the Intrinsically Disordered HIV-1 Tat Protein.

[103] Doniach, S. (2001) Changes in Biomolecular Conformation Seen by Small Angle X-ray Scattering, *Chem. Rev. 101*, 1763-1778.

[104] Sibille, N., and Pernado, P. (2012) Structural characterization of intrinsically disordered proteins by the combined use of NMR and SAXS, *Biochem. Soc. Trans. 40*.

[105] Kikhney, A. G., and Svergun, D. I. (2015) A practical guide to small angle X-ray scattering (SAXS) of flexible and intrinsically disordered proteins, *FEBS Lett. 589*, 2570-2577.

[106] Van Der Meer, B. W., Coker, G. I., and Chen, S. Y. S. (1994) *Resonance energy transfer*, Wiley-VCH:Berlin.

[107] Ferreon, A. C., Gambin, Y., Lemke, E. A., and Deniz, A. A. (2009) Interplay of α-synuclein binding and conformational switching probeb by single-molecule fluorescence, *Proc. Natl. Acad. Sci. USA 106*, 5645-5650.

[108] Trexler, A. J., and Rhoades, E. (2013) Function and dysfunction of α-Synuclein: Probing conformational changes and aggregation by single molecule fluorescence., *Mol. Neurobiol. 47*, 622-631.

[109] Brucale, M., Schuler, B., and Samori, B. (2014) Single-molecule studies of intrinsically disordered proteins, *Chem. Rev. 114*, 3281-3317.

[110] Barre-Sinoussi, F., Chermann, J. C., Rey, F., Nugeyre, M. T., Chamaret, S., Gruest, J., Dauguet, C., Axler-Blin, C., Vezinet-Brun, F., Rouzioux, C., Rozenbaum, W., and Montagnier, L. (1983) Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS), *Science 220*, 868-871.

[111] UNAIDS. (2016) Global HIV Statistics.

[112] T., C.-Y. O., and T., G. R. (2011) Update on human immunodeficiency virus (HIV)-2 infection, *Clin. Infect. Dis. 52*, 780-787.

[113] Ekouevi, D. K., Tchounga, B. K., Coffie, P. A., Tegbe, J., Anderson, A. M., Gottlieb, G. S., Vitoria, M., Dabis, F., and Eholie, S. P. (2014) Antiretroviral therapy response among HIV-2 infected patients: a systematic review., *BMC Infect. Dis. 14*.

[114] Maartens, G., Celum, C., and Lewin, S. R. (2014) HIV infection: epidemiology, pathogenesis, treament, and prevention, *Lancet 384*, 258-271.

[115] Moore, M. D., and Hu, W. S. (2009) HIV-1 RNA dimerization: It takes two to tango., *AIDS Rev. 11*, 91-102.

[116] Wain-Hobson, S., Sonigo, P., Danos, O., and al., e. (1985) Nucleotide sequence of the AIDS virus, LAV, *Cell 40*, 9-17.

[117] Ratner, L., Haseltine, W., Patarca, R., and al., e. (1985) Complete nucleotide sequence of the AIDS virus, HTLV-III, *Nature 313*, 277-284.

[118] https://www.hiv.lanl.gov. (2016) HIV sequence database.

[119] Garzon, M. T., Lidon-Moya, M. C., Barrera, F. N., Prieto, A., Gomez, J., Mateu, M. G., and Neira, J. L. (2004) The dimerization domain of the HIV-1 capsid protein binds a capsid protein-derived peptide: abiophysical characterization, *Protein Sci. 13*, 1512-1523.

[120] Yan, N., Regalado-Magdos, A. D., Stiggelbout, B., Lee-Kirsch, M. A., and Lieberman, J. (2010) The cytosolic exonuclease TREX1 inhibits the innate immune response to human immunodeficiency virus type 1, *Nat. Immunol. 11*, 1005-1013.

[121] Rasaiyaah, T., Tan, C. P., Fletcher, A. J., Price, A. J., and Blondeau, C. (2013) HIV-1 evades innate immune recognition through specfic cofactor recruitment, *Nature 503*, 402-405.

[122] Lahaye, X., Satoh, T., Gentili, M., Cerboni, S., Conrad, C., and al., e. (2013) The capsids of HIV-1 and HIV-2 determine immune detection of the viral cDNA by the innate sensor cGAS in dendritic cells., *Immunity 39*, 1132-1142.

[123] (2008) HIV Sequence Compendium 2008, (Kuiken, C., Leitner, T., Foley, B., Hahn, B., Marx, P., McCutchan, F., Wolinsky, S., and Korber, B., Eds.).

[124] Salgado, G. F., Vogel, A., Marquant, R., Feller, S. E., Bouaziz, S., and Alves, I. D. (2009) The role of membranes in the organization of HIV-1 Gag p6 and Vpr: p6 shows high affinity

for membrane bilayers which substantially increases the interaction between p6 and Vpr., *J. Med. Chem. 52*, 7157-7162.

[125] Zhu, P., Liu, J., Bess, J. J., Chertova, E., Lifson, J. D., Grisé, H., Ofek, G. A., Taylor, K. A., and Roux, K. H. (2006) Distribution and three-dimensional structure of AIDS virus envelope spikes., *Nature 441*, 847-852.

[126] Burton, D. R. (2006) Structural biology: Images from the surface of HIV, *Nature 441*, 817-818.

[127] Gentile, M., Adrian, T., Scheidler, A., Ewald, M., Dianzani, F., Pauli, G., and Gelderblom, H. R. (1994) Determination of the size of HIV using adenovirus type 2 as an internal length marker, *J. Virol. Methods. 48*, 43-52.

[128] Konstantinov, I., Stefanov, Y., Kovalevsky, A., Voronin, Y., and Company, V. S. (2011) Human Immunodeficiency Virus 3D, *Science 331*, 848-849.

[129] Chan, D. C., and Kim, P. S. (1998) HIV Entry and Its Inhibition, *Cell 93*, 681-684.

[130] Craigie, R. (2001) HIV Intergrase, a brief overview from chemistry to therapeutics, *J. Biol. Chem. 276*, 23213-23216.

[131] Barre-Sinoussi, F., Ross, A. L., and Delfraissy, J.-F. (2013) Past, present and future: 30 years of HIV research, *Nature Reviews Microbiology 11*, 877-883.

[132] Cantin, R., Fortin, J. F., Lamontagne, G., and Tremblay, M. (1997) The presence of host-derived HLA-DR1 on human immunodeficiency virus type 1 increases viral infectivity., *J. Virol. 71*, 1922-1930.

[133] Paquette, J. S., Fortin, J. F., Blanchard, L., and Tremblay, M. J. (1998) Level of ICAM-1 surface expression on virus producer cells influences both the amount of virion-bound

host ICAM-1 and human immunodeficiency virus type 1 infectivity., *J. Virol. 72*, 9329-9336.

[134] Saifuddin, M., Hedayati, T., Atkinson, J. P., Holquin, M. H., Parker, C. J., and Spear, G. T. (1997) Human immunodeficiency virus type 1 incorporates both glycosyl phosphatidylinositol-anchored CD55 and CD59 and integral membrane CD46 at levels that protect from complement-mediated destruction., *J. Gen. Virol. 78*, 1907-1911.

[135] Fischl, M. A., Richman, D. D., Grieco, M. H., Gottlieb, M. S., Volberding, P. A., Laskin, O. L., Leedom, J. M., Groppman, J. E., Mildvan, D., Schooley, R. T., and al., e. (1987) The efficacy of azidothymidine (AZT) in the treament of patients with AIDS and AIDS-related complex. A double-blind, placebo-controlled trial., *N. Engl. J. Med. 317*, 185-191.

[136] Mitsuya, H., Weinhold, K. J., Furman, P. A., Clair, M. H. S., Lehrman, S. N., Gallo, R. C., Bolognesi, D., Barry, D. W., and Broder, S. (1985) 3'-Azido-3'-deoxythymidine (BW A509U): an antiviral agent that inhibits the infectivity and cytopathic effect of human T-lymphotropic virus type III/lymphadenopathy-associated virus in vitro., *Proc. Natl. Acad. Sci. USA 82*, 7096-7100.

[137] Hammer, S. M., Squires, K. E., Hughes, M. D., Grimes, J. M., Demeter, L. M., Currier, J. S., Eron, J. J. J., Feinberg, J. E., Balfour, H. H. J., Deyton, L. R., Chodakewitz, J. A., and Fischl, M. A. (1997) A controlled trial of two nucleoside analogues plus indinavir in persons with human immunodeficiency virus infection and CD4 cell counts of 200 per cubic millimeter or less. AIDS Clinical Trials Group 320 Study Team., *N. Engl. J. Med. 337*, 725-733.

[138] Palella, F. J. J., Delaney, K. M., Moorman, A. C., Loveless, M. O., Fuhrer, J., Saltten, G. A., Aschman, D. J., and Holmberg, S. D. (1998) Declining morbidity and mortality among

patients with advanced human immunodeficiency virus infection. HIV Outpatient Study Investigators., *N. Engl. J. Med. 338*, 853-860.

[139] Kao, S.-Y., Calman, A. F., Luciw, P. A., and Peterlin, B. M. (1987) Anti-termination of transcription within the long terminal repeat of HIV-1 by tat gene product, *Nature 330*, 489-493.

[140] Li, L., Li, H. S., Pauza, D., Bukrinsky, M., and Zhao, R. Y. (2005) Roles of HIV-1 auxiliary proteins in viral pathogenesis and host-pathogen interactions, *Cell Research 15*, 923-934.

[141] Romani, B., Engelbrecht, S., and Glashoff, R. H. (2010) Functions of Tat: the versatile protein of human immunodeficiency virus type 1, *Journal of General Virology 91*, 1-12.

[142] Liang, C., and Wainberg, M. A. (2002) The role of Tat in HIV-1 replication: an activator and/or a suppressor?, *AIDS Rev. 4*, 41-49.

[143] Ott, M., Schnolzer, M., Garnica, J., Fischle, W., Emiliani, S., and al., e. (1999) Acetylation of the HIV-1 Tat protein by p300 is important for its transcriptional activity, *Curr. Biol. 9*, 1489-1492.

[144] D’Orso, I., and Frankel, A. D. (2009) Tat acetylation modulates assembly of a viral-host RNA-protein transcriptional complex, *Proc. Natl. Acad. Sci. USA 106*, 3101-3106.

[145] Dingwall, C., Ernberg, I., J.Gait, M., and M.Green, S. (1990) HIV-1 tat protein stimulates transcription by binding to a U-rich bulge in the stem of the TAR RNA structure, *The EMBO Journal 9*, 4145-4153.

[146] Cordingley, M. G., L.LaFemina, R., Callahan, P. L., Condra, J. H., Sardana, V. V., Graham, D. J., Nguyen, T. M., LeGrow, K., Gotlib, L., Schlabach, A. J., and al., e. (1990) Sequence-

specific interaction of Tat protein and Tat peptides with the transactivation-responsive sequence element of human immunodeficiency virus type 1 in vitro, *Proc. Natl. Acad. Sci. USA 87*, 8985-8989.

[147] Fujinaga, K., Irwin, D., Huang, Y., and Taube, R. (2004) Dynamics of Human Immunodeficiency Virus ranscription: P-TEFb Phosphorylates RD and Dissociates Negative Effectors from the Transactivation Response Element, *Mol. Cell. Biol 24*, 787-795.

[148] D'Orso, I., Jang, G. M., Pastuszak, A. W., Faust, T. B., Quezada, E., Booth, D. S., and Frankela, A. D. (2012) Transition Step during Assembly of HIV Tat:P-TEFb Transcription Complexes and Transfer to TAR RNA, *Molecular and Cellular Biology 32*, 4780-4793.

[149] Easley, R., Duyne, R. V., Coley, W., Guendel, I., Dadgar, S., Kehn-Hall, K., and Kashanchi, F. (2010) Chromatin dynamics associated with HIV-1 Tat activated transcription, *Biochim. Biophys. Acta. 1799*, 275-285.

[150] Kiernan, R. E., Vanhulle, C., Schiltz, L., Adam, E., Xiao, H., Maudoux, F., Calomme, C., Burny, A., Benkirane, M., and Lint, C. V. (1999) HIV-1 Tat transcriptional activity is regulated by acetylation, *The EMBO Journal 18*, 6106-6118.

[151] Mushinova, Y. R., Sheval, E. V., Dib, C., Germini, D., and Vassetzky, Y. S. (2016) Functional roles of HIV-1 Tat protein in the nucleus, *Cell. Mol. Life Sci. 73*.

[152] Mediouni, S., Darque, A., Baillat, G., and al., e. (2012) Antiretroviral therapy does not block the secretion of human immunodeficiency virus tat protein, *Infect. Disord. Drug. Targets. 12*, 81-86.

[153] Ferrari, A., Pellegrini, V., Arcangeli, C., and al., e. (2003) Caveolae-mediated internalization of extracellular HIV-1 tat fusion proteins visualized in real time., *Mol. Ther. 8*, 284-294.

[154] Debaisieux, S. n., Rayne, F., Yezid, H., and Beaumelle, B. (2012) The Ins and Outs of HIV-1 Tat, *Traffic 13*, 10.

[155] De Falco, G., Bellan, C., Lazzi, S., Claudio, P., La Sala, D., Cinti, C., Tosi, P., Giordano, A., and Leoncini, L. (2003) Interaction between HIV-1 Tat and pRb2/p130: a possible mechanism in the pathogenesis of AIDS-related neoplasms., *Oncogene 22*, 6214-6219.

[156] Prakash, O., Tang, Z. Y., He, Y. E., Ali, M. S., Coleman, R., Gill, J., Farr, G., and Samaniego, F. (2000) Human Kaposi's sarcoma cell-mediated tumorigenesis in human immunodeficiency type 1 tat-expressing transgenic mice., *J. Nalt. Cancer Inst. 92*, 721-728.

[157] Nunnari, G., Smith, J. A., and Daniel, R. (2008) HIV-1 Tat and AIDS-associated cancer: targeting the cellular anti-cancer barrier?, *J. Exp. Clin. Cancer Res. 27*.

[158] Hudson, L., Liu, J., Nath, A., Jones, M., Raghavan, R., Narayan, O., Male, D., and Everall, I. (2000) Detection of the human immunodeficiency virus regulatory protein tat in CNS tissues, *J. Neurovirol 6*, 145-155.

[159] Maubert, M. E., Pirrone, V., Rivera, N. T., Wigdahl, B., and Nonnemacher, M. R. (2015) Interaction between Tat and Drugs of Abuse during HIV-1 infection and central nervous system disease., *Front Microbiol. 6*.

[160] Shojania, S., and O’Neil, J. D. (2006) 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*, 8347–8356.

[161] Li, L., Dahiya, S., Kortagere, S., Aiamkitsumrit, B., Cunningham, D., Pirrone, V., Nonnemacher, M. R., and BrianWigdahl. (2012) Impact of Tat Genetic Variation on HIV-1 Disease, *Advances in Virology 2012*, 28.

[162] Jeang, K.-T., Xiao, H., and Rich, E. A. (1999) Multifaceted Activities of the HIV-1 Transactivator of Transcription, Tat, *J. Biol. Chem. 274*, 28837-28840.

[163] Strazza, M., Pirrone, V., BrianWigdahl, and Nonnemacher, M. R. (2011) Breaking down the barrier: The effects of HIV-1 on the blood-brain barrier, *Brain Research 1399*, 96-115.

[164] Huigen, M. C. D. G., Kamp, W., and Nottet, H. S. L. M. (2004) Multiple effects of HIV-1 trans-activator protein on the pathogenesis of HIV-1 infection, *Eur. J. Clin. Invest 34*, 57-66.

[165] Tahirov, T. H., Babayeva, N. D., Varzavand, K., Cooper, J. J., Sedore, S. C., and Price, D. H. (2010) Crystal structure of HIV-1 Tat complexed with human P-TEFb, *Nature 465*, 747-753.

[166] Campbell, G. R., Pasquier, E., Watkins, J., and Bourgarel-Rey, V. (2004) The Glutamine-rich Region of the HIV-1 Tat Protein Is Involved in T-cell Apoptosis, *J. Biol. Chem. 279*, 47197-48204.

[167] Hocine Yezid, Karidia Konate, Sole`ne Debaisieux, Anne Bonhoure, and Beaumelle, B. (2009) Mechanism for HIV-1 Tat Insertion into the Endosome Membrane, *J. Biol. Chem. 284*, 22736-22746.

[168] López-Huertas, M. R., Mateos, E., Cojo, M. S. d., Gómez-Esquer, F., and Díaz-Gil, G. (2013) The Presence of HIV-1 Tat Protein Second Exon Delays Fas Protein-mediated Apoptosis in CD4+ T Lymphocytes, *J. Biol. Chem. 288*, 7626-7644.

[169] Barillari, G., Gendelman, R., Gallo, R. C., and Ensoli, B. (1993) The Tat protein of human immunodeficiency virus type 1, a growth factor for AIDS Kaposi sarcoma and cytokine-activated vascular cells, induces adhesion of the same cell types by using integrin receptors recognizing the RGD amino acid sequence, *PNAS 90*, 7941-7945.

[170] Neuveut, C., Scoggins, R. M., Camerini, D., Markham, R. B., and Jeang, K.-T. (2003) Requirement for the Second Coding Exon of Tat in the Optimal Replication of Macrophage-Tropic HIV-1, *J. Biomed. Sci*, 651-660.

[171] Lopez-Huertas, M. R., Callejas, S., Abia, D., Mateos, E., Dopazo, A., Alcami, J., and Coiras, M. (2010) Modifications in host cell cytoskeleton structure and function mediated by intracellular HIV-1 Tat protein are greatly dependent on the second coding exon, *Nucleic Acids Research 37*, 3287-3307.

[172] Mahlknecht, U., Dichamp, I., Varin, A., Lint, C. V., and Herbein, G. (2007) NF-kB-dependent control of HIV-1 transcription by the second coding exon of Tat in T cells, *J. Leukocyte Biol. 83*, 718-727.

[173] Smith, S. M., Pentlicky, S., Klase, Z., Singh, M., Neuveut, C., Lu, C.-y., Marvin S. Reitz, J., Yarchoan, R., Marx, P. A., and Jeang, K.-T. (2003) An in Vivo Replication-important Function in the Second Coding Exon of Tat Is Constrained against Mutation despite Cytotoxic T Lymphocyte Selection, *Journal of Biological Chemistry 278*, 44816-44825.

[174] Pincus, S. H., Messer, K. G., Nara, P. L., Blattner, W. A., Colclough, G., and Reitz, M. (1994) Temporal Analysis of the Antibody Response to HIV Envelope Protein in HIV-infected Laboratory Workers, *Journal of of Clinical Investigation, Inc. 93*, 2505-2513.

[175] Manish K. Johri, Mishra, R., Chhatbar, C., Salini, K. U., and Singh, S. K. (2011) Tits and bits of HIV Tat protein, *Expert Opin. Bio. Ther. 2011*, 269-283.

[176] Jäger, S., Cimermancic, P., Gulbahce, N., Johnson, J. R., McGovern, K. E., Clarke, S. C., Shales, M., Mercenne, G., Pache, L., Li, K., Hernandez, H., Jang, G. M., Roth, S. L., Akiva, E., Marlett, J., Stephens, M., D'Orso, I., Fernandes, J., Fahey, M., Mahon, C., O'Donoghue, A. J., Todorovic, A., Morris, J. H., Maltby, D. A., Alber, T., Cagney, G., Bushman, F. D., Young, J. A., Chanda, S. K., Sundquist, W. I., Kortemme, T., Hernandez, R. D., Craik, C. S., Burlingame, A., Sali, A., Frankel, A. D., and Krogan, N. J. (2012) Global landscape of HIV–human protein complexes, *Nature 481*, 365–370.

[177] Serrière, J., Dugua, J.-M., Bossus, M., Verrier, B., Haser, R., Gouet, P., and Guillon, C. (2011) Fab′-Induced Folding of Antigenic N-Terminal Peptides from Intrinsically Disordered HIV-1 Tat Revealed by X-ray Crystallography, *Journal of Molecular Biology 405*, 33-42.

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

[179] Nicholas Lyle, Das, R. K., and Pappu, R. V. (2013) A quantitative measure for proteini conformational heterogeneity, *J. Chem. Phys. 139*.

[180] Henzler-Wildman, K., and Kern, D. (2007) Dynamic personalities of proteins, *Nature 450*, 964 - 972.

[181] Karplus, M., and Petsko, G. A. (1990) Molecular dynamics simulations in biology., *Nature 347*, 631-639.

[182] Dobson, C. M. (2003) Protein folding and misfolding, *Nature 426*, 884-890.

[183] Smock, R. G., and Gierasch, L. M. (2009) Sending signals dynamically, *Science 324*, 198-203.

[184] Watson, H. C. (1969) The stereochemistry of the protein Myoglobin, *Prog. Stereochem. 4*, 299.

[185] Baldwin, A. J., and Kay, L. E. (2009) NMR spectroscopy brings invisible protein states into focus, *Nature Chemical Biology 5*, 808 - 814.

[186] Mulder, F. A. A., Mittermaier, A., Hon, B., Dahlquist, F. W., and Kay, L. E. (2001) Studying excited states of proteins by NMR spectroscopy, *Nature*.

[187] Mittag, T., Schaffhausen, B., and Gunther, U. L. (2003) Direct observation of protein-ligand interaction kinetics, *Biochemistry 42*, 11128-11136.

[188] Tang, C., Iwahara, J., and Clore, G. M. (2006) Visualization of transient encounter complexes in protein-protein association., *Nature 444*, 383-386.

[189] Iwahara, J., and Clore, G. M. (2006) Detecting transient intermediates in macromolecular binding by paramagnetic NMR, *Nature 440*, 1227-1230.

[190] Sugase, K., Lansing, J. C., Dyson, H. J., and Wright, P. E. (2007) Tailaoring relaxation dispersion experiments for fast-associating protein complexes, *J. AM. CHEM. SOC. 129*, 13406-13407.

[191] Sugase, K., Dyson, H. J., and Wright, P. E. (2009) Mechanism of coupled folding and binding of an intrinsically disordered protein, *Nature 447*, 1021-1025.

[192] Korzhnev, D. M., Bezsonova, I., Lee, S., Chalikian, T. V., and Kay, L. E. (2009) Alternate binding modes for a ubiquitin-SH3 domain interaction studied by NMR spectroscopy, *J. Mol. Biol. 386*, 391-405.

[193] Cole, R., and Loria, J. P. (2002) Evidence for flexibility in the function of ribonuclease A. , *Biochemistry 41*, 6072-6081.

[194] Watt, E. D., Shimada, H., Kovrigin, E. L., and Loria, J. P. (2007) The mechanism of rate-limiting motions in enzyme function, *PNAS 104*, 11981-11986.

[195] Vallurupalli, P., and Kay, L. E. (2006) Complementarity of ensemble and single-molecule measures of protein motion: a relaxation dispersion NMR study of an enzyme complex, *PNAS 103*, 11910-11915.

[196] Korzhnev, D. M., Religa, T. L., Lundstrom, P., Fersht, A. R., and Kay, L. E. (2007) The folding pathway of an FF domain: characterization of an on-pathway intermediate state under folding conditions by (15)N, (13)C(alpha) and (13)C-methyl relaxation dispersion and (1)H/(2)H-exchange NMR spectroscopy, *J. Mol. Biol. 372*, 497-512.

[197] Tang, Y., Gery, M. J., McKnight, J., Palmer, A. G. I., and Raleigh, D. P. (2006) Mulistate folding of the villin headpiece domain, *J. Mol. Biol. 355*, 958-976.

[198] Choy, W. Y., Zhou, Z., Bai, Y., and Kay, L. E. (2005) An 15N NMR spin relaxation dispersion study of the folding of a pair of engineered mutants of apocytochrome b562, *J. AM. CHEM. SOC. 127*, 5066-5072.

[199] S., F. J., van den Bedem, H., Samelson, A. J., Lang, P. T., Holton, J. M., Echols, N., and Alber, T. (2011) Accessing protein conformational ensembles using room-temperature X-ray crystallography, *PNAS 108*, 16247-16252.

[200] Leopold, P. E., Montal, M., and Onuchic, J. N. (1992) Protein folding funnels: A kinetic approach to the sequence-structure relationship., *PNAS 89*, 8721-8725.

[201] Onuchic, J. N., and Wolynes, P. G. (2004) Theory of protein folding., *Curr. Opin. Struct. Biol. 14*, 70-75.

[202] Webb, P. A., Perisic, O., Mendola, C. E., Backer, J. M., and Williams, R. L. (1995) The crystal structure of a human nucleoside diphosphate kinase, *J. Mol. Biol. 251*, 574-587.

[203] Drobnak, I., De Jonge, N., Haesaerts, S., Vesnaver, G., Loris, R., and Lah, J. (2013) Energetic basis of uncoupling folding from binding for an intrinsically disordered protein, *J. AM. CHEM. SOC. 135*, 1288-1294.

[204] Burger, V. M., Gurry, T., and Stultz, C. M. (2014) Intrinsically disodered proteins: Where computation meets experiment, *Polymers 6*, 2684-2719.

[205] Metzger, A. U., Bayer, P., Willbold, D., Hoffmann, S., Frank, R. W., Goody, R. S., and Rösch, P. (1997) The interaction of HIV-1 Tat(32-72) with its target RNA: a fluorescence and nuclear magnetic resonance study., *Biochem. Biophys. Res. Commun. 241*, 31-36.

[206] Mujtaba, S., He, Y., Zeng, L., Farooq, A., Carlson, J. E., Ott, M., Verdin, E., and Zhou, M. M. (2002) Structural basis of lysine-acetylated HIV-1 Tat recognition by PCAF bromodomain., *Mol. Cell 9*, 575-586.

[207] Purcell, E. M., Torrey, H. C., and Pound, R. V. (1946) Resonance absoption by nuclear magnetic moments in sold, *Phys. Rev. 69*, 37-38.

[208] Bloch, F. (1946) Nuclear Induction, *Phys. Rev. 70*, 460-474.

[209] Proctor, W. G., and Yu, F. C. (1950) The dependence of a nuclear magnetic resonance frequency upon chemical compound, *Phys. Rev. 77*, 717-717.

[210] Overhauser, A. W. (1953) Polarization of nuclei in metals, *Phys. Rev. 92*, 411-415.

[211] Nelson, F. A., and Weaver, H. E. (1964) Nuclear magnetic resonance spectroscopy in super-con-ducting magnetic fields, *Science 146*, 223-232.

[212] Ernst, R. R., and Anderson, W. A. (1966) Application of Fourier transformed spectroscopy to magnetic resonance., *Rev. Sci. Instrum. 37*, 93-102.

[213] Aue, W. P., Bartholdi, E., and Ernst, R. R. (1976) Two-dimensional spectroscopy. Application to nuclear magnetic resonance., *J. Chem. Phys. 64*, 2229-2246.

[214] Williamson, M. P., Havel, T. F., and Wuthrich, K. (1985) Solution conformation of proteinase inhibitor IIA from bull seminal plasma by ^{1}H nuclear magnetic resonance and distance geometry, *J. Mol. Biol. 182*, 295-315.

[215] Bhattacharya, A. (2010) Breaking the billion-hertz barrier, *Nature 463*, 605-606.

[216] Kenjiro Hashi, Shinobu Ohki, Matsumoto, S., Nishijima, G., Goto, A., Kenzo Deguchi, Kazuhiko Yamada, Takashi Noguchi, Shuji Sakai, Masato Takahashi, Yoshinori Yanagisawa, Seiya Iguchi, Toshio Yamazaki, Hideaki Maeda, Ryoji Tanaka, Takahiro Nemoto, Hiroto Suematsu, Takashi Miki, Kazuyoshi Saito, and Shimizu, T. (2015) Achivement of 1020 MHz NMR, *Journal of Magnetic Resonance 256*.

[217] Pervushin, K., Riek, R., Wider, G., and Wüthrich, K. (1997) Attenuated T2 relaxation by mutual cancellation of dipole-dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of very large biological macromolecules in solution., *PNAS 94*, 12366-12371.

[218] Kay, L. E. (2016) New views of functionally dynamic proteins by solution NMR spectroscopy, *J. Mol. Biol. 428*, 323-331.

[219] Ortega, G., Pons, M., and Millet, O. Protein Functional Dynamics in Multiple Timescales as Studied by NMR Spectroscopy, In *Advances in Protein Chemistry and Structural Biology*.

[220] Lopes, P. E. M., Guvench, O., and Alexander D. MacKerell, J. (2015) Current status of protein force fields for molecular dynamics simulations., *Methods Mol. Biol. 1215*, 47-71.

[221] Lipari, G., and Szabo, A. (1982) Model-free approach to the interpretation of nuclear magnetic-resonance relaxation in macromolecules. 2. Analysis of experimental results., *J. AM. CHEM. SOC. 104*, 4559-4570.

[222] Lippari, G., and Szabo, A. (1982) Model-Free Approach to the Interpretation of Nuclear Magnetic Resonance Relaxation in Macromolecules. 1. Theory and Range of Validity, *J. AM. CHEM. SOC. 104*, 4546-4559.

[223] Camilloni, C., Sahakyan, A. B., Holliday, M. J., JIsern, N. G., Zhang, F. L., Eisenmeisser, E. Z., and Vendruscolo, M. (2014) Cyclophilin A catalyzes proline isomerization by an electrostatic handle mechanism, *PNAS 111*, 10203-10208.

[224] Dyson, H. J., and Wright, P. E. (2004) Unfolded proteins and protein folding studied by NMR, *Chem. Rev. 104*, 3607-3622.

[225] Ellman, J. A., Volkman, B. F., Mendel, D., Schultz, P. G., and Wemmer, D. E. (1992) Site-specific isotopic labeling of proteins for NMR studies., *J. AM. CHEM. SOC. 114*, 7959-7961.

[226] Krishnarjuna, B., Jaipuria, G., Thakur, A., D'Silva, P., and Atreya, H. S. (2011) Amino acid selective unlabeling for sequence specific resonance assignments in proteins, *J. Biomol. NMR 49*, 39-51.

[227] Schubert, M., Labudde, D., Leitner, D., Oschikinat, H., and Schmieder, P. (2005) A modified strategy for sequence specific assignment of protein NMR spectra based on amino acid type selective experiments., *J. Biomol. NMR 31*, 115-128.

[228] Schubert, M., Oschkinat, H., and Schmieder, P. (2001) MUSIC and aromatic residues: Amino acid type-selective 1H-15N correlations, part III., *Journal of Magnetic Resonance 153*, 186-192.

[229] Schubert, M., Oschkinat, H., and Schmieder, P. (2001) Amino acid type-selective 1H-15N correlations for Arg and Lys., *20*, 379-384.

[230] Schubert, M., Oschkinat, H., and Schmieder, P. (2001) MUSIC, selective pulses and tuned delays: Amino acid type-selective 1H-15N correlations, part II., *Journal of Magnetic Resonance 148*, 61-72.

[231] Schubert, M., Oschkinat, H., and Schmieder, P. (1999) MUSIC in triple resonance experiments: Amino acid type selective 1H-15N correlations., *Journal of Magnetic Resonance 141*, 34-43.

[232] Mobli, M., and Hoch, J. C. (2014) Nonuniform sampling and non-Fourier signal processing methods in multidimensional NMR, *Progress in Nuclear Magnetic Resonance Spectroscopy 83*, 21-41.

[233] Mobli, M., Maciejewski, M. W., Schuyler, A. D., Stern, A. S., and Hoch, J. C. (2012) Sparse sampling methods in multidimensional NMR, *Phys. Chem. Chem. Phys. 14*, 10835-10843.

[234] Kupce, E., and Freeman, R. (2004) Projection-reconstruction techquie for speeding up multidimensional NMR spectroscopy, *J. AM. CHEM. SOC. 126*, 6429-6440.

[235] Hiller, S., Fiorito, F., Wuthrich, K., and Wider, G. (2005) Automated projection spectroscopy, *Proc. Natl. Acad. Sci. USA 102*, 10876-10881.

[236] Kupce, E. (2015) NMR with multiple receivers, *Top Curr. Chem. 335*, 71-96.

[237] Zawadzka-Kazimierczuk, A., Kozminski, W., Sanderova, H., and Krasny, L. (2012) High dimensional and high resolution pulse sequences for backbone resonance assignment of intrinsically disordered proteins., *J. Biomol. NMR 52*, 329-337.

[238] Zerko, S., and Kozminski, W. (2015) Six- and seven-dimensional experiments by combination of sparse random sampling and projection spectroscopy dedicated for backbone resonance assignment of intrinsically disordered proteins, *J. Biomol. NMR 63*, 283-290.

[239] Dziekanski, P., Grudziaz, K., Jarvoll, P., Kozminski, W., and Zawadzka-Kazimierczuk, A. (2015) ^{13}C-detected NMR experiments for automatic resonance assignment of IDPs and multiple-fixing SMFT processing, *J. Biomol. NMR 62*, 179-190.

[240] Dunker, A. K., Oldfield, C. J., Meng, J., Romero, P., Yang, J. Y., Chen, J. W., Vacic, V., Obradovic, Z., and Uversky, V. N. (2008) The unfoldomics decade: an update on intrinsically disordered proteins, *BMC Genomics 9*, Suppl 2():S1.

[241] Brutscher, B., Felli, I. C., Gil-Caballero, S., Hosek, T., Kummerle, R., Piai, A., Pierattelli, R., and Solyom, Z. (2015) NMR Methods for the Study of Instrinsically Disordered Proteins Structure, Dynamics, and Interactions: General Overview and Practical Guidelines, In *Intrinsically disordered proteins studied by NMR spectroscopy* (Felli, I. C., and Pierattelli, R., Eds.), pp 49-122, Springer International Publishing.

[242] Rule, G. S., and Hitchens, T. K. (2006) *Fundalmentals of Protein NMR Spectroscopy*, Springer, The Neitherlands.

[243] Wishart, D. S., Sykes, B. D., and Richards, F. M. (1991) The Chemical Shift Index: A Fast and Simple Method for the Assignment of Protein Secondary Structure through NMR Spectroscopy, *Biochemistry 31*, 1647-1651.

[244] Mielke, S. P., and Krishnan, V. V. (2009) Characterization of protein secondary structure from NMR chemical shifts, *Progress in Nuclear Magnetic Resonance Spectroscopy 54*, 141-165.

[245] Wishart, D. S., and Sykes, B. D. (1994) The 13C Chemical-Shift Index: A simple method for the identification of protein secondary structure using 13C chemical-shift data, *J Biomol NMR 4*, 171-180.

[246] Dios, A. C. d., Pearson, J. G., and Oldfield, E. (1993) Secondary and tertiary structural effects on protein NMR chemical shifts: an *ab initio* approach., *Science 260*, 1491-1496.

[247] Kjaergaard, M., Brander, S., and Poulsen, F. M. (2011) Random coil chemical shift for intrinsically disordered proteins: effects of temperature and pH., *J. Biomol. NMR 49*, 139-149.

[248] Wishart, D. S., Bigam, C. G., Holm, A., Hodges, R. S., and Sykes, B. D. (1995) ^{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*, 67-81.

[249] Schwarzinger, S., Kroon, G. J., Foss, T. R., Wright, P. E., and Dyson, H. J. (2000) Random coil chemical shifts in acidic 8 M urea: implementation of random coil shift data in NMRView., *J. Biomol. NMR 18*, 43-48.

[250] Braun, D., Wider, G., and Wuthrich, K. (1994) Sequence-corrected 15N random coil chemical-shifts, *J. AM. CHEM. SOC. 116*, 8466-8469.

[251] Bienkiewicz, E. A., and Lumb, K. J. (1999) Random-coil chemical shifts of phosphorylated amino acids., *J. Biomol. NMR 15*, 203-206.

[252] Schwarzinger, S., Kroon, G. J., Foss, T. R., Chung, J., Wright, P. E., and Dyson, H. J. (2001) Sequence-dependent correction of random coil NMR chemical shifts, *J. AM. CHEM. SOC. 123*, 2970-2978.

[253] M. Kjaergaard, F. M. P. (2011) Sequence correction of random coil chemical shifts: correlation between neighbor correction factors and changes in the Ramachandran distribution., *J. Biomol. NMR 50*, 157-165.

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

[255] Simone, A. D., Cavalli, A., Hsu, S. T., Vranken, W., and Vendruscolo, M. (2009) Accurate random coil chemical shifts from an analysis of loop regions in native states of proteins, *J. AM. CHEM. SOC. 131*, 16332-16333.

[256] Tamiola, K., Acar, B., and Mulder, F. A. A. (2010) Sequence-Specific Random Coil Chemical Shifts of Intrinsically Disordered Proteins, *J. AM. CHEM. SOC. 132*, 18000-18003.

[257] Wang, L., Eghbalnia, H. R., and Markley, J. L. (2006) Probabilistic approach to determining unbiased random-coil carbon-13 chemical shift values from the protein chemical shift database., *J. Biomol. NMR 35*, 155-165.

[258] Ulrich, E. L., Akutsu, H., Doreleijers, J. F., Harano, Y., Ioannidis, Y. E., Lin, J., Livny, M., Mading, S., Maziuk, D., Miller, Z., Nakatani, E., Schulte, C. F., Tolmie, D. E., Wenger, R. K., Yao, H., and Markley, J. L. (2008) BioMagResBank, *Nucleic Acids Res. 36*, D402-D408.

[259] Ting, D., Wang, G., Shapovalov, M., Mitra, R., Jordan, M. I., and Jr., R. L. D. (2010) Neighbor-dependent Ramachandran probability distributions of amino acids developed from a hierarchical Dirichlet process model., *PLos Comput. Biol. 6*.

[260] MacArthur, M. W., and Thornton, J. M. (1991) Influence of proline residues on protein conformation, *J. Mol. Biol. 218*, 397-412.

[261] Wang, Y., and Jardetzky, O. (2002) Probability-based protein secondary structure identification using combined NMR chemical-shift data., *Protein Sci. 11*, 852-861.

[262] Mulder, F. A. A., and Filatov, M. (2010) NMR chemical shift data and ab initio shielding calculations: emerging tools for protein structure determination, *Chem. Soc. Rev. 39*, 578-590.

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

[264] Marsh, J. A., Signh, V. K., Jia, Z., and Forman-Kay, J. D. (2006) Sensitivity of secondary structure propensities to sequence differences between alpha- and gama-synuclein: Implications for fibrillation, *Protein Science 15*, 2795-2804.

[265] Karplus, M. (1959) Contact electron - spin coupling of nuclear magnetic moments, *J. Chem. Phys. 30*, 11-15.

[266] Case, D. A. (2000) Interpretation of chemical shifts and coupling constants in macromolecules, *Curr. Opin. Struct. Biol. 10*, 197-203.

[267] Gapsys, V., Narayanan, R. L., Xiang, S., Groot, B. L. d., and Zweckstetter, M. (2015) Improved validation of IDP ensembles by one-bond Ca–Ha scalar couplings, *J. Biomol. NMR 63*, 299-307.

[268] Morin, S. (2011) A practical guide to protein dynamics from 15N spin relaxation in solution, *Progress in Nuclear Magnetic Resonance Spectroscopy 59*, 18.

[269] Kay, L. E., Torchia, D. A., and Bax, A. (1989) Backbone dynamics of proteins as studied by ^{15}N inverse detected heteronuclear NMR spectroscopy: Application to Staphylococcal Nuclease, *Biochemistry 28*, 8972-8979.

[270] Morris, G. A., and Freeman, R. (1979) Enhancement of nuclear magnetic resonance signals by polarization transfer., *Journal of the American Chemical Society 101*, 760-762.

[271] Pinheiro, A. S., Marsh, J. A., Forman-Kay, J. D., and Peti, W. (2010) Structural Signature of the MYPT1-PP1 Interaction, *J. AM. CHEM. SOC. 133*.

[272] Kay, L. E., Torchia, D. A., and Bax, A. (1989) Backbone dynamics of proteins as studied by 15N inverse detected heteronuclear NMR spectroscopy: application to staphylococcal nuclease., *Biochemistry 28*, 8972-8979.

[273] Farrow, N. A., Zhang, O., Szabo, A., Torchia, D. A., and Kay, L. E. (1995) Spectral density function mapping using 15N relaxation data exclusively, *Journal of Biomolecular NMR 6*, 153-162.

[274] Ishima, R., and Nagayama, K. (1995) Protein Backbone Dynamics Revealed by Quasi Spectral Density Function Analysis of Amide N-15 Nuclei, *Biochemistry 34*, 3162-3171.

[275] Clore, G. M., Szabo, A., Bax, A., Kay, L. E., Driscoll, P. C., and Gronenborn, A. M. (1990) Deviations from the Simple Two-Parameter Model-Free Approach to the Interpretation of Nitrogen-15 Nuclear Magnetic Relaxation of Proteins, *J. AM. CHEM. SOC. 112*, 4989-4991.

[276] d'Auvergne, E. J., and Gooley, P. R. (2008) Optimisation of NMR dynamic models I. Minimisation algorithms and their performance within the model-free and Brownian rotational diffusion spaces, *J Biomol NMR 40*, 107-119.

[277] d'Auvergne, E. J., and Gooley, P. R. (2008) Optimisation of NMR dynamic models II. A new methodology for the dual optimisation of the model-free parameters and the Brownian rotational diffusion tensor, *J Biomol NMR 40*, 121-133.

[278] d'Auvergne, E. J. (2006) Protein dynamics: A study of the model-free analysis of NMR relaxation data, In *Russell Grimwade Department of Biochemistry and Molecular Biology*, p 379, Melbourne.

[279] Kempf, J. G., and Loria, J. P. (2003) Protein dynamics from solution NMR: theory and applications, *Cell Biochem. Biophys. 37*, 187-211.

[280] Palmer, A. G. I. (1997) Probing molecular motion by NMR., *Curr. Opin. Struct. Biol. 7*, 732-737.

[281] Bieri, M., d'Auvergne, E. J., and Gooley, P. R. (2011) relaxGUI: a new software for fast and simple NMR relaxation data analysis and calculation of ps-ns and μs motion of proteins., *J Biomol NMR 50*, 147-155.

[282] Khan, S. N., Charlier, C., Augustyniak, R., Salvi, N., Dejean, V., Bodenhausen, G., Lequin, O., Pelupessy, P., and Ferrage, F. (2015) Distribution of Pico and Nanosecond Motions in Disordered Proteins from Nuclear Spin Relaxation, *Biophysical Journal 109*, 988-999.

[283] Prompers, J. J., and Bruschweiler, R. (2002) General framework for studying the dynamics of folded and nonfolded proteins by NMR relaxation spectroscopy and MD simulation., *J. AM. CHEM. SOC. 124*, 4522-4534.

[284] Modig, K., and Poulsen, F. M. (2008) Model-independent interpretation of NMR relaxation data for unfolded proteins: the acid-denatured state of ACBP, *J. Biomol. NMR 42*, 163-177.

[285] Buevich, A. V., and Baum, J. (1999) Dynamics of Unfolded Proteins: Incorporation of Distributions of Correlation Times in the Model Free Analysis of NMR Relaxation Data, *J. Am. Chem. Soc. 121*, 2.

[286] Buevich, A. V., Shinde, U. P., Inouye, M., and Baum, J. (2001) Backbone dynamics of the natively unfolded pro-peptide of subtilisin by heteronuclear NMR relaxation studies, *J Biomol NMR 20*, 233-249.

[287] Peng, J. W., and Wagner, G. (1992) Mapping of Spectral Density Functions Using Heteronuclear NMR Relaxation Measurements, *J. Magn. Reson. 98*, 308-332.

[288] Peng, J. W., and Wagner, G. (1992) Mapping of the spectral densities of nitrogen-hydrogen bond motions in Eglin c using heteronuclear relaxation experiments, *Biochemistry 31*, 8571-8586.

[289] Farrow, N. A., Zang, O., Forman-Kay, J. D., and Kay, L. E. (1995) Comparison of the backbone dynamics of a folded and an unfolded SH3 domain existing in equilibrium in aqueous buffer., *Biochemistry 34*, 868-878.

[290] Krızova, H., Žıdek, L., Stoneb, M. J., Novotny, M. V., and Sklenar, V. (2003) Temperature-dependent spectral density analysis applied to monitoring backbone dynamics of major urinary protein-I complexed with the pheromone 2-sec-butyl-4,5-dihydrothiazole, *Journal of Biomolecular NMR*.

[291] Korzhnev, D. M., and Kay, L. E. (2007) Probing invisible, low-populated States of protein molecules by relaxation dispersion NMR spectroscopy: an application to protein folding., *ACCOUNTS OF CHEMICAL RESEARCH 41*, 442-451.

[292] III, A. G. P., Kroenke, C. D., and Loria, J. P. (2001) Nuclear Magnetic Resonance Methods for Quantifying Microsecond-to-Millisecond Motions in Biological Macromolecules, *Methods in Enzymology 339*, 204 - 238.

[293] Millet, O., Loria, J. P., Kroenke, C. D., Pons, M., and III, A. G. P. (2000) The static magnetic field dependence of chemical exchange linebroadening defines the NMR chemical shift time scale., *J. AM. CHEM. SOC. 122*, 2867-2877.

[294] Ishima, R. (2011) Recent Developments in 15N NMR Relaxation Studies that Probe Protein Backbone Dynamics, *Top Curr. Chem. 326*, 99-122.

[295] Korzhnev, D. M., Salvatella, X., Vendruscolo, M., Nardo, A. A. D., Davidson, A. R., Dobson, C. M., and Kay, L. E. (2004) Low-populated folding intermediates of Fyn SH3 characterized by relaxation dispersion NMR *Nature*.

[296] Carver, J. P., and Richards, R. E. (1972) General 2-site solution for chemical exchange produced depenence of t2 upon carr-purcell pulse separation., *J. Magn. Reson. 6*, 89-105.

[297] Ishima, R., and Torchia, D. A. (1999) Estimating the time scale of chemical exchange of proteins from measurements of transverse relaxation rates in solution, *J. Biomol. NMR 14*, 369-372.

[298] Hvidt, A., Johansen, G., Linderstrøm-Lang, K., and Vaslow, F. (1955) Exchange of deuterium and^{18}O between water and other substances. I. Methods., *C. R. Trav. Lab. Carlsberg Chim. 29*, 129-157.

[299] Hvidt, A., and Linderstrøm-Lang, K. (1955) Exchange of deuterium and ^{18}O between water and other substances. III. Deuterium exchange ofshort peptides, Sanger's A-chain and insulin. , *C. R. Trav. Lab. Carlsberg Chim. 29*, 385-402.

[300] Skinner, J. J., Lim, W. K., Bedard, S., Black, B. E., and Englander, S. W. (2012) Protein hydrogen exchange: Testing current models, *Protein Sci. 21*, 987-995.

[301] Englander, S. W., Mayne, L., Bai, Y., and Sosnick, T. R. (1996) Hydrogen exchange:The modem legacy of Linderstrom-Lang, *Protein Science 6*.

[302] Perrin, C. L., and Lollo, C. P. (1983) Mechanisms of NH Proton Exchange in Amides and Proteins: Solvent Effects and Solvent Accessibility, *J. AM. CHEM. SOC. 106*.

[303] Perrin, C. L. (1989) Proton exchange in amides: Surprises from simple systems, *Acc. Chem. Res. 22*, 268-275.

[304] Molday, R. S., Englander, S. W., and Kallen, R. G. (1972) Primary structure effects on peptide group hydrogen exchange, *Biochemistry 11*, 150-158.

[305] Bai, Y., Milne, J. S., Mayne, L., and Englander, S. W. (1993) Primary structure effects on peptide group hydrogen exchange, *Proteins 17*, 75-86.

[306] Linderstrøm-Lang, K., and Schellman, J. A. (1959) *Protein structure and enzyme activity.*, New York: Academic Press.

[307] Linderstrøm-Lang, K. (1958) *Deuterium exchange and protein structure*, London: Methuen.

[308] Hvidt, A., and Nielsen, S. O. (1966) Hydrogen exchange in proteins, *Adv. Protein Chem. 21*, 287-386.

[309] Skinner, J. J., Lim, W. K., Bedard, S., Black, B. E., and Englander, S. W. (2012) Protein dynamics viewed by hydrogen exchange, *Protein Sci. 21*, 996-1005.

[310] Connelly, G. P., Bai, Y., Jeng, M.-F., and Englander, S. W. (1993) Isotope Effects in Peptide Group Hydrogen Exchange, *PROTEINS: Structure, Function, and Genetics 17*, 87-92.

[311] Li, A., and Daggett, V. (1995) Investigation of the solution structure of chymotrypsin inhibitor 2 using molecular dynamics: Comparison to x-ray crystallographic and NMR data, *Protein Eng. 8*, 1117-1128.

[312] Sheinerman, F. B., and Brooks, C. L. (1998) Molecular picture of folding of a small α/β protein., *Proc. Natl. Acad. Sci. USA 95*, 1562-1567.

[313] Garcia, A. E., and Hummer, G. (1999) Conformational dynamics of cytochrome c: Correlation to hydrogen exchange., *Proteins 36*, 175-191.

[314] Petrunk, A. A., and al., e. (2013) Molecular dynamics simulations provide atomistic insight into hydrogen exchange mass spectrometry experiments., *J. Chem. Theory Comput. 9*, 658-669.

[315] Shan, Y., Arkhipov, A., Kim, E. T., Pan, A. C., and Shaw, D. E. (2013) Transitions to catalytically inactive conformations in EGFR kinase., *Proc. Natl. Acad. Sci. USA 110*, 7270-7275.

[316] Kossiakoff, A. A. (1982) Protein dynamics investigated by the neutron diffraction-hydrogen exchange technique., *Nature 296*, 713-721.

[317] Levitt, M. (1981) Molecular dynamics of hydrogen bonds in bovine pancreatic trypsin inhibitor protein. , *Nature 294*, 379-380.

[318] Milne, J. S., Mayne, L., Roder, H., Wand, A. J., and Englander, S. W. (1998) Determinants of protein hydrogen exchange studied in equine cytochrome c, *Protein Sci. 7*, 739-745.

[319] Bahar, I., Wallqvist, A., Covell, D. G., and Jernigan, R. L. (1998) Correlation between native-state hydrogen exchange and cooperative residue fluctuations from a simple model, *Biochemistry 37*, 1067-1075.

[320] Hernandez, G., Anderson, J. S., and LeMaster, D. M. (2012) Experimentally assessing molecular dynamics sampling of the protein native state conformational distribution., *Biophys. Chem. 163-164*, 21-34.

[321] Tompa, P. (2010) *Structure and Function of Intrinsically Disordered Protein*, Taylor and Francis Group, United States of America.

[322] Li, M., and Song, J. (2007) The N- and C-termini of the human Nogo molecules are intrinsically unstructured: bioinformatics, CD, NMR characterization, and functional implications., *Proteins 68*, 100-108.

[323] Ritter, C., Maddelein, M. L., Siemer, A. B., Luhrs, T., Ernst, M., Meier, B. H., Saupe, S. J., and Riek, R. (2005) Correlation of structural elements and infectivity of the HET-s prion., *Nature 435*, 844-848.

[324] Csizmok, V., Felli, I. C., Tompa, P., Banci, L., and Bertini, I. (2008) Structural and dynamic characterization of intrinsically disordered human securin by NMR spectroscopy., *J. AM. CHEM. SOC. 130*, 16873-16879.

[325] Thapar, R., Mueller, G. A., and Marzluff, W. F. (2004) The N-terminal domain of the Drosophila histone mRNA binding protein, SLBP, is intrinsically disordered with nascent helical structure., *Biochemistry 43*, 9390-9400.

[326] Wagner, G., and Wuthrich, K. (1982) Amide proton exchange and surface conformation of the basic pancreatic trypsin inhibitor in solution: studies with two-dimensional nuclear magnetic resonance., *J. Mol. Biol. 160*, 343-361.

[327] Gal, M., Schanda, P., Brutscher, B., and Frydman, L. (2007) UltraSOFAST HMQC NMR and the Repetitive Acquisition of 2D Protein Spectra at Hz Rates, *J. AM. CHEM. SOC. 129*, 1372-1377.

[328] Roder, H., Elöve, G. A., and Englander, S. W. (1988) Structural characterization of folding intermediates in cytochrome c by H-exchange labelling and proton NMR., *Nature 335*, 700-704.

[329] Hwang, T.-L., Zijl, P. C. M. v., and Mori, S. (1997) Accurate quantitation of water–amide proton exchange rates using the Phase-Modulated CLEAN chemical EXchange (CLEANEX-PM) approach with a Fast-HSQC (FHSQC) detection scheme, *Journal of Biomolecular NMR 11*.

[330] Chevelkov, V., Xue, Y., Rao, D. K., Forman-Kay, J. D., and Skrynnikov, N. R. (2009) 15N H/D-SOLEXSY experiment for accurate measurement of amide

solvent exchange rates: application to denatured drkN SH3, *J Biomol NMR 46*, 18.

[331] Hwang, T.-L., Mori, S., Shaka, A. J., and Zijl, P. C. M. v. (1997) Application of Phase-Modulated CLEAN Chemical EXchange Spectroscopy (CLEANEX-PM) to Detect Water-Protein Proton Exchange and Intermolecular NOEs, *J. Am. Chem. Soc. 119*.

[332] Shang, Z. G., Swapna, G. V. T., Rios, C. B., and Montelione, G. T. (1997) Sensitivity enhancement of triple-resonance protein NMR spectra by proton evolution of multiple-quantum coherences using a simultaneous ^{1}H and ^{13}C constant-time evolution period. , *J. AM. CHEM. SOC. 119*, 9274-9278.

[333] Hitchens, T. K., McCallum, S. A., and Rule, G. S. (1999) A J^{CH}-modulated 2D (HACACO)NH pulse scheme for quantitative measurement of ^{13}Cα-^{1}Hα couplings in ^{15}N, ^{13}C-labeled proteins., *J. Magn. Reson. 140*, 281-284.

[334] Hansen, P. E. (2000) Isotope effects on chemical shifts of proteins and peptides., *Magn. Reson. Chem. 38*, 1-10.

[335] McConnell, H. M. (1958) Reaction rates by nuclear magnetic resonance., *J. Chem. Phys. 28*, 430-431.

[336] Jeener, J., Meier, B. H., Bachmann, P., and Ernst, R. R. (1979) Investigation of exchange processes by 2-dimensional NMR spectroscopy., *J. Chem. Phys. 71*, 4546-4553.

[337] McCammon, J. A., Gelin, B. R., and Karplus, M. (1977) Dynamics of folded proteins, *Nature 267*, 585-590.

[338] Eckhardt, W., Alexander Heinecke, Bader, R., Brehm, M., Hammer, N., Huber, H., Kleinhenz, H.-G., Jadran Vrabec, Hasse, H., Horsch, M., Bernreuther, M., Glass, C. W., Niethammer, C., Bode, A., and Bungartz, H.-J. (2013) *Supercomputing*, Springer Berlin Heidelberg, Berlin.

[339] Brooks, B. R., and al., e. (2009) CHARMM: The biomolecular simulation program, *J. Comput. Chem. 30*, 1545-1614.

[340] Plimpton, S. (1995) Fast Parallel Algorithms for Short-Range Molecular Dynamics, *J. Comp. Phys. 117*, 1-19.

[341] Phillips, J. C., and al., e. (2005) Scalable molecular dynamics with NAMD, *J. Comput. Chem. 26*, 1781-1802.

[342] Eastman, P., and *al., e.* (2013) OpenMM 4: A reusable, extensible, hardware independent library for high performance molecular simulation., *J. Chem. Theory Comput. 9*, 461-469.

[343] Abraham, M. J., Spoel, D. v. d., Lindahl, E., Hess, B., and team, a. t. G. d. (2014) GROMACS User Manual version 5.0.4.

[344] Born, M., and Oppenheimer, R. (1927) "Zur Quantentheorie der Molekeln" [On the Quantum Theory of Molecules], *Annalen der Physik 389*, 457-484.

[345] Leach, A. R. (2001) *Molecular Modelling - Principles and applications*, 2 ed., Pearson Education Limited, Harlow, Essex, England.

[346] Henriques, J. (2016) Modeling and simulation of intrinsically disordered proteins, In *Department of Chemistry*, p 70, Lund University, Lund, Sweden.

[347] Allen, M. P., and Tildesley, D. J. (1991) *Computer simulation of liquids*, Oxford University Press, Great Britain.

[348] Frenkel, D., and Smit, B. (2002) *Understanding Molecular Simulation From Algorithms to Applications*, Academic Press.

[349] Verlet, L. (1967) Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard−Jones Molecules, *Phys. Rev. 159*, 98-103.

[350] Berendsen, H. J. C., Postma, J. P. M., Gunsteren, W. F., and Hermans, J. (1981) *Interaction models for water in relation to protein hydration*, Reidel, Dordrecht, Holland.

[351] Berendsen, H. J. C., Grigera, J. R., and Straatsma, T. P. (1987) The missing term in effective pair potentials, *J. Phys. Chem. 91*, 6269-6271.

[352] Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., and Klein, M. L. (1983) Comparison of simple potential function for simulating liquid water, *J. Chem. Phys. 79*, 926-935.

[353] Jorgensen, W. L., and Madura, J. D. (1985) Temperature and size dependence for monte carlo simulations of TIP4P water., *Mol. Phys. 56*, 1381-1392.

[354] Abascal, J. L. F., and Vega, C. (2005) A general purpose model for the condensed phases of water: TIP4P/2005., *J. Chem. Phys. 123*, 234505.

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

[356] Henriques, J., and Skepö, M. (2016) Molecular Dynamics Simulations of Intrinsically Disordered Proteins: On the Accuracy of the TIP4P-D Water Model and the Representativeness of Protein Disorder Models., *J. Chem. Theory Comput. 12*, 3407-3415.

[357] Piana, S., Lindorff-Larsen, K., and Shaw, D. E. (2011) How Robust Are Protein Folding Simulations with Respect to Force Field Parameterization?, *Biophys. J. 100*, L47-L49.

[358] Lopes, P. E. M., Guvench, O., and Alexander D. MacKerell, J. (2015) Current status of protein force fields for molecular dynamic., *Methods Mol. Biol. 1215*, 47-71.

[359] Guvench, O., and Alexander D. MacKerell, J. (2008) Comparison of protein force fields for molecular dynamics simulations, *Methods Mol. Biol. 443*, 63-88.

[360] Wang, W., Ye, W., Jiang, C., Luo, R., and Chen, H. F. (2014) New force field on modeling intrinsically disordered proteins, *Chem. Biol. Drug Des. 84*, 253-269.

[361] Song, D., Wang, W., Ye, W., Ji, D., Luo, R., and Chen, H.-F. (2017) ff14IDPs force field improving the conformation sampling of intrinsically disordered proteins., *Chem. Biol. Drug Des. 89*, 5-15.

[362] Huang, J., Rauscher, S., Nawrocki, G., Ran, T., Feig, M., Groot, B. L. d., Grubmüller, H., and Jr, A. D. M. (2017) CHARMM36m: an improved force field for folded and intrinsically disordered proteins., *Nature Methods 14*, 71-73.

[363] Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A., and Haak, J. R. (1984) Molecular-Dynamics with Coupling to an External Bath, *J. Chem. Phys. 81*, 3684-3690.

[364] Kutzner, C., Páll, S., Fechner, M., Esztermann, A., Groot, B. L. d., and Grubmüller, H. (2015) Best bang for your buck: GPU nodes for GROMACS biomolecular simulations, *J. Comput. Chem. 36*, 1990-2008.

[365] Nakajima, N., Nakamura, H., and Kidera, A. (1997) Multicanonical Ensemble Generated by Molecular Dynamics Simulation for Enhanced Conformational Sampling of Peptides, *J. Phys. Chem. B 101*, 817-824.

[366] Ikebe, J., Umezawa, K., Kamiya, N., Sugihara, T., Yonezawa, Y., Takano, Y., Nakamura, H., and Higo, J. (2011) Theory for Trivial Trajectory Parallelization of Multicanonical Molecular Dynamics and Application to a Polypeptide in Water, *J. Comput. Chem. 32*, 1286-1297.

[367] Bernardi, R. C., Melo, M. C. R., and Schulten, K. (2015) Enhanced Sampling Techniques in Molecular Dynamics Simulations of Biological Systems, *Biochim. Biophys. Acta. 1580*, 872-877.

[368] Frankel, A. D., and Pabo, C. O. (1988) Cellular uptake of the tat protein from human immunodeficiency virus., *Cell 55*, 1189-1193.

[369] Tong, K. I., Yamamoto, M., and Tanaka, T. (2008) A simple method for amino acid selective isotope labeling of recombinant proteins in E. coli, *J Biomol NMR 42*, 9.

[370] Hiroaki, H., Umetsu, Y., Nabeshima, Y.-i., Hoshi, M., and Kohda, D. (2011) A simplified recipe for assigning amide NMR signals using combinatorial 14N amino acid inverse-labeling, *J. Struct. Funct. Genomics 12*, 167-174.

[371] Gold, V., and Lowe, B. M. (1967) Measurement of solvent isotope effects with glass electrode. I. The ionic product of D2O and D2O-H2O mixtures., *J. Chem. Soc.*, A:936-943.

[372] Covington, A. K., Paabo, M., Robinson, R. A., and Bates, R. G. (1968) Use of glass electrode in deuterium oxide and relation between standardized pD (pa$_d$) scale and operational pH in heavy water., *Anal. Chem. 40*, 700-706.

[373] Ikura, M., Kay, L. E., and Bax, A. (1990) A novel approach for sequential assignment of ^{1}H, ^{13}C, and ^{15}N spectra of proteins: heteronuclear triple-resonance three-dimensional NMR spectroscopy. Application to calmodulin., *Biochemistry 29*, 4659-4667.

[374] Kay, L. E., Keifer, P., and Saarinen, T. (1992) Pure absorption gradient enhanced heteronuclear single quantum correlation spectroscopy with improved sensitivity., *J. AM. CHEM. SOC. 114*, 10663-10665.

[375] Yamazaki, T., Lee, W., Arrowsmith, C. H., Muhandiram, D. R., and Kay, L. E. (1994) A Suite of Triple Resonance NMR Experiments for the Backbone Assignment of 15N, 13C, 2H Labeled Proteins with High Sensitivity, *J. AM. CHEM. SOC. 116*, 11655-11666.

[376] Grzesiek, S., and Bax, A. (1992) Correlating backbone amide and side chain resonances in larger proteins by multiple relayed triple resonance NMR, *J. AM. CHEM. SOC. 114*, 6291-6293.

[377] Shaka, A. J., Keller, J., Frenkiel, T., and Freeman, R. (1983) An improved sequence for broadband decoupling: WALTZ-16, *J. Magn. Reson. 52*, 335-338.

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

[379] Goddard, T. D., and Kneller, D. G. SPARKY 3, University of California, San Francisco.

[380] Lee, W., Tonelli, M., and Markley, J. L. (2015) NMRFAM-SPARKY: enhanced software for biomolecular NMR spectroscopy, *Bioinformatics 31*, 1325-1327.

[381] Vuister, G. W., and Bax, A. (1993) Quantitative J Correlation: A New Approach for Measuring Homonuclear Three-Bond J(HNHA) Coupling Constants in 15N-Enriched Proteins, *J. AM. CHEM. SOC. 115*, 7772-7777.

[382] Spyracopoulos, L. (2006) A suite of Mathematica notebooks for the analysis of protein main chain 15N NMR relaxation data, *J Biomol NMR 36*, 215-224.

[383] Neudecker, P., Lundstrom, P., and Kay, L. E. (2009) Relaxation Dispersion NMR Spectroscopy as a Tool for Detailed Studies of Protein Folding, *Biophysical Journal 96*, 2045–2054.

[384] Ishima, R. CPMG Relaxation Dispersion, In *Protein Dynamics: Methods and Protocols* (Livesay, D. R., Ed.).

[385] Bieri, M., and Gooley, P. R. (2011) Automated NMR relaxation dispersion data analysis using NESSY, *BMC Bioinformatics 12*.

[386] Hanwell, M. D., Curtis, D. E., Lonie, D. C., Vandermeersch, T., Zurek, E., and Hutchison, G. R. (2012) Avogadro: An advanced semantic chemical editor, visualization, and analysis platform., *Journal of Cheminformatics 4*.

[387] Ewald, P. (1921) Die Berechnung optischer und elektrostatischer Gitterpotentiale, *Ann. Phys. 369*, 253-287.

[388] Muchmore, D. C., McIntosh, L. P., Russell, C. B., Anderson, D. E., and Dahlquist, F. W. (1989) Expression and Nitrogen-15 Labeling of Proteins for Proton and Nitrogen-15 Nuclear Magnetic Resonance *Methods in Enzymology 177*.

[389] Schwarzinger, S., Kroon, G. J. A., Foss, T. R., Chung, J., Wright, P. E., and Dyson, H. J. (2001) Sequence-Dependent Correction of Random Coil NMR Chemical Shifts, *J. AM. CHEM. SOC. 123*, 2970-2978.

[390] Uversky, V. N. (2009) Intrinsically Disordered Proteins and Their Environment: Effects of Strong Denaturants, Temperature, pH, Counter Ions, Membranes, Binding Partners, Osmolytes, and Macromolecular Crowding, *The Protein Journal 28*, 305-325.

[391] Yang, Y., and Igumenova, T. I. (2013) The C-Terminal V5 Domain of Protein Kinase Cα Is Intrinsically Disordered, with Propensity to Associate with a Membrane Mimetic, *PLOS ONE 8*, e65699.

[392] Hughes, S., and Graether, S. P. (2011) Cryoprotective mechanism of a small intrinsically disordered dehydrin protein, *Protein Science 20*, 42-50.

[393] Ágoston, B. S., Kovács, D. n., Tompa, P. t., and Perczel, A. s. (2011) Full backbone assignment and dynamics of the intrinsically disordered dehydrin ERD14, *Biomol NMR Assign 5*, 189-193.

[394] Anand, K., Schulte, A., Vogel-Bachmayr, K., Scheffzek, K., and Geyer, M. (2008) Structural insights into the Cyclin T1–Tat–TAR RNA transcription activation complex from EIAV, *Nature Structural & Molecular Biology 15*, 1287-1292.

[395] Obradovic, Z., Peng, K., Vucetic, S., Radivojac, P., Brown, C., and Dunker, A. K. (2003) Predicting intrinsic disorder from amino acid sequence., *Proteins 53*, 566 - 572.

[396] Peng, K., Vucetic, S., Radivojac, P., Brown, C. J., Dunker, A. K., and Obradovic, Z. (2005) Optimizing Intrinsic Disorder Predictors with Protein Evolutionary Information., *Journal of Bioinformatics and Computational Biology 3*, 35-60.

[397] Peng, K., Radivojac, P., Vucetic, S., Dunker, A. K., and Obradovic, Z. (2006) Length-Dependent Prediction of Protein Intrinsic Disorder, *BMC Bioinformatics 7*, 208.

[398] Obradovic, Z., Peng, K., Vucetic, S., Radivojac, P., and Dunker, A. K. (2005) Exploiting Heterogeneous Sequence Properties Improves Prediction of Protein Disorder, *Proteins 61*, 176-182.

[399] Dosztányi, Z., Csizmók, V., Tompa, P., and Simon, I. (2005) The Pairwise Energy Content Estimated from Amino Acid Composition Discriminates between Folded and Intrinsically Unstructured Proteins *J. Mol. Biol. 347*, 827-739.

[400] T., I., and K., K. (2007) PrDOS: prediction of disordered protein regions from amino acid sequence., *Nucleic Acids Res. 35*.

[401] Macraild, C. A., Zachrdla, M., Andrew, D., Krishnarjuna, B., and al., e. (2015) Conformational Dynamics and Antigenicity in the Disordered Malaria Antigen Merozoite Surface Protein 2, *PLOS ONE*.

[402] Dayie, K. T., Wagner, G., and Lefevre, J. F. (1996) Theory and practice of nuclear spin relaxation in proteins, *Annu. Rev. Phys. Chem. 47*, 243-282.

[403] Lappalainen, I., Hurley, M. G., and Clarke, J. (2008) Plasticity Within the Obligatory Folding Nucleus of an Immunoglobulin-like Domain, *J. Mol. Biol. 375*, 547-559.

[404] Schurr, J. M., Babcock, H. P., and Fujimoto, B. S. (1994) A test of the model-free formulas. Effects of anisotropic rotational diffusion and dimerization., *J. Magn. Reson. B 105*, 211-224.

[405] Kosol, S., Contreras-Martos, S., Cedeño, C., and Tompa, P. (2013) Structural Characterization of Intrinsically Disordered Proteins by NMR Spectroscopy, *Molecules 18*, 10802-10828.

[406] Schneider, R., Maurin, D., Communie, G., Kragelj, J., Hansen, D. F., Ruigrok, R. W. H., Jensen, M. R., and Blackledge, M. (2015) Visualizing the Molecular Recognition Trajectory of an Intrinsically Disordered Protein Using Multinuclear Relaxation Dispersion NMR., *J. AM. CHEM. SOC. 137*, 1220-1229.

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

[408] Jones, D. P., and Go, Y. M. (2011) Mapping the cysteine proteome: analysis of redox-sensing thiols., *Curr. Opin. Chem. Biol. 15*, 103-112.

[409] Klepeis, J. L., Lindorff-Larsen, K., Dror, R. O., and Shaw, D. E. (2009) Long-timescale molecular dynamics simulations of protein structure and function, *Curr. Opin. Struct. Biol. 19*, 120-127.

[410] Ostermeir, K., and Zacharias, M. (2013) Advanced replica-exchange sampling to study the flexibility and plasticity of peptides and proteins., *Biochim. Biophys. Acta. 1834*, 847-853.

[411] Smilgies, D. M., and Folta-Stogniew, E. (2015) Molecular weight–gyration radius relation of globular proteins: a comparison of light scattering, small-angle X-ray scattering and structure-based data., *J. Appl. Crystallogr. 48*, 1604-1606.

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

[413] Touw, W. G., Baakman, C., Black, J., Beek, T. A. H. t., Krieger, E., Joosten, R. P., and Vriend, G. (2015) A series of PDB related databases for everyday needs., *Nucleic Acids Res. 43* D365-D368.

[414] Rogers, J. M., Steward, A., and Clarke, J. (2012) Folding and Binding of an Intrinsically Disordered Protein: Fast, but Not 'Diffusion-Limited', *J. Am. Chem. Soc. 135*, 8.

[415] Bienkiewicz, E. A., Adkins, J. N., and Lumb, K. J. (2002) Functional Consequences of Preorganized Helical Structure in the Intrinsically Disordered Cell-Cycle Inhibitor p27Kip1, *Biochemistry 41*, 752-759.

[416] ChenYu, Niu, X., Jin, F., Liu, Z., Jin, C., and Lai, L. (2016) Structure-based Inhibitor Design for the Intrinsically Disordered Protein c-Myc, *Scientific Reports 6*, 22298.

[417] Ambadipudi, S., and Zweckstetter, M. (2016) Targeting intrinsically disordered proteins in rational drug discovery., *Expert Opin. Drug Discov. 11*, 65-77.

[418] Schulze-Gahmen, U., Lu, H., Zhou, Q., and Alber, T. (2014) AFF4 binding to Tat-P-TEFb indirectly stimulates TAR recognition of super elongation complexes at the HIV promoter, *eLife*.

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

[420] (2012) *Intrinsically Disordered Protein Analysis*, Vol. 1, Humana Press.

[421] Konrat, R. (2014) NMR contributions to structural dynamics studies of intrinsically disordered proteins., *J. Magn. Reson. 241*, 74-85.

Appendix A

Resonance Assignments for His-tagged Tat$_{101}$

Table A.1. Resonance assignments of the His-tagged Tat$_{101}$ from HIV -1 determined at pH 4.0 and 293 K.

Position	Residue type	Chemical shift (ppm)					
		H_α	C_α	C_β	C′	N	H_N
1	MET	.	.	.	.	.	.
2	GLY	.	43.6	.	170.45	.	.
3	SER	4.524	58.34	64.22	174.65	115.6	8.681
4	SER	4.396	58.46	64.04	174.55	117.9	8.481
5	HIS	4.665	55.33	29.02	174.25	120	8.556
6	HIS	4.651	55.32	29.3	174.25	119.2	8.54
7	HIS	4.666	55.39	29.4	174.25	120	8.716
8	HIS	4.669	55.46	29.49	174.25	120.4	8.784
9	HIS	4.661	55.65	29.48	174.25	120.9	8.792
10	HIS	4.72	55.52	29.53	174.25	121.5	8.769
11	SER	4.497	58.43	64.14	174.55	118.7	8.564
12	SER	4.49	58.73	64.16	174.95	118.4	8.563
13	GLY	3.958	45.4	.	173.85	110.6	8.424
14	LEU	4.365	55.26	42.59	177.25	121.8	8.115
15	VAL	4.404	59.99	.	174.55	123.2	8.188
16	PRO	.	63.25	32.26	177.05	.	.
17	ARG	4.317	56.59	31.05	177.15	122.1	8.49
18	GLY	3.982	45.41	.	174.25	110.4	8.481
19	SER	4.413	58.54	64.13	174.45	115.5	8.208
20	HIS	4.716	55.51	29.06	174.15	120.2	8.599
21	MET	4.319	55.49	33.19	175.85	121.8	8.391
22	GLU	4.635	54.23	.	.	123.5	8.475
23	PRO	.	63.16	32.24	176.75	.	.

24	VAL	4.03	62.13	33.16	175.75	120.5	8.182
25	ASP	4.854	51.79	.	175.25	125.9	8.387
26	PRO	.	64.01	32.3	177.65	.	.
27	ARG	4.143	57.17	30.33	176.95	118.8	8.377
28	LEU	4.287	55.08	42.32	177.05	120.2	7.87
29	GLU	4.208	54.07	.	174.45	120.4	7.94
30	PRO	.	64.27	31.81	176.85	.	.
31	TRP	4.605	57.47	28.77	176.35	117.6	7.45
32	LYS	4.167	56.29	33.14	175.75	122.3	7.609
33	HIS	4.861	53.32	.	172.35	119.2	8.096
34	PRO	.	63.76	32.27	177.75	.	.
35	GLY	4.012	45.42	.	174.35	109.8	8.601
36	SER	4.467	58.4	64.17	174.35	115.4	8.186
37	GLN	4.611	53.95	.	174.05	123	8.426
38	PRO	.	.	.	.	.	.
39	LYS	4.341	56.7	33.13	177.05	121.9	8.52
40	THR	4.309	61.81	70.19	174.25	115	8.068
41	ALA	4.363	52.72	19.52	177.65	126.4	8.36
42	CYS	4.559	58.72	28.17	175.05	118.9	8.39
43	THR	4.356	62.18	69.93	174.45	116.4	8.264
44	ASN	4.712	53.53	39.03	175.15	121	8.408
45	CYS	4.44	58.77	28.13	174.35	119.2	8.231
46	TYR	4.585	58.25	38.73	175.65	122.6	8.276
47	CYS	4.435	58.52	28.26	174.25	121.1	8.109
48	LYS	4.252	56.84	33.37	176.55	124.3	8.317
49	LYS	4.312	56.65	33.26	176.55	123.5	8.351
50	CYS	4.457	58.66	28.25	174.45	120.6	8.378
51	CYS	4.466	58.57	28.23	174.05	121.7	8.369
52	PHE	4.598	57.94	39.96	175.45	122.8	8.259
53	HIS	4.642	55.28	29.27	173.85	120.8	8.453
54	CYS	4.424	58.65	28.29	174.35	120.6	8.349
55	GLN	4.32	56.21	29.74	175.95	123.3	8.576
56	VAL	4.08	62.66	33.13	175.85	121.9	8.224
57	CYS	4.461	58.47	28.33	174.15	122.9	8.327
58	PHE	4.679	57.91	39.85	175.55	123.5	8.348
59	ILE	4.193	61.23	39.03	176.15	122.9	8.103
60	THR	4.29	62.15	69.99	174.55	118.9	8.156
61	LYS	4.273	56.65	33.24	176.25	124.2	8.282
62	ALA	4.28	52.69	19.29	177.75	125.2	8.277
63	LEU	4.306	55.52	42.72	178.15	121.5	8.192
64	GLY	3.931	45.56	.	174.35	109.1	8.32

65	ILE	4.141	61.42	39.1	176.45	119.9	7.931
66	SER	4.434	58.28	64.02	174.45	119.2	8.299
67	TYR	4.537	58.52	39.03	176.55	122.7	8.219
68	GLY	3.894	45.65	.	174.25	110	8.334
69	ARG	4.296	56.49	31	176.65	120.6	8.134
70	LYS	4.28	56.62	33.17	176.65	122.8	8.322
71	LYS	4.274	56.55	.	.	122.8	8.324
72	ARG	.	.	.	.	.	.
73	ARG	.	56.3	31.03	176.25	123.3	8.498
74	GLN	4.337	55.85	30.04	175.85	122.6	8.511
75	ARG	4.32	56.26	31.21	176.25	123.6	8.525
76	ARG	.	56.15	31.17	176.15	123.3	8.498
77	ARG	4.603	54.17	.	173.95	124.5	8.513
78	PRO	.	.	.	.	.	.
79	PRO	.	63.18	32.19	177.15	.	.
80	GLN	4.324	56.19	29.94	176.75	121	8.544
81	GLY	4	45.44	.	174.35	110.5	8.495
82	SER	4.436	58.73	64.07	174.95	115.6	8.288
83	GLN	4.401	56.11	29.54	176.35	122.2	8.54
84	THR	4.27	62.18	69.97	174.45	114.8	8.128
85	HIS	4.717	55.36	29.11	174.15	120.6	8.525
86	GLN	4.362	56.08	29.73	175.95	122.5	8.473
87	VAL	4.136	62.49	33.13	176.15	122.2	8.326
88	SER	4.479	58.21	63.99	174.65	119.7	8.426
89	LEU	4.396	55.4	42.53	177.55	125.1	8.406
90	SER	4.407	58.56	64.05	174.55	116.5	8.269
91	LYS	4.339	56.29	33.16	176.35	123.5	8.345
92	GLN	4.591	53.91	.	174.15	122.8	8.359
93	PRO	.	63.25	32.27	176.75	.	.
94	ALA	4.315	52.76	19.38	177.95	124.4	8.459
95	SER	4.425	58.24	64.1	174.25	114.9	8.243
96	GLN	4.649	53.9	.	174.15	123	8.339
97	PRO	.	63.38	32.29	177.05	.	.
98	ARG	4.331	56.35	31.17	176.95	121.6	8.5
99	GLY	3.931	45.21	.	173.45	110	8.369
100	ASP	4.908	52.31	.	174.85	121.2	8.311
101	PRO	.	63.78	32.24	177.45	.	.
102	THR	4.356	62.12	70.03	175.05	113.1	8.253
103	GLY	4.099	45.19	.	172.05	111	8.118
104	PRO	.	63.42	32.29	177.45	.	.
105	LYS	4.284	56.66	33.07	176.95	121.4	8.467

106	GLU	4.349	56.38	30.01	176.45	121.4	8.352
107	SER	4.424	58.61	64.05	174.65	117.5	8.38
108	LYS	4.337	56.43	33.13	176.45	123.2	8.312
109	LYS	4.256	56.62	33.2	176.55	122.5	8.246
110	LYS	.	56.62	33.2	176.55	122.5	8.246
111	VAL	4.092	62.49	32.96	176.25	122.3	8.214
112	GLU	4.36	56.32	30.03	176.15	125	8.474
113	ARG	4.326	56.37	33.26	176.35	122.1	8.379
114	GLU	4.456	56.47	29.72	176.35	123	8.405
115	THR	4.311	62.11	70.04	174.65	114.7	8.163
116	GLU	4.308	56.46	29.72	176.25	122.8	8.401
117	THR	4.342	61.93	70.03	174.15	114.7	8.163
118	ASP	4.93	52.47	.	173.75	123.5	8.376
119	PRO	.	63.41	32.25	176.95	.	.
120	VAL	4.115	62.44	32.88	175.45	119.4	8.197
121	ASP	4.507	54.6	.	.	126.6	8.035

Appendix B

Gene sequence of the full-length HIV-1 Tat$_{101}$

CATCATCATCATCATCACAGCAGCGGCCTGGTGCCGCGCGGCAGCCATATGGAACC
GGTCGACCCGCGTCTGGAACCATGGAAACACCCCGGGTCCCAGCCGAAAACCGCGT
GCACCAACTGCTACTGCAAAAAATGCTGCTTCCACTGCCAGGTTTGCTTCATCACCA
AAGCCCTAGGTATCTCTTACGGCCGTAAAAAACGTCGTCAGCGACGTCGTCCGCCGC
AGGGATCCCAGACTCATCAAGTTTCCTTGTCCAAGCAACCGGCGTCTCAGCCGCGTG
GTGACCCGACCGGTCCGAAAGAATCTAAAAAAAAAGTTGAACGTGAAACCGAAACC
GACCCGGTTGAC