Structures and functions of the C-Terminal domain of HIV-1 integration

Oyindamola Oladosu

UNIVERSITÉ DE STRASBOURG

ÉCOLE DOCTORALE DES SCIENCES DE LA VIE ET DE LA SANTÉ

[UMR 7104]

THÈSE présentée par :

[Oyindamola OLADOSU]

soutenue le : **16 Mai 2017**

pour obtenir le grade de : **Docteur de l'université de Strasbourg**

Discipline/ Spécialité : Biophysique et Biologie Structurale

Structures et Fonctions du Domaine C-Terminal de L'Intégrase du VIH-1

THÈSE dirigée par :

M. RUFF Marc	Directeur de Recherche, université de Strasbourg

RAPPORTEURS :

Mme. ANDREOLA Marie-Line	Directeur de Recherche, université de Bordeaux
M. BOUAZIZ Serge	Directeur de Recherche, université de Paris Descartes

AUTRES MEMBRES DU JURY :

Mme. BERNACCHI Serena	Chargé de Recherche, université de Strasbourg
M. NEGRONI Matteo	Directeur de Recherche, université de Strasbourg
M. PARISSI Vincent	Directeur de Recherche, université de Bordeaux

For Toyin and Toyin
You made me everything I am today
I hope I have made you proud

Acknowledgements

I would like to start by thanking God for bringing me this far and making all of this possible. I am nothing without Him.

A big thank you to all of the members of my jury: Dr. Marie-Line Andreola, Dr. Serge Bouaziz, Dr. Serena Bernacchi, Dr. Matteo Negroni and Dr. Vincent Parissi, for taking the time to read and evaluate my work.

I am forever indebted to my thesis supervisor Marc Ruff for allowing me to join his team and giving me the opportunity to work on this project. Marc, you are such a brilliant scientist and it was a huge privilege to learn from you. Thank you for all your support and for being an overall super awesome boss.

I thank the PhD program through the LABEX and SIDACTION for funding my thesis.

I would like to especially thank Raphael Recht and Bruno Kieffer for countless hours spent helping me with NMR and discussing about my project. I would also like to thank Yves Nomine for helping me with CD experiments, and for the helpful discussions about my results. I would like to thank the Molecular Biology platform, especially Paola Rossolillo for producing so many expression vectors used for my projects.

To all the past and present members of the Ruff team: BenOlas (Benoit and Nicolas), Julien (JuJu), Karine and Sylvia (just realized I don't have nicknames for you guys) and Eduardo. I am grateful for the opportunity to have worked with you over the past 3.5 years. You guys taught me pretty much everything in the lab. Thank you so very much for all the scientific discussions, for always answering my stupid questions, and for always being there to lend a helping hand. You are all super brilliant, and each one of you contributed immensely to my time in the lab. I am also thankful for all your contributions towards the completion of this manuscript.

A special shout out to my friends Chantal (ChanChan), Francesca and Lorraine, for all their support and always being there to hear me rant and listen to my "business ideas" during lunch. I'm especially thankful that you guys help me with everything from taxes to hospital visits to making phone calls. Arielle, thanks for being the president of the "fat ass club". It was nice having another 'sista' at the IGBMC. Thanks to Fabrice and Hubert for all your

help, from talking about my project during lunch, to helping me move apartments, you guys were always there, and I really appreciate it.

Dr. Nicaise Ndembi, thank you for introducing me to HIV research and allowing me to work in your lab at IHVN. That experience helped shape my scientific interests and pushed me to do this PhD.

Sam, I could write a whole book about what an incredible friend you have been to me over the past few years. I can't believe I was fronting on you at first. I came to Strasbourg for a PhD and I also got a sister in the process. You have been there for me every step of way and I am forever thankful to you.

To my mamas Aunt Deola and Mama Hart, you are both so inspiring. I'm blessed to have you in my life.

To my friends: Ona, Kunle, Wole, Brendan, Bolaji, Abby, Faith, Gloria, thank you for always being a phone call away and for keeping me grounded.

Bimpe and Deoti, the best sisters in the world. You girls inspire me so much and push me to be the best version of myself. Thank you for everything.

Finally, to my husband Valentine, thank you for all your support. None of this would have been possible without you. I'm blessed to be on this journey with you!

Abstract en Français

L’Integrase du VIH est une ADN recombinase catalysant deux réactions qui permettent l'intégration de l'ADN viral dans l'ADN hôte. L’intégrase du VIH comprend 3 domaines : N-terminal impliqué dans la réaction de « 3' processing » et le transfert de brin, le domaine catalytique contenant le site actif et le domaine C-terminal liant l'ADN non-spécifiquement (CTD). Des recherches récentes mettent en évidence l'importance du CTD dans la liaison avec d'autres protéines virales comme la transcriptase inverse.

L'intégration du génome viral dans le génome de l'hôte nécessite un ciblage dans des régions spécifiques de la chromatine. Des facteurs de ciblage tels que BET ou LEDGF ont été découverts comme jouant un rôle dans la sélection des sites d'intégration. Ces facteurs reconnaissent des modifications spécifiques des histones. Cependant, les processus d'association avec le nucléosome et l'intégration dans la chromatine ne sont pas pleinement compris. L’IN-CTD de l’intégrase possède un domaine « SH3 like » avec une structure en tonneau beta. Cette structure est compatible avec la superfamille « Royal Domain » qui reconnaît les histones modifiées. Nos collaborateurs (Vincent Parissi, Bordeaux) ont montré que la queue N-terminale mono-methylé de l’histone H4 (H4K20Me1) interagit spécifiquement avec le CTD.

Une caractéristique du VIH est sa grande variabilité génétique. Cette souplesse, qui permet une diversité antigénique importante est centrale pour l'adaptation virale à la réponse immunitaire et pour échapper aux antiviraux. L'inconvénient de la variation génétique est la perte de la fonctionnalité des protéines nécessaire pour l'infectivité virale. La façon dont l'équilibre est maintenu entre ces exigences divergentes est essentielle pour notre compréhension de l'évolution virale. L'examen de ces questions peut fournir de nouvelles perspectives sur la biologie du VIH et sur l'identification de cibles pour les antirétroviraux.

Les chimères inter-groupes entre le groupe M (sous-type A2) et le groupe O ont fourni la preuve qu'un motif $N_{222}K_{240}N_{254}K_{273}$ dans le groupe M, remplacé par le motif $K_{222}Q_{240}K_{254}Q_{273}$ dans le groupe O, est important pour l'intégration (Matteo Negroni, IBMC). Nous sommes intéressés par les différences structurales entre le CTD de ces différents groupes et sous-types.

Le but de ma thèse était de comprendre les rôles et l'importance du domaine C-terminal de l'intégrase dans deux contextes : l'intégration dans la chromatine et la coévolution, avec l'objectif de comprendre le rôle de la multimerisation dans la fonction de l'intégrase.

La première partie de mon projet a porté sur l'élucidation du rôle du CTD dans les interactions avec les histones en comprenant les bases structurales de l'interaction entre le CTD et H4K20me1. La procédure globale impliquait des études biochimiques, fonctionnelles et structurales.

En utilisant la thermophorèse, j'ai montré que l'IN-CTD se lie préférentiellement au peptide H4K20Me1 avec un Kd de 0,8 μM. Ces résultats ont été confirmés à l'aide de la NOESY-RMN où un changement du signal NOE sur le peptide indiquait la formation d'un complexe CTD/H4K20Me1.

Les spectres HSQC ^{15}N ont montré que les résidus 271-288 étaient désordonnés. En outre, les résultats de RMN ont indiqué que la protéine est plus désordonnée dans 150mM NaCl comparativement à 1M NaCl et moins oligomérisée à pH 8 par rapport à pH 7. En utilisant le CTD marqué isotopiquement, j'ai attribué avec succès 75% des résidus et calculé des structures RMN montrant une topologie semblable aux structures publiées. J'ai observé des interactions dépendant du pH avec H4K20Me1. Lors de la liaison du peptide à pH 8, les spectres HSQC suggèrent un changement de conformation conduisant à une oligomérisation plus importante. En quantifiant les changements de déplacements chimiques observés pendant le titrage peptidique, j'ai identifié les résidus affectés à pH 7 et à pH 8.

En utilisant la cristallographie, j'ai obtenu des structures du CTD à pH 6.5, 7 et 7.5. En utilisant la structure à pH 7 (1.5 Å), j'ai cartographié les résidus dont les déplacements chimiques ont été perturbés par la liaison au peptide et observé qu'ils étaient situés dans la partie N et C terminale. Plusieurs de ces résidus sont hydrophobe suggérant qu'ils forment un patch hydrophobe pour la fixation de H4K20Me1. Nos expériences de docking suggèrent que la lysine méthylée peut interagir avec ce patch.

Sur la base de ces données, je propose que la liaison de l'IN-CTD au peptide se produit via le patch hydrophobe, induisant des changements de conformation (oligomères), et provoquant des changements dans l'exposition au solvant de certains résidus suggérant que la propension à la multimerisation du CTD joue un rôle dans la fonction de l'IN.

Le but de la deuxième partie de mon projet était de résoudre la structure du CTD sauvage et chimères afin de comprendre l'importance du motif $N_{222}K_{240}N_{254}K_{273}$ pour la fonction de l'IN et son rôle dans la coévolution. À cette fin, j'ai purifié des constructions CTD 220-270 sauvage ainsi que des chimères inter groupe. J'ai obtenu des cristaux et résolu les structures de la protéine sauvage et des mutants sélectionnés.

J'ai obtenu une structure à 1.7 Å du CTD A2, qui montre trois surfaces de contact possibles pour la dimérisation. La structure du mutant K240Q/N254K (2 Å) était similaire à la structure du sauvage, sauf dans les boucles et les régions flexibles. Les résultats suggèrent que le motif $N_{222}K_{240}N_{254}K_{273}$ non conservé joue un rôle dans la formation d'oligomères d'ordre supérieur. En particulier, K240/N254 sont importants pour la dimérisation. Comme l'IN doit oligomériser afin d'être actif, ce motif est important la fonction de l'IN.

Globalement, les résultats de mon projet indiquent que l'IN-CTD joue un rôle important, en contribuant à la formation de multimères d'ordre supérieur importants pour la fonction de l'IN. Le projet souligne l'importance d'une approche de biologie structurale intégrée pour répondre aux questions biologiques. Les résultats obtenus dans ce projet sont importants pour comprendre les relations structure/fonction de l'intégrase pendant le cycle de vie du VIH-1. Les méthodes développées au cours de ce projet peuvent être utilisées pour cribler de nouveaux inhibiteurs de conformation du l'IN du VIH-1.

Abstract in English

HIV Integrase is a DNA recombinase that catalyzes two endonucleolytic reactions that allow the viral DNA integration into host DNA for replication and subsequent viral protein production. HIV Integrase consists of 3 structural and functional domains: The N-terminal zinc domain involved in 3' processing and strand transfer, the catalytic core domain which contains the active site, and the C-terminal domain that binds DNA non-specifically. Recent research highlights the importance of the CTD in binding with other viral proteins such as Reverse Transcriptase.

The integration of the viral genome into the host's genome requires targeting into specific regions of the chromatin. Cellular targeting factors such as BET or LEDGF/p75 have been discovered to play a role in the integration site selection because they recognize and bind specific histone modifications. However, the process of nucleosome association and chromatin integration is not yet fully understood. The C-terminal domain of HIV Integrase possess a SH3 like domain made up of approximately 60 amino acids that displays a β-barrel fold. This SH3 fold is consistent with the Royal Domain superfamily that recognize post translationally modified histone tails. Our collaborators (Vincent Parissi, Bordeaux) show that the monomethylated N-terminal tail of H4K20Me1 interacts specifically with the C-terminal Domain of HIV integrase.

A main feature of the human immunodeficiency virus (HIV) is its great variability, reflecting a remarkable genetic flexibility. This flexibility, which allows antigenic variation, is central for viral adaptation to the immune response mounted by the host and to escape antiviral treatments. The drawback of genetic variation can be the loss of proteins and nucleic acids functionality, necessary to ensure viral infectivity. How the balance is kept between these two divergent requirements is central for our understanding of viral evolution. Addressing these issues can provide new insights into HIV biology and on the identification of potential targets for antiviral strategies.

Inter-group chimeric constructs between group M (subtype A2) and group O provided evidence that a non-conserved motif, $N_{222}K_{240}N_{254}K_{273}$ in group M, replaced by $K_{222}Q_{240}K_{254}Q_{273}$ in group O, is important for integration (Matteo Negroni, IBMC). We are interested in the structural differences between the CTD of these different groups and subtypes.

The aim of my thesis was to understand the roles and importance of the C-terminal domain of HIV-1 Integrase in two contexts: chromatin integration, and co-evolution, with the overall purpose of understanding the role of multimerization in IN function.

The first part of my project focused on elucidating the role of the IN-CTD in histone interactions, by understanding the structural basis of the interaction between IN-CTD and mono-methylated N-terminal tail of histone H4K20 (H4K20me1). The overall procedure involved biochemical, *in-vitro* functional and structural studies.

Using Microscale Thermophoresis, my results show that the CTD preferentially binds to the fluorescent mono-methylated H4K20 peptide with a K_d of 0.8μM. These results were confirmed using NOESY-NMR, where a change in NOE signal on the peptide indicated complex formation between the CTD and H4K20Me1.

^{15}N HSQC spectra showed that the residues 271-288 of the CTD were disordered. Additionally, NMR results indicated that the protein is more disordered in 150mM NaCl, compared to 1M NaCl and less oligomerized at pH 8, compared to pH 7. Using isotopically labeled CTD, I successfully assigned 75% of the protein residues, and calculated NMR structures, whose overall topology was similar to published structures. I also observed pH dependent interactions with H4K20Me1. Upon binding of the peptide at pH 8, the HSQC spectra suggest a change to a more oligomerized protein conformation. By quantifying the chemical shift changes observed during peptide titration, I was able to identify the residues affected by peptide binding at pH 7 and pH 8.

Using X-ray crystallography, I obtained high-resolution structures of the isolated IN-CTD at pH 6.5, 7 and 7.5. Using the crystal structure of the IN-CTD at pH 7 (1.5 Å), I mapped the residues whose chemical shifts were perturbed by peptide binding, and observed that they were mostly located on the N and C terminus of the IN-CTD. Interestingly, several of these residues are hydrophobic in nature, suggesting that they form a hydrophobic patch, which serves as the binding surface for H4K20Me1. Our preliminary docking experiments suggest that the methylated lysine is in close proximity and may interact with this patch.

Based on this data, I propose that binding on IN-CTD to the peptide occurs via the hydrophobic patch, inducing conformational changes (oligomerization), and causing changes

in solvent exposure of some residues. This suggests that the propensity for multimerization of the IN-CTD plays a role in IN function.

The goal of the second part of my project was to solve the structure of IN-CTD from WT and chimeric constructs in order to understand the importance of the motif $N_{222}K_{240}N_{254}K_{273}$ for IN function and its role in co-evolution. To this end, I purified HIS-IN-CTD 220-270 from WT and inter-group chimeric constructs. I obtained crystals and solved the structure of selected WT and mutant proteins.

I obtained a high-resolution structure (1.7Å) of the Group A2 WT IN-CTD, which presented three possible surface contacts for dimerization. The structure of the K240Q/N254K mutant (2Å) was similar in structure to the WT, except in loops and flexible regions. Results suggest that the non-conserved $N_{222}K_{240}N_{254}K_{273}$ motif plays a role in the formation of higher order oligomers in the IN-CTD. In particular, K240/N254 are important for dimerization. As IN needs to be oligomerized in order to be active, this motif is important for IN function

Overall, results from my project indicate that the IN-CTD plays an important role, by contributing to the formation of higher order multimers that are important for IN functionality. Project highlights the importance of an integrated structural biology approach to answering biological questions. Insights gained from this project are important for understanding structure/function relationships of Integrase during the HIV-1 life cycle. The methods developed (NMR and X-ray) during this project can be used to screen for new conformational inhibitors for HIV-IN.

Table of Contents

TABLE OF ABBREVIATIONS

A	
AIDS	Acquired Immune Deficiency Syndrome
B	
βME	β-mercaptoethanol
BET	Bromodomain and Exter terminal domain
C	
CA	Capsid
CCD	Catayltic Core domain
CCR5	Chemokine (C-C motif) receptor 5
CDC	Center for Diseases Control
CDK9	Cyclin dependent kinase 9
CD4	Cluster of Differentiation 4
CTD	C-terminal Domain
CRFs	Circulating Recombinant Forms
CRM1	Chromosome region maintenance 1
CXCR4	Chemokine (C-X-C motif) receptor 4
D	
D2O	Deuteriom Oxide
dsDNA	double stranded Deoxyribonucleic Acid
dNTPs	deoxynucleotides
E	
Env	envelope
F	
FDA	Food and Drug Administration
FIV	Feline Immunodeficiency Virus
G	
Gag	group specific antigen
GST	Glutathione S- Transferase
H	
H4K20Me1	monomethylated H4K20 peptide
HAART	Highly Active Anti-Retroviral Therapy
HIV	Human Immunodeficiency Virus
I	
IBD	Integrase binding domain
IN	Integrase
IPTG	IsoPropyl β-D-1-ThioGalactopyranoside
K	

Kd	Dissociation Constant

L

LEDGF	Lens Epithelium Derived Growth Factor

M

mRNA	messenger RNA
MA	Matrix
MHz	megahertz
MLV	Murine Leukemia Virus
MST	Microscae thermophoresis
MVV	Maedi-Visna Virus
MWCO	Molecular Weight Cut Off

N

Nef	Negative Regulatory Factor
NC	Nucleocapsid
NMR	Nuclear Magnetic Resonance
NNRTIs	Non-nucleoside reverse transcriptase inhibitors
NRTIs	Nucleoside reverse transcriptase inhibitors
NOEs	Nuclear Overhauser Effects

P

P24	protein 24
P3C	Protease 3C precision
PCR	Polymerase Chain Reaction
PIC	Pre-integration complex
PFV	Prototype foamy virus
Pol	Polymerase
Ppm	parts per million
PR-	Protease

R

RNAse1	Ribonuclease 1
RNA	Ribonucleic Acid
RT	Reverse Transcriptase
RTC	Reverse transcriptase complex

S

SH3	Src Homology 3
SIV	Simian Immunodeficiency Virus
STC	Strand transfer complex
SDS-PAGE	Sodium Dodecyl Sulfate Polyacrylamide (SDS-PAGE) Electrophoresis

T

TAR	transactivation response protein
Tat	transactivator of transcription
TCC	target capture complex
TE	Tris-EDTA
trNOESY	Transferred Nuclear Overhauser Effect Spectroscopy

U

U3	Unique sequence element at the 3'end of the viral RNA
U5	Unique sequence element at the 5'end of the viral RNA
URFs	Unique Recombinant forms

V

Vif	Viral infectivity factor
Vpr	Viral Protein R
Vpu	Viral Protein U

W

WaterLOGSY	Water-Ligand Observed via Gradient Spectroscopy

TABLE OF FIGURES

CHAPTER 1 - INTRODUCTION

Origin and Epidemiology of HIV

Human Immunodeficiency Virus (HIV) is the causative agent of AIDS (Acquired Immune Deficiency Syndrome). According to the WHO, by the end of 2015, there were 36.7 million people living with HIV, with the highest population of 25.5 million in Africa. Recent advancements in HIV testing and treatments have led to an overall decline in new infections with 18.2 million people living with HIV on anti-retroviral therapy, and 1.1 million AIDS-related deaths in 2015 (W.H.O 2016). However, with 2.1 million new infections in 2015 (approximately 5700 new infections daily), the lack of a preventive vaccine or cure, and high HIV drug resistance rates, HIV remains a global public health threat.

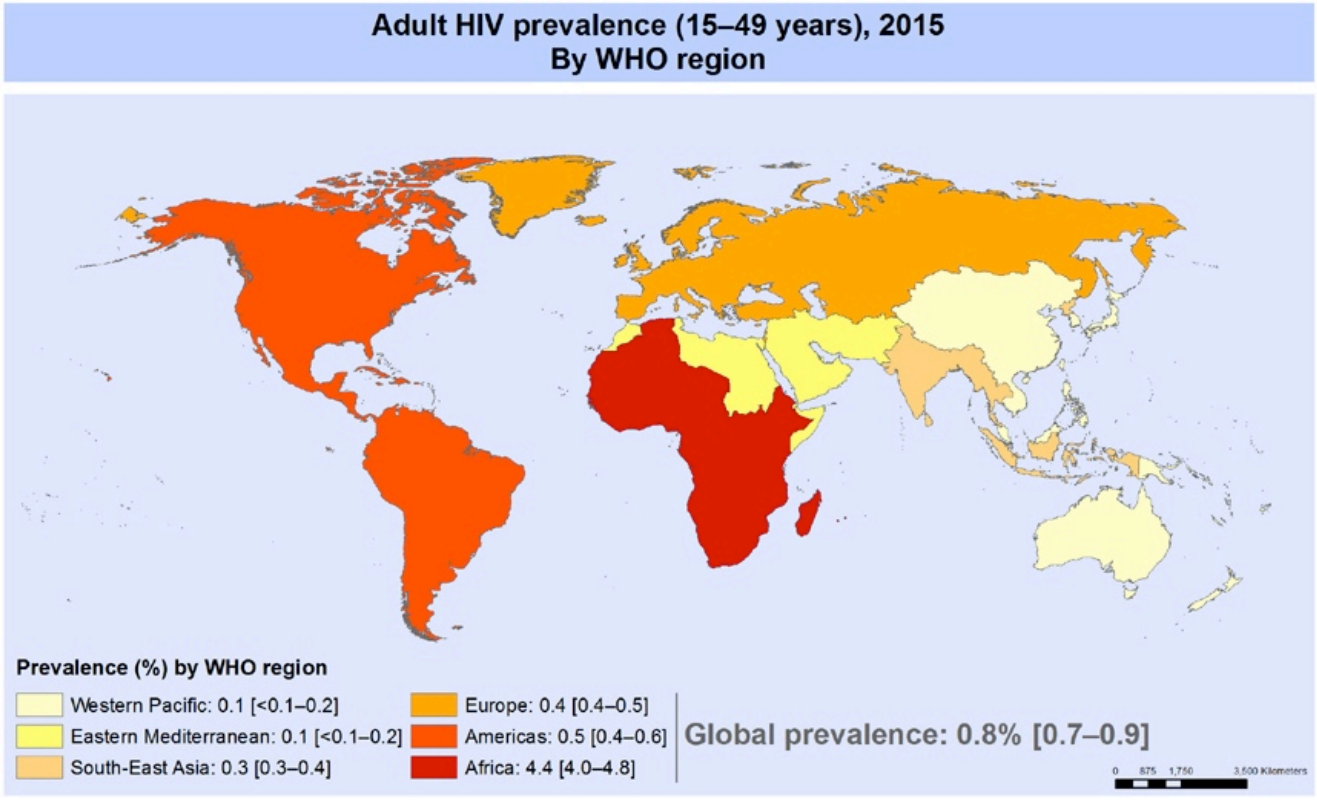


Figure 1: World map showing HIV prevalence by region.

AIDS was recognized as a disease in 1981 (Sharp and Hahn 2011), and in 1983 (Barre-Sinoussi, Chermann et al. 1983), HIV-1 was determined to be the cause of AIDS. HIV can be transmitted through three transmission routes (Shaw and Hunter 2012): mucosal transmission through body fluids such as sperm and vaginal fluid during unprotected sex, parenteral transmission through blood transfusion and needle sharing, and mother to child transmission

in utero, during childbirth or breastfeeding. With 80% of adults acquiring HIV through mucosal surfaces, it is considered to be a sexually transmitted virus (Cohen, Shaw et al. 2011).

HIV infection occurs in 3 stages if untreated (Coffin and Swanstrom 2013, AIDS.gov 2015, AIDSinfo 2016). The acute infection stage develops within 2-4 weeks of infection, with some people showing flu-like symptoms such as fever, headache, sore throat and body rash. HIV replicates and spreads rapidly in this stage, depleting CD4 levels, and compromising the immune system. The process of seroconversion occurs, where the body develops antibodies against HIV, which are detectable for HIV test. In the chronic infection stage or the asymptomatic stage, the virus goes into latency and continues to multiply at low levels, without showing symptoms in some people. It can take up to 10 years before the disease progresses to the final stage of symptomatic infection. Symptoms at this stage include chronic diarrhea, persistent cough and weight loss. By this stage, the immune system is completely weakened, with a CD4 count of less than 200 cells/m^3, making the patient prone to opportunistic infections such as tuberculosis, malaria and cancers. Without treatment, the lifespan of people living with AIDS is about 3 years. Three types of tests can detect HIV infection: antibody tests that detect antibodies against HIV in the blood, usually with test kits, a combination test that detects viral p24 antigens and antibodies, and a nucleic acid test that detects viral load levels in the blood (CDC 2016).

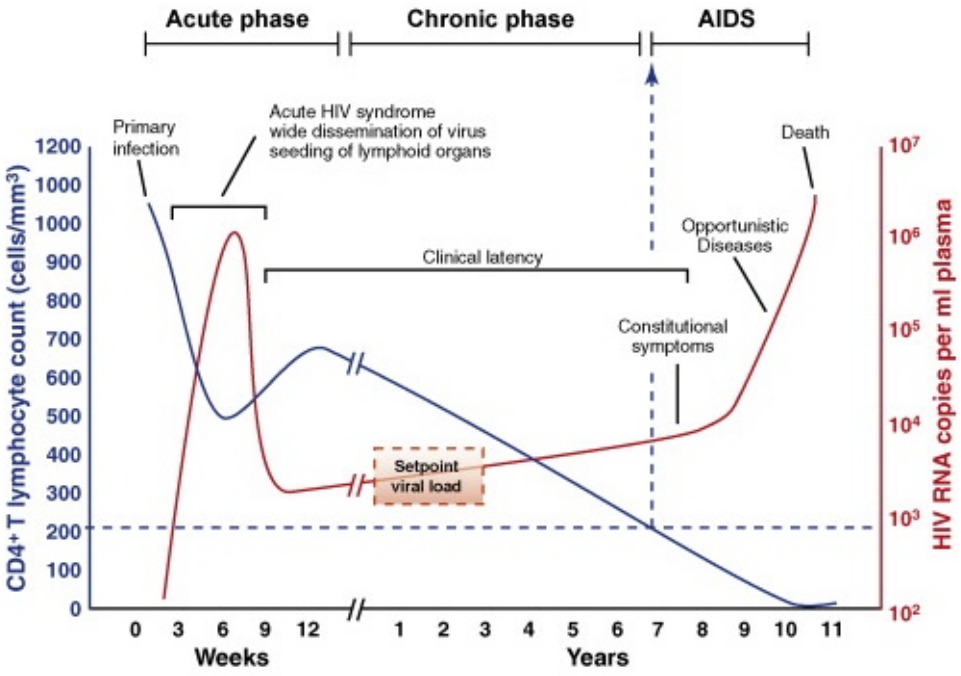


Figure 2: Typical course of HIV progression. *An and Winkler, Trends in Genetics, 2010*

There is currently no cure for HIV, and the best treatment is prevention. For people living with HIV, HAART (Highly Active Antiretroviral Treatment), a combination therapy using multiple classes of antiretroviral regimens that target different steps of the HIV life cycle is used to treat HIV. HAART therapy offers several benefits including viral load reduction, disease progression delay and opportunistic infection prevention. Additionally, HAART reduces side effects by reducing drug dosage, and decreases the emergence of drug resistant viruses by reducing the ability of the virus to adapt and mutate. There are currently six classes of antiretrovirals on the market.

HIV Phylogeny

HIV emerged in the late 19th/early 20th century from non-human primates through cross-species zoonosis of Simian Immunodeficiency Virus (Hahn, Shaw et al. 2000, Sharp and Hahn 2011). HIV belongs to the Lentivirus genus in the Retroviridae family. Lentiviruses are characterized by long incubation periods, and are typically enveloped viruses. Retroviruses contain a positive single stranded RNA genome, and contain the reverse transcriptase protein, which allow them to transcribe their single stranded RNA into double stranded DNA once inside the host cell. Other lentiviruses include SIV, Feline Immunodeficiency Virus (FIV), and Maedi-Visna Virus (MVV) (Clapham and McKnight 2002). HIV is also an enveloped virus, budding off from the host cell enveloped by fragment of the host cell membrane.

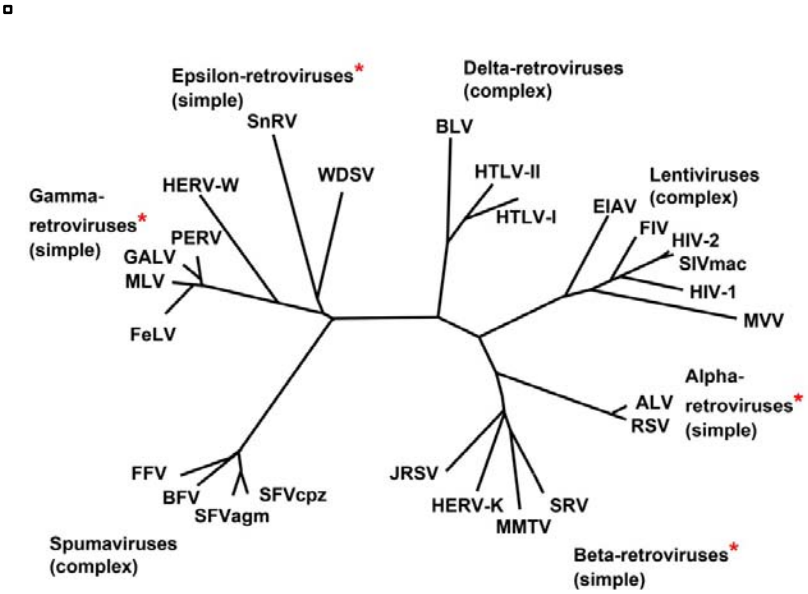


Figure 3: Members of the Retroviridae Family. *François Charles Javaugue, VIH, ed. Hermann, 2014*

There are two classes of HIV: HIV-1 and HIV-2. HIV-1 is classified into 4 groups – M, N, O, and P, while HIV-2 is classified into groups A-H. Group M is the most widely spread group, responsible for approximately 33 million infections worldwide (Hemelaar 2012). Group M is further sub-divided into subtypes A, B, C, D, F, G, H, J, K. There are also circulating recombinant forms (CRFs) and unique recombinant forms (URFs) of HIV-1. Subtype B is most common subtype in Group M. Group O infections are found predominantly in Western-Central Africa. HIV Group M and N originated from SIVcpz in the chimpanzee *Pan troglodytes troglodytes* in West–Central Africa (Gao, Bailes et al. 1999). HIV Group O and P originated from SIVgor in Western lowland gorillas in Cameroon (Van Heuverswyn, Li et al. 2006, Plantier, Leoz et al. 2009). HIV-2 originated from SIVsm in sooty mangabey monkeys (*Cercocebus atys*) in West Africa (Wertheim and Worobey 2009).

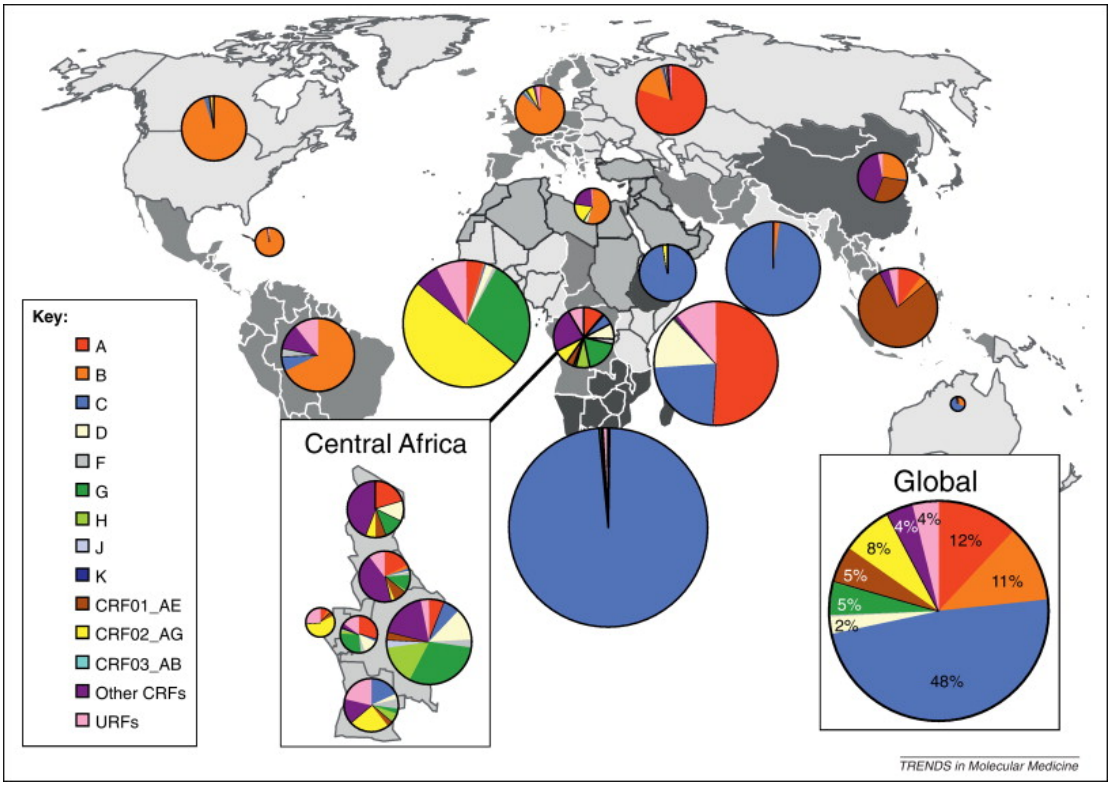


Figure 4: World map showing global distribution of HIV groups and subtypes. *J. Hemelaar, Trends in Molecular Medicine, 2012*

HIV Structure

The HIV virion is about 120nm in diameter with a spherical shape. The lipid bilayer, derived from the host cell, contains host membrane proteins including antigens, major histocompatibility complex, and ubiquitin (Arthur, Bess et al. 1992). On the surface, the envelope consists of trimers of glycoprotein gp120, attached to gp41 via the transmembrane protein(Ganser-Pornillos, Yeager et al. 2008). On the inside, the matrix (MA) shell lines the inner membrane and maintains the structural integrity of the virus. There is a fullerene cone capsid (CA) in the center, and a nucleocapsid (NC) that contains the viral genome - two single unspliced positive RNA strands of approximately 9.2kb in size, tightly bound to viral proteins essential for the life cycle such as reverse transcriptase (RT), integrase (IN). Lysine tRNA is bound to the viral RNA, and acts as a reverse transcription primer. Protease (PR) is also found within the nucleocapsid. Some accessory proteins like Nef, Vif and Vpr are packaged in the virus, while other accessory proteins like Rev, Tat and Vpu are not packaged in the virus.

HIV Genome Description

The HIV genome contains nine genes (Frankel and Young 1998, Watts, Dang et al. 2009). Three main genes: Gag, Pol and Env encode for structural proteins, viral enzymes and envelope proteins respectively. Gag encodes for core structural proteins including matrix, capsid and nucleocapsid. Pol encodes for viral enzymes including reverse transcriptase, integrase and protease. Env encodes for envelope protein gp160, which is spliced to gp120 and gp41. Tat and Rev encode for regulatory proteins involved in transcription activation and RNA splicing and export. Vif, Vpr, Vpu and Nef encode for accessory proteins involved in the synthesis regulation, viral RNA processing and other functions.

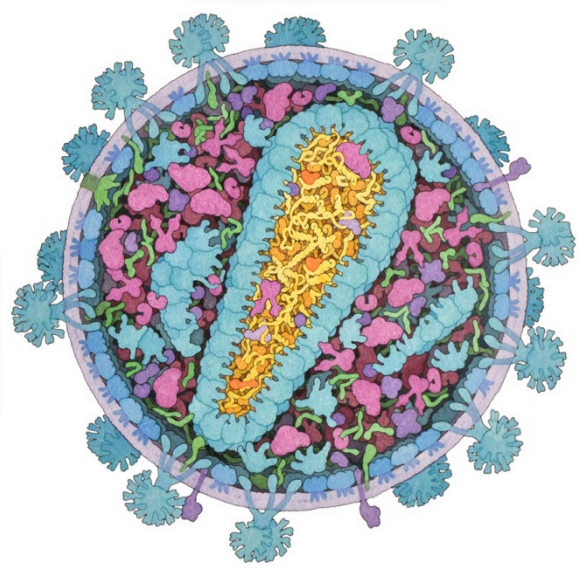


Figure 5: Cross section of Mature HIV-1 virion Structural proteins are shown in blue, viral enzymes in magenta, accessory proteins in green and viral RNA in yellow. Host proteins and tRNA are shown in purple. *David S. Goodsell,* http://hive.scripps.edu/resources.html

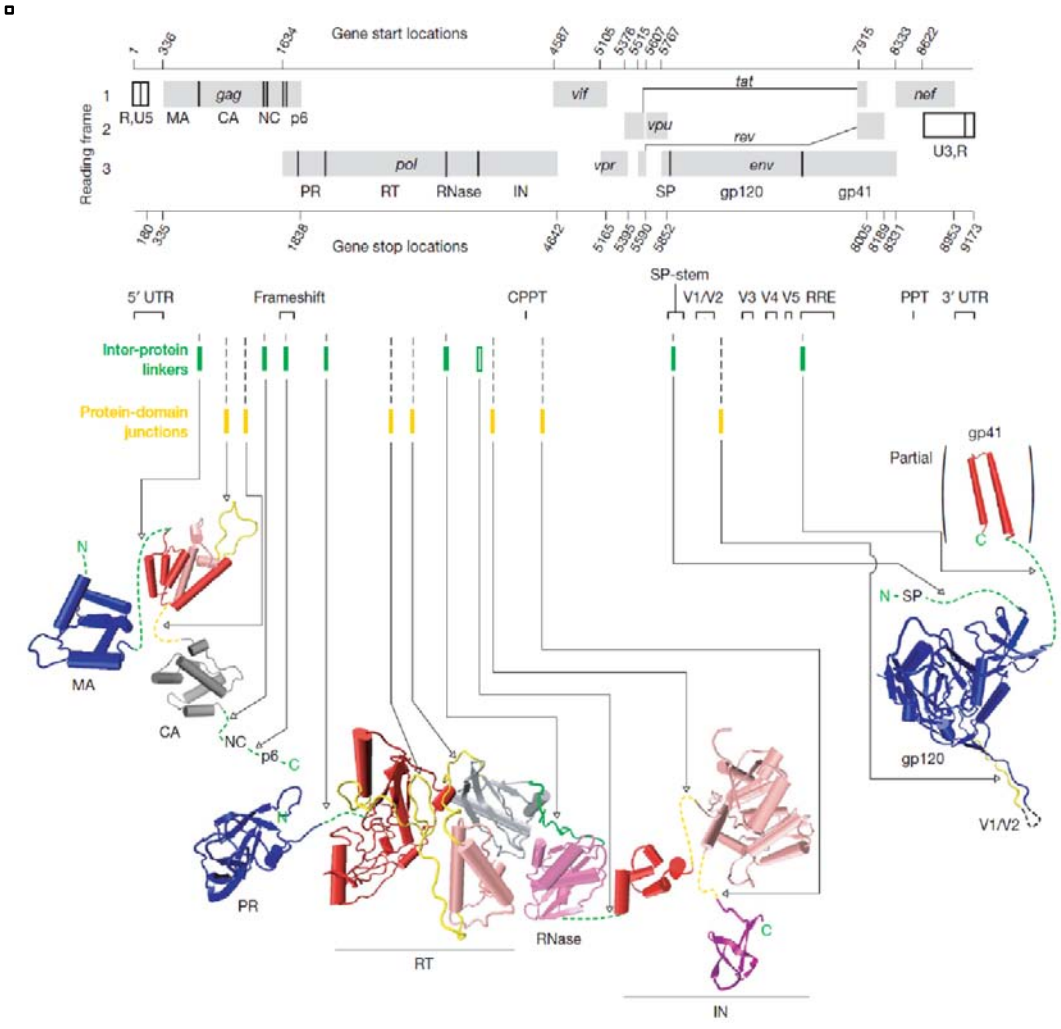


Figure 6: Organization of HIV genome. *Adapted from Watts et al., Nature, 2009*

HIV life cycle

The life cycle of HIV can be broken down into 7 main steps namely binding, fusion, reverse transcription, integration, transcription and translation, assembly and budding. Each of these steps is described briefly below.

Binding

HIV-1 primarily targets T-helper lymphocytes, monocytes, macrophages and dendritic cells that contain the CD4+ receptor and a co-receptor such as CCR5 and CXCR4 (Clapham and McKnight 2002, Liu, Bartesaghi et al. 2008). The spikes on the HIV envelope contain trimers of glycoprotein gp120 non-covalently bound to gp41 (Rizzuto, Wyatt et al. 1998).

Fusion

Upon binding to CD4 cells, a conformational change is induced in gp120, exposing a binding site for the co-receptor. Binding to the co-receptor induces further structural changes in gp120, exposing the hydrophobic fusion peptide on gp41 and causing insertion of the fusion peptide into the cell membrane and subsequent fusion of the viral and host cell membranes (Kwong, Wyatt et al. 1998, Moscoso, Sun et al. 2011, Munro, Gorman et al. 2014).

Reverse Transcription

After viral entry, the capsid shell is partially dissolved, allowing for the formation of the reverse transcriptase complex (RTC) (Fassati and Goff 2001, Forshey, von Schwedler et al. 2002, Peng, Muranyi et al. 2014). This complex consists of viral proteins including reverse transcriptase, capsid, integrase as well as host proteins. During reverse transcription, reverse transcriptase catalyzes the formation of double stranded DNA using the single strand RNA of the virus as a template. This double stranded DNA is rich in uracil to prevent auto-integration (Yan, O'Day et al. 2011). Viral proteins such as nucleocapsid have been shown to improve reverse transcription by improving the binding of primer tRNA Lys 3 to viral RNA (Barat, Lullien et al. 1989). Furthermore, Vif has also been shown to contribute to reverse transcription by increasing the polymerization rate (Cancio, Spadari et al. 2004).

Integration

Following the synthesis of double stranded DNA; the RTC is transformed into the Pre-Integration complex (PIC). The HIV-1 PIC consists of cellular proteins such as LEDGF (Raghavendra, Shkriabai et al. 2010), viral proteins such as nucleocapsid, matrix, Vpr and

reverse transcriptase, and integrase, the core viral protein in the PIC (Matreyek and Engelman 2013). More recently, capsid has also been shown to be in the PIC (Hulme, Kelley et al. 2015). Integrase is responsible for catalyzing the two main interactions (3' processing and strand transfer) that allow for DNA integration. While in the cytoplasm, integrase catalyzes a 3' processing reaction (Miller, Farnet et al. 1997). The PIC travels along microtubules to reach the nuclear envelope, and crosses the nuclear pore complex to reach the genome of non-dividing cells. Vpr promotes nuclear localization of viral DNA during nuclear import (Heinzinger, Bukinsky et al. 1994). Following the active transport, integrase catalyzes the strand transfer reaction, allowing for integration of the viral DNA into the host DNA. Host machinery repairs any gaps, and the viral DNA is replicated along with host DNA (Craigie and Bushman 2012). Post integration, the virus can go into latency, establishing reservoirs that make HIV incurable so far (Dahabieh, Battivelli et al. 2015).

Transcription and Translation

Transcription can occur pre- and post- integration. Pre-integration transcription involves the production of viral regulatory proteins such as Tat, Rev and Nef that interact with cellular proteins to regulate viral transcription and translation (Sloan and Wainberg 2011). Tat initiates post integration transcription from the U3 promoter in the upstream LTR by binding to the viral transactivation response protein (TAR) (Bieniasz, Grdina et al. 1998). This leads to the recruitment the cyclin dependent kinase (CDK9) and cyclin T1 to the TAR. CyclinT1 binds to Tat, increasing its affinity and specificity for RNA. CDK9 mediated phosphorylation of the C-terminal domain of RNA polymerase II stimulates transcription elongation (Fujinaga, Cujec et al. 1998, Isel and Karn 1999, Zhou and Rana 2002).

Newly synthesized unspliced or partially spliced viral mRNAs are exported via Rev or Gag mediated pathways. In the Rev mediated pathway, Rev binds to Rev Response Element (RRE) and recruits cellular nuclear export factors such as chromosome region maintenance 1 (CRM1) and RanGTP to export mRNAs to the cytoplasm through the nuclear pore for the synthesis of Gag and Gag-Pol precursors (Jain and Belasco 2001, Rausch and Grice 2015). Env precursor (gp160) is synthesized on the rough endoplasmic reticulum (Checkley, Luttge et al. 2011). In the Gag mediated pathway, interaction between the nuclear export signal of matrix^{Gag} and CRM1 results in mRNA export (Parent 2011). Rev, Gag, CRM1 and RanGTP are re-imported into the nucleus.

Assembly and Budding

Nascent virions are assembled at the plasma membrane, containing viral RNA, Gag and Gag-Pol, viral proteins such as Vif, Nef, Vpr, and host proteins (Freed 2015) . Following synthesis, Env (gp120 and gp41) trimers are exported to the plasma membrane by secretory pathway. CD4 proteins are also synthesized in the endoplasmic reticulum. Vpu interacts with CD4 to promote Env packaging by inducing CD4 degradation (Le Noury, Mosebi et al. 2015). N-terminal myristoylation of Gag contributes to the membrane association (Morikawa, Hockley et al. 2000). Interactions between proline-rich motifs in Gag and cellular class E vacuolar protein sorting (VPS) proteins cause the virion to be pinched off the plasma membrane (Ren and Hurley 2011). Tetherin, a cellular protein that restricts viral budding, is inhibited by Vpu (McNatt, Zang et al. 2013).

Maturation

About 2400 copies of Gag bud to form an immature particle with 2 copies of unspliced viral genome (Carlson, Briggs et al. 2008). During maturation, non-infectious immature virions are converted to infectious virions by the proteolysis of precursor proteins by protease (Pettit, Moody et al. 1994). Gag and Gag-Pol precursor are processed to produce structural proteins (Matrix, Capsid, Nucleocapsid), and enzymes (Protease, Integrase and Reverse Transcriptase) (Pettit, Moody et al. 1994, Briggs, Riches et al. 2009).

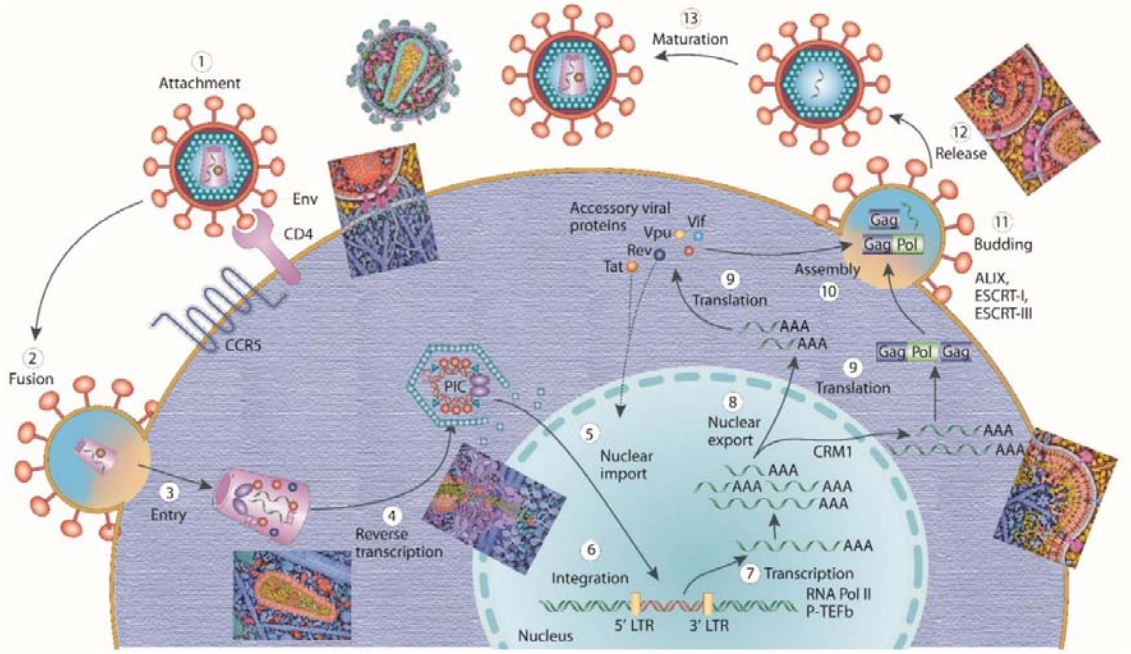


Figure 7: The life cycle of HIV-1. Inserts show pictorial representations of the virus at each step. Adapted from *Engelman and Cherapanov, Nature Reviews, Microbiology, 2012, and David S. Goodsell* http://hive.scripps.edu/resources.html

HIV Inhibitors

There are currently six classes of inhibitors approved by the Food and Drug Administration (FDA) for the treatment of HIV infections (Cihlar and Fordyce 2016, F.D.A 2016). Due to the high rates of drug resistance of HIV to inhibitors, it is important that new inhibitors are designed frequently. There is an emerging class of HIV inhibitors known as maturation inhibitors that inhibit the protease mediated Gag cleavage process (Wang, Lu et al. 2015). HAART (Highly Active Anti Retroviral Therapy), a combination of 3 or more different classes of antiretrovirals is used to improve efficacy and reduce drug resistance rates.

Entry Inhibitors: Maraviroc (Pfizer) is a CCR5 antagonist approved for HIV treatment by the FDA in 2007. It functions by binding to CCR5, preventing the binding of gp120 to the co-receptor (Henrich and Kuritzkes 2013).

Fusion Inhibitors: Enfuvirtide (Roche) was approved by the FDA in 2003 and was the first fusion inhibitor on the market. It works as a peptide mimetic, locking a conformation of gp41, and preventing the structural changes that allow for membrane fusion between host and viral cells (Greenberg and Cammack 2004).

Reverse Transcriptase Inhibitors: RT inhibitors were the first class of inhibitors to be approved by the FDA. They are divided into two sub-classes: nucleoside reverse transcriptase inhibitors (NRTIs) and non-nucleoside reverse transcriptase inhibitors (NNRTIs). They both function by inhibiting the reverse transcription process. NRTIs are nucleoside analogs that compete with dNTPs and inhibit reverse transcriptase when incorporated in the nascent viral DNA chain. NNRTIS bind to reverse transcriptase and act as allosteric inhibitors, preventing conformational changes necessary for activity (Sluis-Cremer, Wainberg et al. 2015). Azidothymidine (AZT), a NNRTI, was the first antiviral drug approved for HIV treatment in 1987 (Fischl, Richman et al. 1987). Since then, there have been 13 NRTIs approved by the FDA. There are currently 6 FDA approved NNRTIs on the market

Integrase Inhibitors: Raltegravir (Merck) was the first FDA approved integrase inhibitor on the market in 2007 (Hicks and Gulick 2009). Dolutegravir (GlaxoSmithkline) and Elvitegravir (Gilead) were approved in 2013 and 2014 respectively. Mg^{2+} ions are necessary for catalytic activity of integrase. These inhibitors prevent the strand transfer process by acting as competitive inhibitors, interfering with Mg^{2+} binding, and preventing the integration of viral DNA into host DNA. Other types of integrase inhibitors, which act as allosteric

inhibitors, binding to the IN/LEDGF complex, and stabilizing IN dimers, are currently under development (Engelman, Kessl et al. 2013, Le Rouzic, Bonnard et al. 2013).

Protease Inhibitors: There are currently 9 protease inhibitors approved by the FDA on the market. Protease inhibitors prevent the formation of mature infectious viruses by binding to the active site of protease and inhibiting the cleavage of HIV protein pre-cursors.(Eron 2000) The FDA approved Saquinavir (Roche), the first protease inhibitor in 1995.

Combination Inhibitors: Recent developments in HAART include the usage of a fixed dose combination of inhibitors from the same class or different classes in one pill, often taken once a day. The first of these inhibitors on the market was Atripla (Bristol-Meyers Squibb, Gilead), a combination of NRTIs and NNRTIs, approved for use in 2006. There are currently six approved combination inhibitors on the market. These inhibitors improve adherence and lower hospitalization risk in patients (Sax, Meyers et al. 2012).

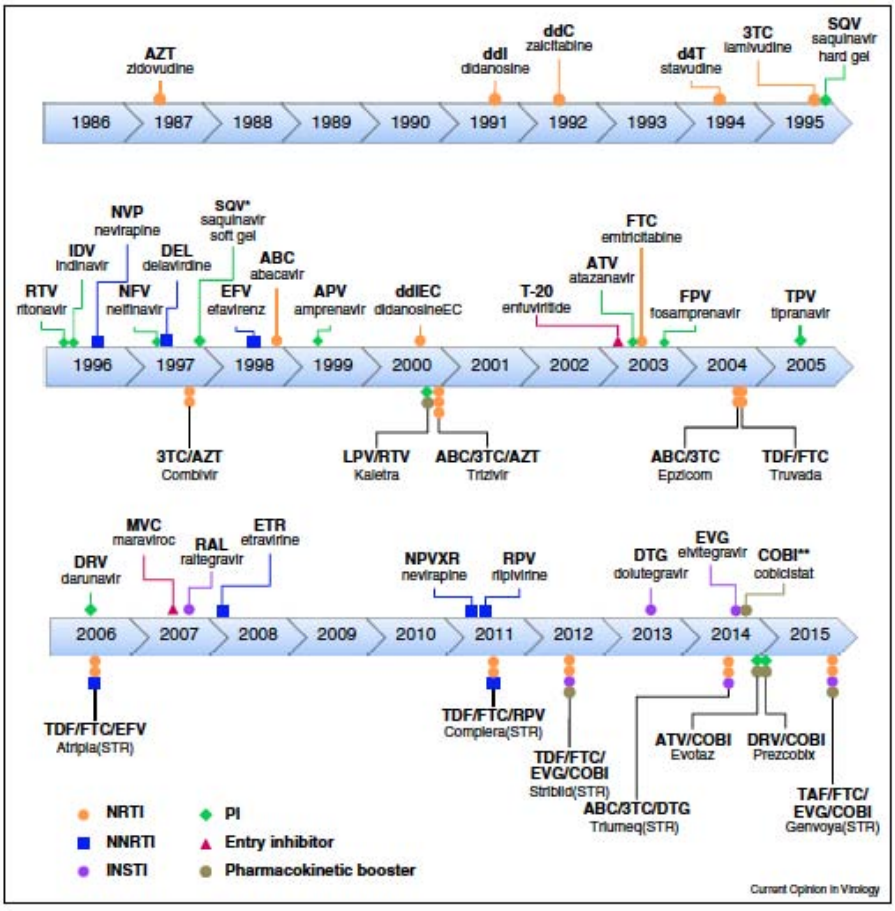


Figure 8: Timeline of HIV-1 inhibitors approved by the FDA. *Cihlar and Fordyce, Current Opinions in Virology, 2016*

HIV Integrase

Integrase is a 288-amino acid (32kDa) protein that is encoded by the end of the pol gene. It is produced as part of the Gag-Pol polyprotein, after which it is cleaved by protease mediated cleavage. It consists of 3 structural and functional domains: the N-terminal zinc finger domain (1-49), which is responsible for protein multimerization, the Catalytic core domain (50-212), which contains the active site of the protein, and the C-terminal domain (213-288) that binds DNA non-specifically.

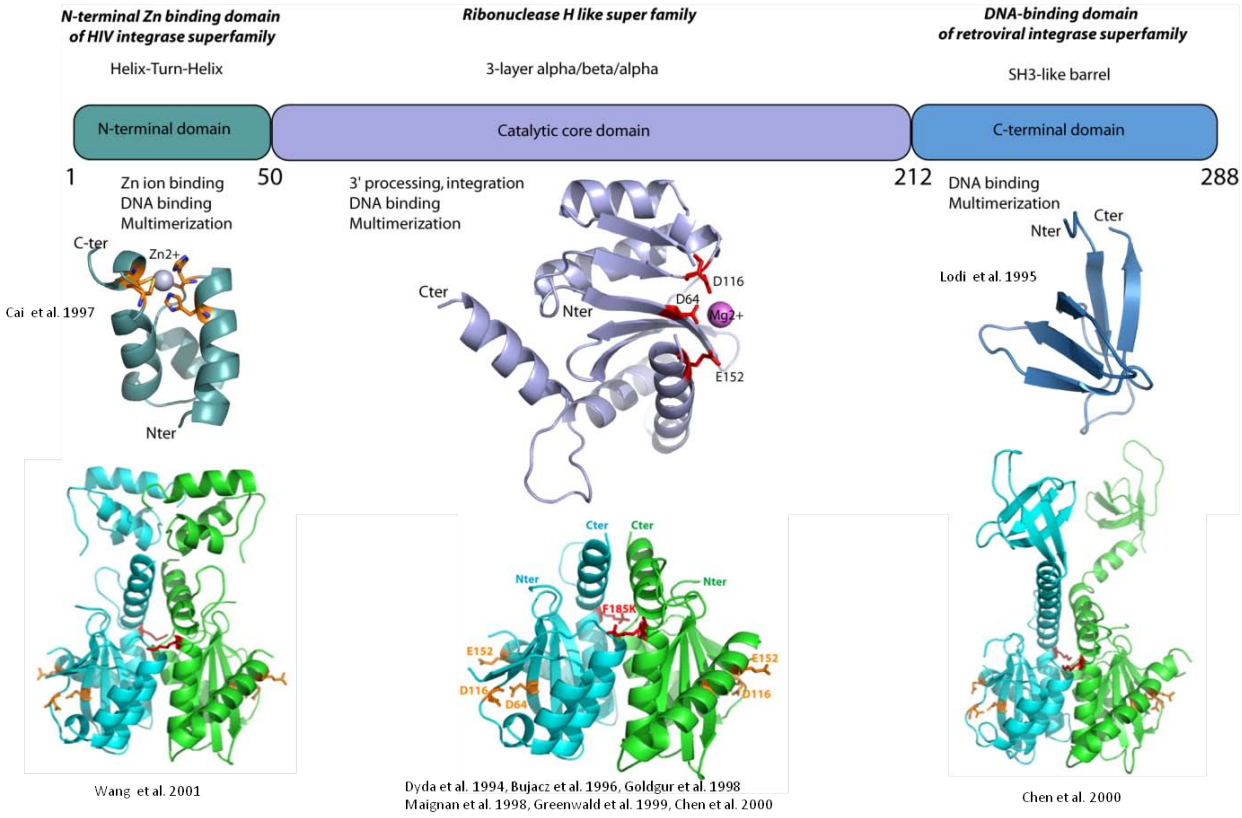


Figure 9: Domain Organization of HIV-1 Integrase showing structures of individual domains

Functions of HIV Integrase

Two reactions are necessary for the covalent integration of viral DNA into host DNA. The full-length integrase is essential for catalyzing the 3' processing and strand transfer reactions (Bushman, Fujiwara et al. 1990). No energy co-factor is required for either reaction to occur.

In addition to its primary roles in integration, integrase may play a role in other steps of the replication cycle, as mutations in integrase affect other steps beyond integration. For example, viruses with mutated integrase display lower levels of reverse transcription and viral particles

with abnormal morphology (Engelman, Englund et al. 1995). Moreover, the use of IN/LEDGF allosteric inhibitors suggests a role for integrase in viral maturation (Le Rouzic, Bonnard et al. 2013)

3’ processing

The pre-integrated viral DNA exists as a linear double stranded DNA fragment. In the cytoplasm, integrase binds to a long terminal repeat sequence on the 5’ and 3’ end, and catalyzes an endonucleolytic cleavage, removing a GT di-nucleotide corresponding to the U5 and U3 end of the viral DNA (Engelman, Mizuuchi et al. 1991). The reaction is a trans-esterification where the phosphodiester bond on the viral DNA is broken by nucleophilic attack. This reaction exposes 3’-OH groups, forming a target capture complex (TCC) that will be covalently joined to target DNA in the strand transfer reaction (Lesbats, Engelman et al. 2016). Divalent cations such as Mn^{2+} and Mg^{2+} are essential for both 3’ processing and strand transfer reactions, with Mg^{2+} being the preferred cation due to better cleavage specificity than Mn^{2+}, with water as the nucleophilic agent (Engelman, Englund et al. 1995).

Strand transfer

Sequentially, the cleaved DNA is inserted into the target DNA during the strand transfer reaction. A nucleophilic attack on the target DNA is catalyzed by integrase, using the Asp and Glu residues of the D, D-35, E motif of the catalytic core domain motif to coordinate two divalent metal ions, thereby activating the 3’-OH groups from the viral DNA to attack the phosphodiester bond in the target DNA (Bushman, Engelman et al. 1993). The exposed 3’-OH groups from the viral DNA are joined to the phosphate ends of the target DNA, allowing integration of the viral DNA and forming the strand transfer complex (STC) (Lesbats, Engelman et al. 2016). Cellular repair proteins such as DNA polymerases and DNA ligases repair both ends (Yoder and Bushman 2000). Integrated viral DNA is then replicated along with cellular DNA.

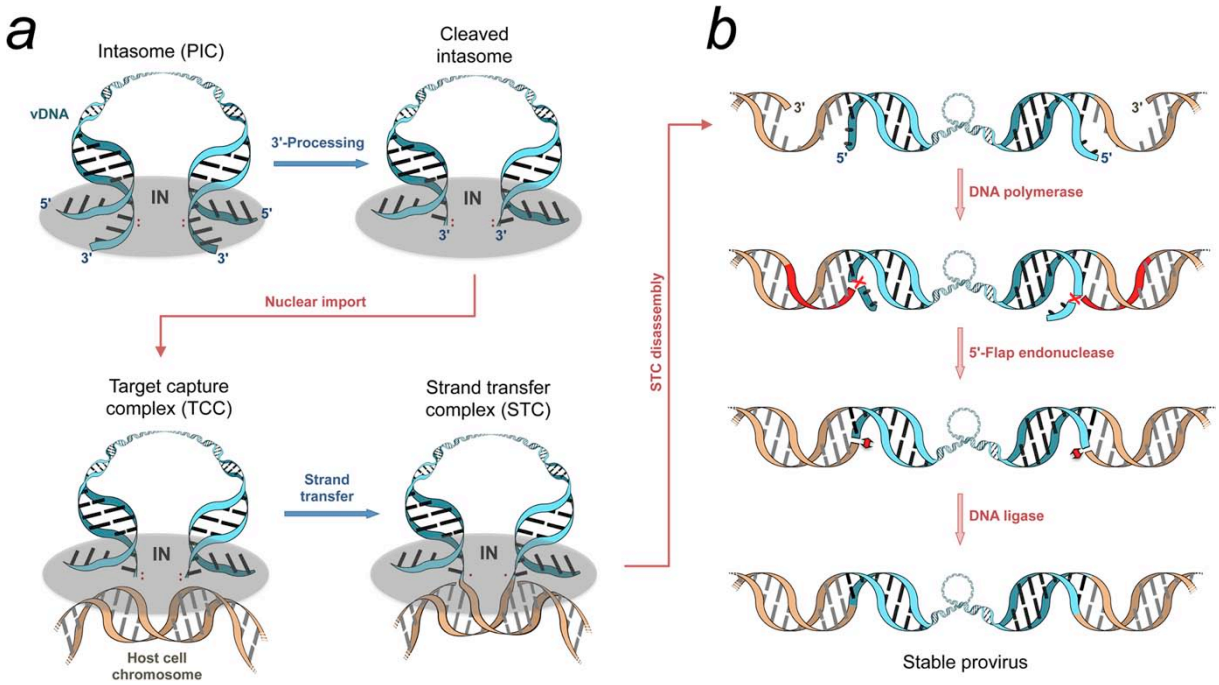


Figure 10: Schematic representation of the 3' processing and strand transfer activities . *Lesbats et al; Chemical Reviews, 2016*

Disintegration

Disintegration is a process that is considered to be the opposite of the strand transfer reaction, whereby integrase reverses the DNA cleavage and ligation reaction. This reaction can be catalyzed by catalytic core domain only or with truncated versions of integrase (IN1-212 or IN52-288) (Chow, Vincent et al. 1992). This reaction has only been observed *in vitro* and there is currently no experimental evidence for this process *in vivo*.

HIV Integrase domains

N-terminal Domain

The N-terminal domain (1-49) contains an HHCC (H12, H16, C40 et C43) motif that resembles a zinc finger, and acts as a zinc-binding domain, with zinc being required for the folding of the isolated N-terminal domain (Bushman, Engelman et al. 1993). Additionally, zinc promotes the multimerization of integrase, which is needed in order to be catalytically active (Zheng, Jenkins et al. 1996). The N-terminal domain of integrase exists as a dimer in solution, with each monomer consisting of four helices (Cai, Zheng et al. 1997) . In the upper

part of the monomer, the HHCC motif coordinates the zinc ion, while a hydrophobic core formed by helices 1, 2 and 3 stabilizes the lower part.

Catalytic Core Domain

The catalytic core domain (50-212) contains the active site of integrase, and contains a D, D-35, E catalytic triad (D64, D116, E152) that is essential for the catalytic activity of the protein (Bushman, Engelman et al. 1993). D64 and D116 coordinate the binding of metallic cofactor (Mg^{2+} or Mn^{2+}), which is required for activity. Mutating any of these residues eliminates enzyme activity. This motif is conserved among all retroviral and retrotransposon integrase proteins (Dyda, Hickman et al. 1994). This domain also contains amino acids that are essential for interactions with viral DNA: Q148, K156 and K159, Y143. Mutating K156 and K159 eliminates the specific interactions with DNA (Jenkins, Esposito et al. 1997). A F185K mutation improves solubility of the catalytic core domain for structural studies (Jenkins, Hickman et al. 1995). The catalytic core domain has a topology that resembles that of ribonuclease H, consisting of a central five-strand β sheet and six helices (from α1 to α6). There is also a disordered loop from residues 141-153. Mutations in the loop G149A and G140A/G149A eliminates catalytic activity in the catalytic core domain by stiffening the loop without interfering with DNA binding (Greenwald, Le et al. 1999).

C-terminal Domain

The C-terminal domain (213 to 288) is the least conserved domain and has been shown to bind viral DNA and non-specific DNA. Triple mutants of the C-terminal domain of HIV-2 integrase (R262D, R263V, K264E) are defective in DNA binding (Lutzke, Vink et al. 1994). Specifically, K264E mutants show reduced binding to DNA. R263K mutants confer low-level resistance to Dolutegravir (Quashie, Mesplède et al. 2012). Additionally, these triple mutants are unable to cleave viral DNA. The structure of the C-terminal domain consists of five β-strands that form 2 anti-parallel β-sheets, with one helical turn between the fourth and fifth β strand (Eijkelenboom, Puras Lutzke et al. 1995, Lodi, Ernst et al. 1995). The fold of the C-terminal domain is similar to the SH3 domains (Src-Homology 3), which display a β-barrel fold consisting of 5 or 6 anti-parallel β-strands. SH3 domains have been shown to be present in proteins involved in protein-protein interactions important for intracellular signaling pathways and DNA binding (Weng, Rickles et al. 1995).

Structure of two domains

Chen et al. solved a two-domain structure of the CCD and CTD (52-288) with C56S, W131D, F139D, F185K, and C280S mutations in order to improve solubility (Chen, Krucinski et al. 2000). Wang et al. solved the structure of the NTD and the CCD (1-212) with W131D, F139D and F185K (Wang, Ling et al. 2001). The IN52-288 crystal structure forms a symmetric dimer where each monomer of the CCD is linked to the CTD by residues 195-220 in helix α6. Similarly, the IN1-212 exists as a dimer, with the interface being mediated by the side chains of R20, K34, Q209, T206 and E212. Unlike the structure of IN1-212 that is compact, the structure of IN52-288 is extended, with the CCD being globular, and the CTD extending away from the catalytic core.

Intrinsic flexibility of Integrase

In order to carry out its diverse functions, Integrase needs to adopt several conformations that allow it to interact with multiple partners in multiple steps of the virus life cycle. Thus, it is a highly flexible protein, whose conformation changes as a function of partner protein, in relation to biological functions being performed. A superposition of the catalytic core domain of structure of integrase from various viruses shows variability between domain organization between viral species (Maillot, Lévy et al. 2013).

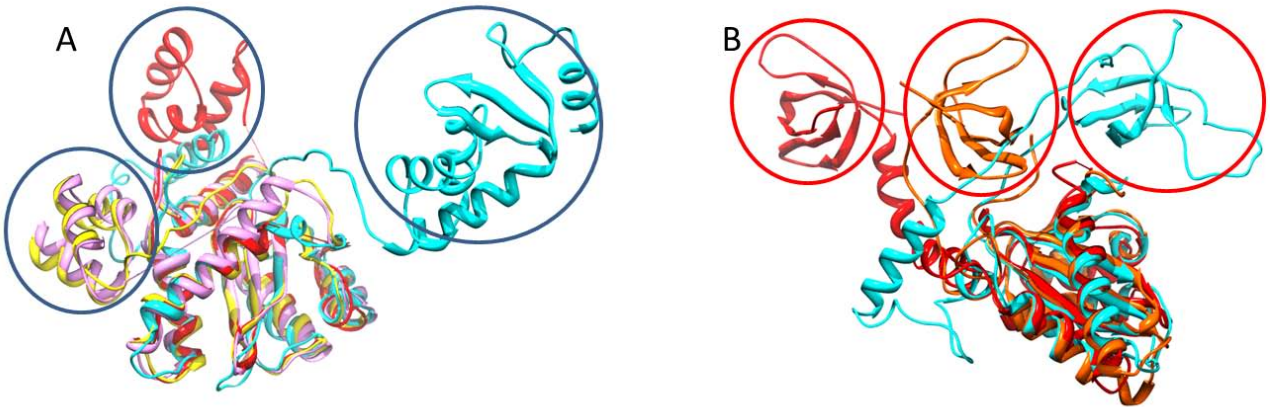


Figure 11: Superposition highlighting IN flexibility. A: Overlay of IN catalytic core domain of MMV (pink), PFV (cyan), HIV-2 (yellow) and HIV-1 (red). N-terminal domains are highlighted in blue. B: Overlay of IN catalytic core domain of RSV (orange), HIV-1 (red) and PFV (cyan). C-terminal domains are highlighted in red. *Maillot et al; PLOS One, 2013*

HIV-1 Intasome structure

Recent advancements in cryoEM have made it possible to solve the structure of the full-length strand transfer complex, with integrase in complex with viral DNA and target DNA (intasome) (Passos, Li et al. 2017). Fusing the DNA binding protein Sso7d to the N-terminal of integrase improved solubility without interfering with activity *in vivo*, and Sso7d-IN was used to assemble intasomes for cryo-EM studies. The HIV intasome consists of a dimer of dimers with four IN protomers (one NTD, one CCD, one CTD each) arranged with two-fold symmetry around the target and viral DNA. A newly identified residue (K46), and previously identified residues (K156, K159 and K160) were confirmed to play roles in viral DNA binding, sequence specificity and catalysis (Jenkins, Esposito et al. 1997, Chen, Weber et al. 2006, Krishnan, Li et al. 2010). Additionally, R231 interacts strongly with target DNA and viral DNA (Serrao, Krishnan et al. 2014). The presence of the host co-factors such as the Integrase binding domain of LEDGF results in the formation of higher order assemblies, where the CTDs are reorganized to engage viral DNA, highlighting the structural flexibility of integrase. Notably, in the CTD, residues L242, I257, and V259 are involved in the formation of CTD-CTD interface. Other residues such as K14, E35, K240, K244 and R269 were also important for higher order oligomers. Mutating residues in the CTD affected strand transfer reactions and virus replication. This data suggests the importance of higher order oligomers for integrase function.

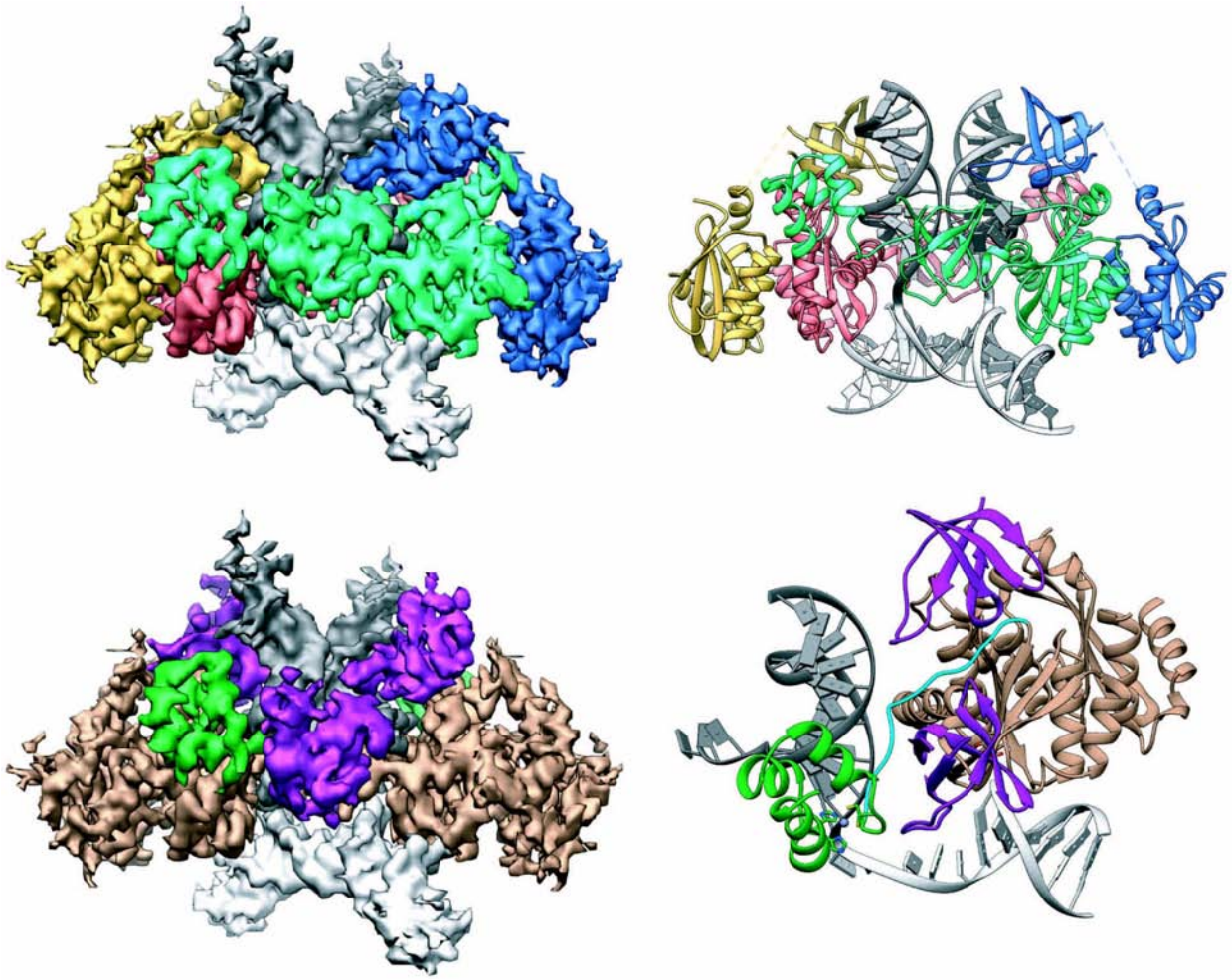


Figure 12: Structure of the STC intasome. Top left: Cryo-EM reconstruction of the STC, showing IN protomers: inner protomers (green and red), outer protomers (yellow and blue), viral DNA (dark grey) and target DNA (light grey). Top right: Atomic model derived from the cryo-EM density. Bottom left: Segmented cryo-EM density showing IN domains: NTD (green), CCD (beige) NTD-CCD linker (blue) CTD (purple), viral DNA (dark grey), target DNA (light grey) Bottom right: Asymmetric subunit of the atomic model, using the same color scheme as in the bottom left. *Passos et al, Science, 2017*

The role of LEDGF in HIV Integration

Several host factors have been shown to be involved in the integration process. One of the most widely studied co-factors is Lens Epithelium Derived Growth Factor (LEDGF/p75). LEDGF/p75 is a 64kDa transcriptional co-activator that belongs to the hepatome derived growth factor (HDGF) related protein (HRP) family (Baid, Upadhyay et al. 2013). It is made up of 550 amino acids and contains two small structural domains. The N-terminal PWWP domain (1-91), along with the nuclear localization signal (178-197) (Maertens, Cherepanov et al. 2004) and a pair of AT-hook motifs (178-197) are responsible for the association of LEDGF to the chromatin (Llano, Vanegas et al. 2006, Turlure, Maertens et al. 2006, Shun, Botbol et al. 2008). The C-terminal (347-429) Integrase Binding Domain (IBD) mediates the

IN/LEDGF interaction (Cherepanov, Devroe et al. 2004, Llano, Saenz et al. 2006). LEDGF/p75 was shown to stimulate the catalytic activities of integrase *in vitro* (Cherepanov, Maertens et al. 2003). HIV-1 infectivity of RNAi LEDGF knockdown cells is severely depleted, due a lack of IN-chromatin association (Llano, Vanegas et al. 2004). The IN/LEDGF interaction has been shown to be specific to lentiviruses only (Busschots, Vercammen et al. 2005, Cherepanov 2007).

The LEDGF-IBD is composed of four long α-helices (α1, α2, α4 and α5) as revealed by NMR spectroscopy (Cherepanov, Sun et al. 2005). A structure of the IN-CCD (F185K)–LEDGF IBD complex has been solved (Cherepanov, Ambrosio et al. 2005). Residues I365, F406 and V408 were shown to be involved in the interaction with IN. D336N is essential for the interaction with IN, as well as for the enzymatic activity of LEDGF. The IN-CCD contains residues involved in the LEDGF interaction (V165, R166, Q168, L172 and K173), and residues directly interacting with LEDGF (A128, A129, W131 and W132). The IN-NTD is important for high affinity IN/LEDGF interactions (Maertens, Cherepanov et al. 2003).

LEDGF has been implicated in the stabilization of the functional tetramer of IN (Cherepanov, Maertens et al. 2003), as well as in 3' processing and strand transfer (Michel, Crucifix et al. 2009). Furthermore, LEDGF protects integrase from proteasomal degradation (Llano, Delgado et al. 2004). Overall, the model for the role of LEDGF in integration postulates that the N-terminal region of LEDGF binds to host chromatin at active transcription sites, and interacts with the PIC through its IBD, allowing to PIC to integrate into transcriptionally active sites, while also stimulating strand transfer (Ciuffi, Llano et al. 2005).

The Ruff team solved the structure of the full-length integrase in complex with LEDGF, and DNA by cryoEM, and proposed a binding mechanism for the IN/LEDGF/DNA complex (Michel, Crucifix et al. 2009). In this model, the IN/LEDGF complex contains 4 integrase and 2 LEDGF molecules, supporting the evidence that IN tetramer is the basic functional unit for integration.

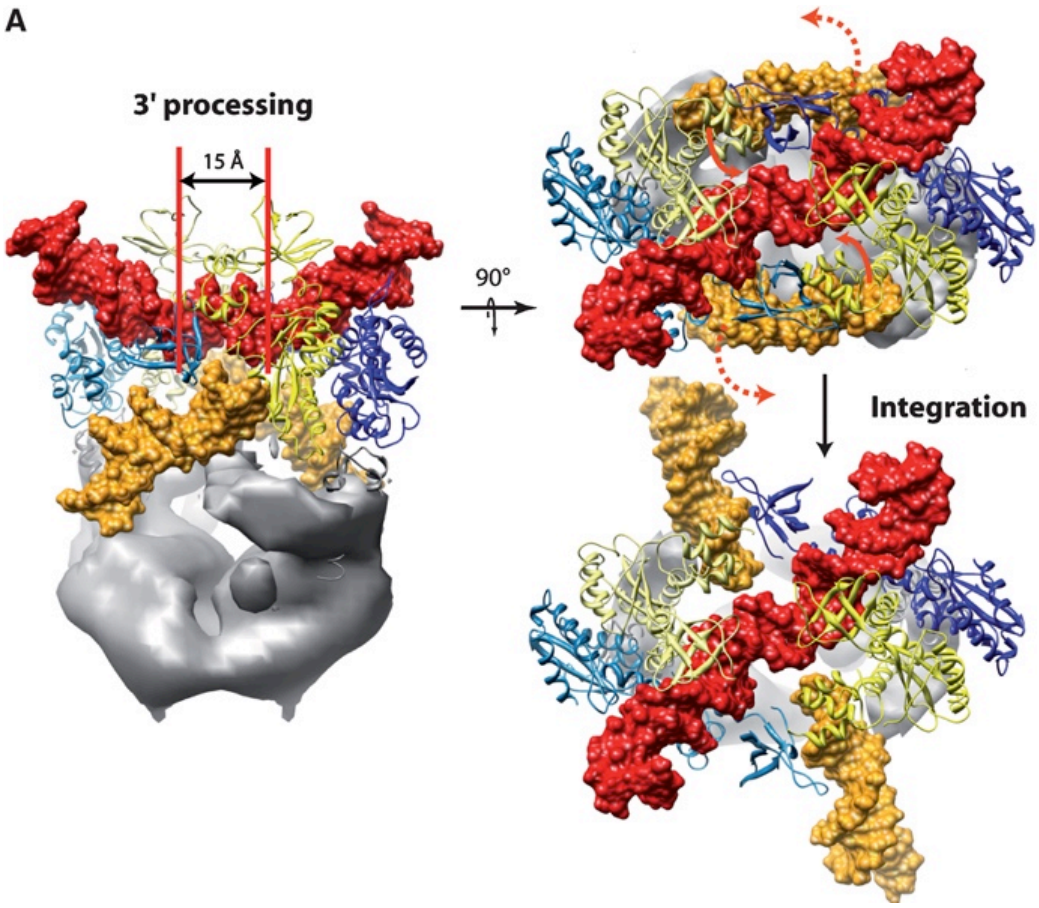


Figure 13: Model for IN/LEDGF interaction in DNA Integration 15Å distance between viral DNA (yellow) and target DNA (red). Red arrows indicate conformational changes in viral DNA for integration to occur. *Michel et al; The EMBO Journal, 2009*

Important factors for Integration Site Selectivity

HIV-1 and other retroviruses preferentially integrate into transcriptionally active sites (Schröder, Shinn et al. 2002, Mitchell, Beitzel et al. 2004). However, HIV-1 can infect both dividing and non-dividing cells (Lewis, Hensel et al. 1992). The ability of HIV-1 and other retroviruses to infect non-dividing cells in the G0 stage makes non-dividing macrophages an important reservoir for HIV-1 in people living with HIV.

HIV enters the nuclear envelope through the nuclear pore complex (NPC) (Bukrinsky, Sharova et al. 1992). Capsid (CA) has been shown to mediate this entry by interacting with NPC components such as Polyadenylation Specificity Factor 6 (CPSF6) (Price, Fletcher et al. 2012, Sowd, Serrao et al. 2016) and Nucleoporins (NUPs) (Ocwieja, Brady et al. 2011, Matreyek, Yücel et al. 2013). The nuclear entry route is the first determinant factor for HIV-1 integration. HIV-1 PIC has been shown to integrate into areas of euchromatin closer to the

nuclear periphery (Di Primio, Quercioli et al. 2013), and HIV-1 integration targets are closely associated with the nuclear pore complex (Marini, Kertesz-Farkas et al. 2015), suggesting that HIV-1 integrates into the chromatin regions it encounters immediately after nuclear translocation.

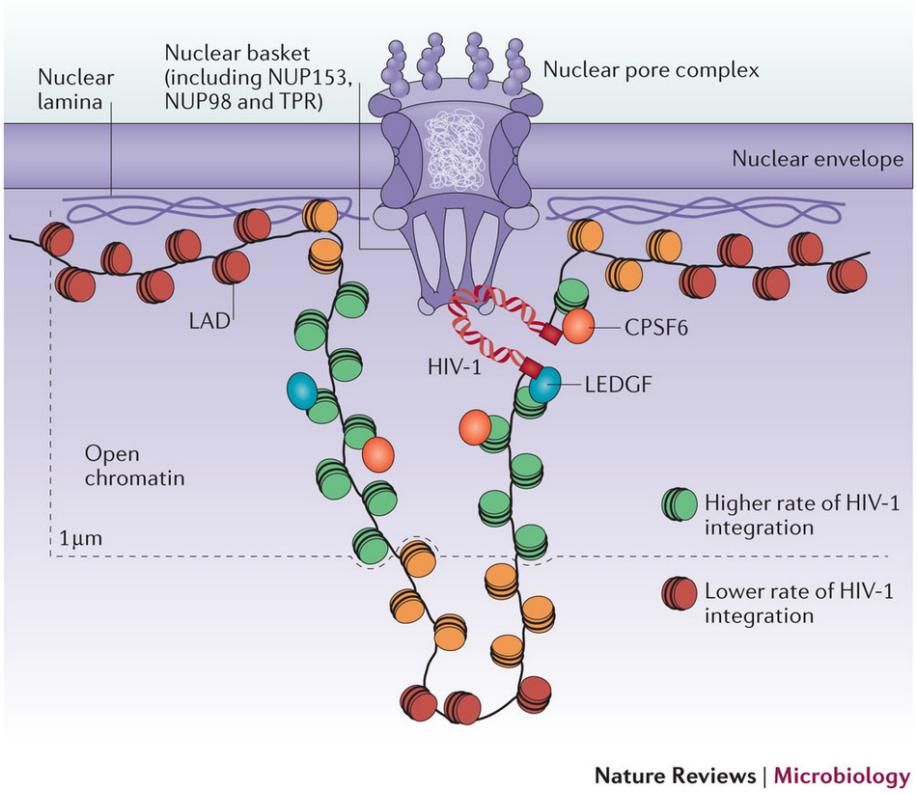


Figure 14: Model of HIV-1 Integration at the Nuclear Pore Complex . *Lusic and Silicano. Nature Reviews Microbiology, 2016*

Additionally, the IN/LEDGF interaction has been shown to be important for targeting transcriptionally active sites, and recognizing transcription associated histone modifications. LEDGF interacts with DNA and tri-methylated H3K36 histones via its PWWP domain (Pradeepa, Sutherland et al. 2012, Eidahl, Crowe et al. 2013, van Nuland, van Schaik et al. 2013), allowing integrase to target actively transcribed genes in this histone mark. HIV integration in LEDGF mutants is decreased, and shifted away from transcriptionally active sites (Ciuffi, Llano et al. 2005, Shun, Raghavendra et al. 2007). In contrast to HIV-1, Gamma retroviruses like MLV use the Bromodomain and Extra terminal domain (BET) proteins for

chromatin targeting and integration near transcription start sites (De Rijck, de Kogel et al. 2013, Gupta, Maetzig et al. 2013, Sharma, Larue et al. 2013)

Furthermore, there is evidence that severely kinked DNA, such as nucleosomal DNA is preferred for integration. Structures of the PFV intasome suggest that there are interactions between the C-terminal domain of the intasome and H2A/H2B nucleosome (Maskell, Renault et al. 2015). PFV intasomes are able to catalyze integration into mononucleosomes. Studies performed using DNA mini-circles to mimic curved nucleosomal DNA show HIV-1 integrase preferentially targets curved DNA, compared to linear DNA (Pasi, Mornico et al. 2016). In addition to nucleosomal target, physical properties of target DNA have been shown to be important for integration, where factors such as the energy required to fit the DNA into the CCD, and DNA wrapping around a nucleosome determine integration sites (Naughtin, Haftek-Terreau et al. 2015). Moreover, intasome architecture and compactness of the chromatin surrounding the nucleosome target are determinants for integration site selectivity (Benleulmi, Matysiak et al. 2015). While PFV and MLV integrate into dense and stable nucleosomes, HIV-1 preferentially integrates into sites with low nucleosome occupancy.

Proposed role of HIV-1 Integrase CTD in Chromatin Association

While cellular co-factors such as LEDGF and BET have been shown to be important for nucleosome targeting, these co-factors are not the only determinants for cellular integration. It is evident that additional interactions between the intasome and nucleosomal DNA need to be further elucidated in order to fully understand HIV-1 integration.

According to recent work done by our collaborators (Vincent Parissi, Bordeaux), another important factor for HIV-1 nucleosomal binding and chromatin integration is histone tails. N-terminal histone tails have been shown to be important for IN interactions with mononucleosomes *in vivo*. Using histone peptide array, it was shown that there is a specific interaction between HIV-1 integrase and the monomethylated N-terminal tail of histone H4. Structurally, the C-terminal domain displays a β-barrel fold, similar to the SH3 fold. This SH3 fold is consistent with members of the Royal Domain superfamily that includes chromo, Tudor, MBP, PWWP and Agenet domains (Chen, Nott et al. 2011). The structures of all members of this superfamily consist of anti-parallel β-strands that form a β-barrel with an aromatic cage located in the groove of the barrel. Members of this superfamily are involved in

recognizing and binding arginine or lysine-methylated ligands. The aromatic cage serves as the binding site for these proteins by interacting with the methylated side chain using electrostatic and hydrophobic contacts. Specifically, Tudor, Chromo, PWWP and MBT domains bind to methylated histone tails (Kim, Daniel et al. 2006, Yap and Zhou 2010).

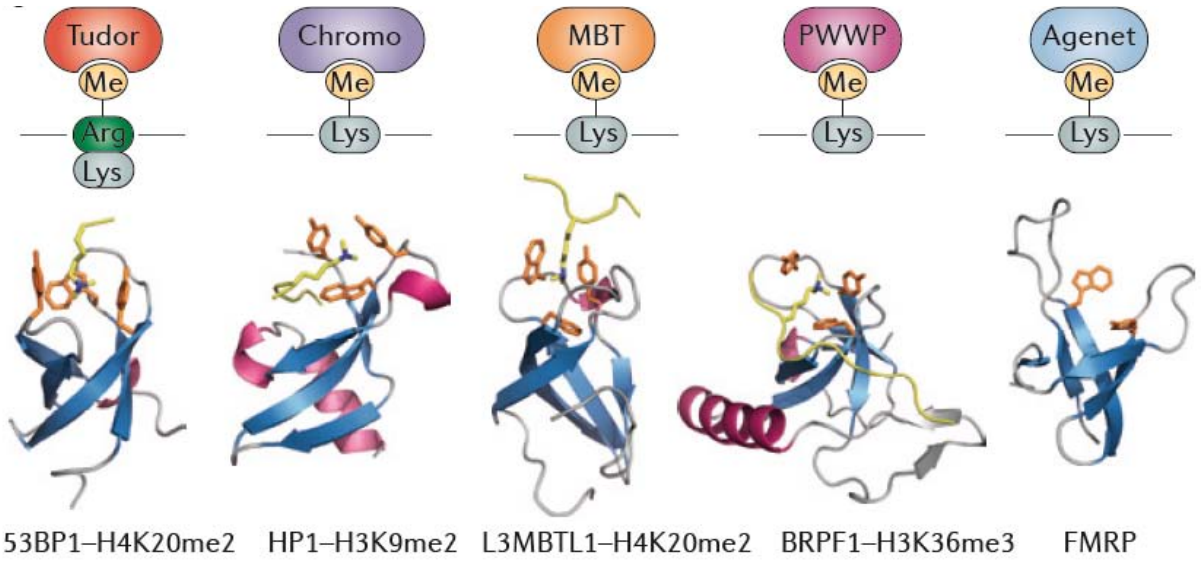


Figure 15: Members of the Royal Domain Superfamily. *Chen et al; Nature Reviews, Molecular Cell Biology, 2011*

Although the CTD of HIV-1 Integrase most closely resembles the Tudor domains, it lacks the aromatic cage consistent with Tudor domains. However, the CTD of HIV-1 Integrase probably functions as a chromodomain *in vivo* to mediate integrase/chromatin associations. Indeed, using the far dot blot assay, our collaborators show that the CTD is responsible for the interactions between HIV-1 integrase and histone H4 mono-methylated at lysine 20 (H4K20).

Additionally, structures of the HIV-1 and PFV integrase (Maskell, Renault et al. 2015, Passos, Li et al. 2017) suggest that contacts between the CTD of the intasomes and target DNA are essential for integration. In the HIV-1 intasome, Residue R231 has been shown to be important for binding to target DNA. The intasome structure also suggests that the CTD is involved in higher order oligomerization that is essential for integrase function.

Genome diversity in HIV

HIV-1 reverse transcriptase lacks a proofreading activity, leading to high mutation rates of about 9.3×10^{-5} mutations per base pair in plasma virus (Cuevas, Geller et al. 2015). There is

high genetic diversity between HIV groups and subtypes. Sequence variation within a subtype in Group M can be as high as 30%, while variation between subtypes in Group M can be as high as 42% (Abecasis, Vandamme et al. 2009), with the highest diversity found in group O. There is also a possibility for superinfection, especially in African regions with different virus variations circulating. This can happen due to simultaneous or sequential infection by two different strains (Piantadosi, Chohan et al. 2007, Powell, Urbanski et al. 2009).

The extensive diversity of the HIV genome has several clinical and biological implications. HIV diversity poses a challenge for HIV diagnosis, especially in resource-limited settings like Africa, where new strains develop faster, making commercial test kits unreliable for HIV diagnosis and sometimes leading to incorrect diagnosis (Aghokeng, Mpoudi-Ngole et al. 2009). Additionally, there is a difference in susceptibility to antiretrovirals between subtypes, leading to differences in resistance pathways (Lessells, Katzenstein et al. 2012) . Several studies in East Africa have shown that some subtypes (subtype D) are more aggressive and show faster disease progression to death when compared to subtype A (Baeten, Chohan et al. 2007, Ssemwanga, Nsubuga et al. 2013). Most importantly, HIV genetic diversity poses a big challenge to HIV vaccine development, making it difficult to design broadly neutralizing antibodies against HIV (Nickle, Rolland et al. 2007, Stephenson, D'Couto et al. 2016).

The high genetic flexibility and diversity displayed by HIV leads to high antigenic variation. HIV-1 mutates in order to adapt to selective pressure from drugs and the immune system. However, it has been shown that these mutations can be detrimental to the virus. Mutations in response to reverse transcriptase inhibitors and protease inhibitors involve changing residues at the active site, causing reduced enzyme efficiency and reduce viral fitness (Mesplede, Quashie et al. 2013, Hu and Kuritzkes 2014). HIV-1 adapts to immune pressures by HLA-associated selection, allowing the virus to avoid recognition by cytotoxic T-lymphocytes (Carlson, Le et al. 2015). These mutations can also reduce viral fitness, and may not be transmitted into a new host (Boutwell, Rowley et al. 2009, Boutwell, Carlson et al. 2013). HIV-1 undergoes compensatory mutations in order to maintain infectivity and fitness. In the case of drug-induced mutations at active sites, amino acid substitutions are made at neighboring sites to improve enzyme activity (Cong, Heneine et al. 2007). However, these mutations do not fully restore viral infectivity, and transmission rates of resistant viruses are much lower.

HIV-1 Co-Evolution

Intra or inter-subtype/group recombination plays in a big role in genetic diversity of HIV-1 and its ability to adapt to selective pressure. Recombination results from template switching during reverse transcription when two copies of RNA from different genomes are present in the same cell. Recombination is not a random event (Smyth, Schlub et al. 2014), as recombination events tend to occur in sites that confer a selective advantage to the virus. According to Archer et al, local sequence identity is an important factor for determining recombination (Archer, Pinney et al. 2008). Additionally, RNA structures are also important for facilitating recombination (Simon-Loriere, Martin et al. 2010). Particularly, HIV-1 prefers to recombine around RNA sequences that encode for inter-protein linkages, suggesting that recombination shifts genes or gene fragments that translate into individual protein domains, thereby increasing the chances of adaptive evolution.

Proposed Importance of CTD in Co-Evolution

Our collaborators (Matteo Negroni, Strasbourg) produced inter-group chimeric constructs between the groups M (subtype A) and O and analyzed their reverse transcription and integration efficiency. By systemically replacing divergent amino acids between subtype A2 and group O, they found residues in the C-terminal domain of integrase that displayed decreased integration activity. Specifically, they found a $N_{222}K_{240}N_{254}K_{273}$ motif in subtype A2 that is replaced by $K_{222}Q_{240}K_{254}Q_{273}$ in group O which appears to be important for IN functionality. Further investigations are needed to understand the importance of this non-conserved motif for IN function and structure.

Project Objectives

Previously, the IN-CTD was the least studied domain of HIV-1 integrase. However, recent studies by our collaborators and others show that it is an equally important domain in several functions of HIV-1 integrase. Therefore, the overall aim of my thesis was to understand interactions in the C-terminal domain of HIV-1 integrase, studying its roles in chromatin integration and co-evolution.

The first project focused on understanding the role of the CTD in the IN/H4K20Me1 interactions from the structural perspective. The main goal was to identify the residues in the CTD of HIV-1 integrase that mediate the interactions with H4K20Me1, as well as observe

any conformational changes in protein structure that may occur upon histone interaction. This information would be essential in order to elucidate the mechanisms of nucleosome association and interaction, providing further insights into the integration process. In the long run, details obtained about these interactions can be used to design novel HIV-1 inhibitors. During the course of this project, I carried out the following processes:

- Solubilized and optimized the purification protocol of isolated IN-CTD constructs in Gateway vectors
- Confirmed the IN-CTD/H4K20Me1interaction using purified IN-CTD and histone peptides
- Perform preliminary structural analysis on IN-CTD and histone peptide using NMR
- Cloned the IN-CTD into pET15b expression vectors to optimize protein expression, solubility and folding
- Solved X-ray structures of IN-CTD
- Obtained crystals and collected x-ray data in the presence of peptide
- Solved NMR structure of IN-CTD alone, and performed titrations in the presence of peptide to determine conformational changes triggered by peptide binding

The overall goal of the second project was to understand co-evolutionary changes in integrase. Since many mutations were observed in the CTD, we aimed to study structural differences in the IN-CTD between groups and subtypes, in order to understand differences observed in activity tests. Specifically, we hoped to gain structural insights into the importance of the $N_{222}K_{240}N_{254}K_{273}$ motif. To achieve the aims of this project, I carried out the following processes:

- Purified IN- CTD from selected WT and mutant constructs
- Obtained crystals and solved the structure of several of the selected WT and mutant protein

CHAPTER 2 – HISTONE ASSOCIATION STUDIES

MATERIALS AND METHODS

DNA cloning

Proteins used in this project were cloned using two different methods. The platform at the IGBMC used the Gateway® strategy to produce several pDEST expression vectors from pENTR vectors with different purification and solubility tags (6X His, GST, Strep). Secondly, restriction enzymes were used to clone protein of interest into the pET15b vector.

Gateway cloning

Background

The Gateway cloning strategy designed by Invitrogen is based on the site-specific recombination properties of lambda bacteriophage (Invitrogen 2003). The major advantage of this system is that it provides a way to clone a DNA sequence from one vector into multiple destination vectors. Lambda recombination occurs between *att*P on the lambda phages and *att*B sites on the E.coli chromosome. Subsequently, recombination between *att*P and *att*B sites gives rise to *att*R and *att*L sites.

Gateway cloning occurs in three main steps. First the gene of interest is amplified by PCR, and flanked by *att*B sites for recombination into a pDONR vector. Subsequently, BP reaction is used to create an entry vector (pENTR) and finally, LR reaction is used to transfer the entry vector into several destination vectors (pDEST) as needed. The *ccd*B gene is used for negative selection following recombination and transformation. The CcdB protein inhibits growth of most bacterial strains such as DH5α by interfering with DNA gyrase. Upon recombination, the gene of interest replaces the *ccd*B gene. In the absence of recombination, cells that take up the un-recombined vectors will not grow.

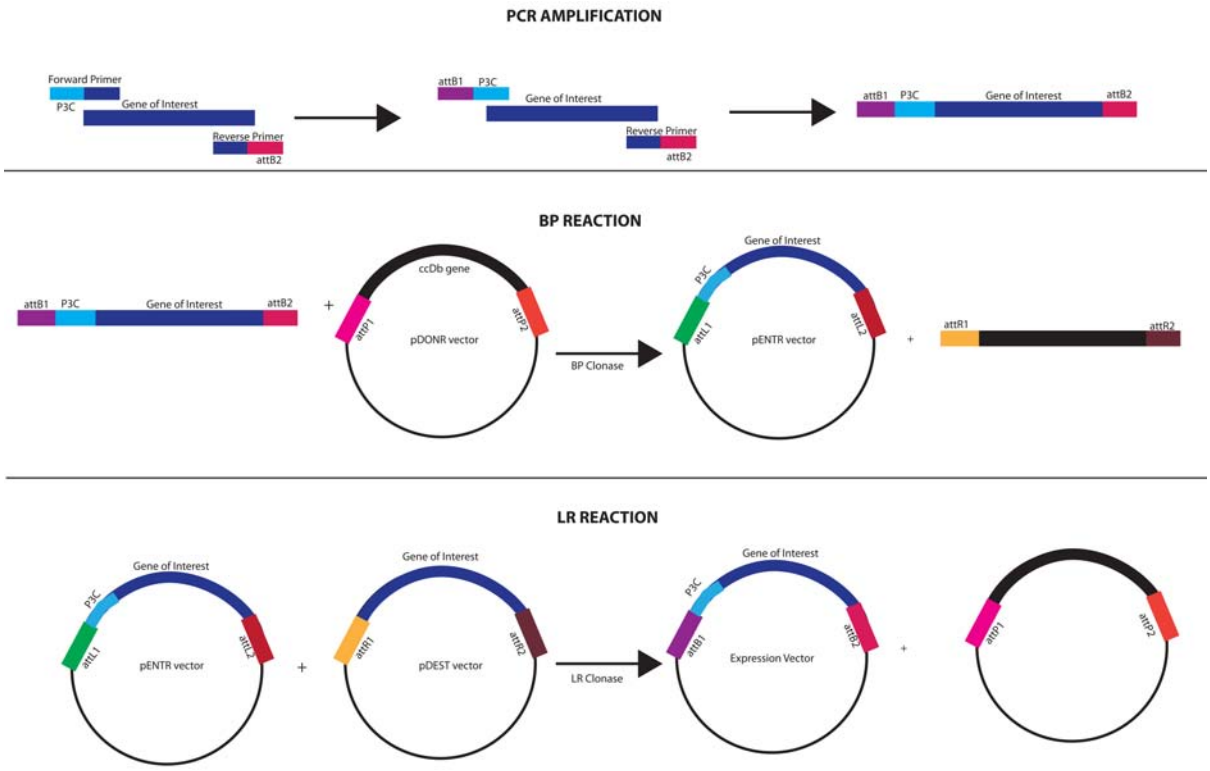


Figure 16: Gateway Cloning Strategy

Protocol

The IGBMC Molecular Biology Service carried out the cloning of all constructs generated using the Gateway strategy. To clone the C-terminal Domain of HIV Integrase using the Gateway strategy, three sets of primers were used:

P3211-F: cttgaagtcctctttcagggacccATTCAAAATTTTCGGGTTTATTACAGG

P3210-R: ggggaccactttgtacaagaaagctgggtctcAATCTTCGTCCTGTCTACTTGC

AttB1-3C**:** ggggacaagtttgtacaaaaaagcaggcttcCTTGAAGTCCTCTTTCAGGGACCC

Round 1 PCR: The first round of PCR was used to amplify the CTD from residues 220 to 288 using the P3211-F as the forward primer with the underlined bases annealing to the start of the integrase sequence, while the non-underlined bases code for the 3C protease-binding site. The reverse primer used was P3210-R, with the underlined bases annealing to the end of the integrase sequence, while the non-underlined bases code for the *att*B2 sequence.

Round 2 PCR: With the products from the first reaction, a second round of PCR was performed using attB1-P3C as the forward primer with the underlined bases annealing to the

3C protease binding site, while the non-underlined bases code for the attB1 sequence. P3210-R was used as the reverse primer to add the attB1 to the sequence.

BP reaction: The PCR product from the previous step was purified, and a BP reaction was carried out to insert the AttB1-3C-INT220-288-AttB2 fragment into the pDONR207 plasmid to generate an entry clone E1441. Following transformation into DH5α, PCR was performed to verify the success of the BP reaction, and PCR fragments were sent to GATC for sequencing.

LR reaction: An LR reaction between E1441 and destination vectors (pHGWA, PDEST15) was used to generate expression vectors.

Restriction - Ligation Cloning

Background

Restriction-Ligation cloning was used to produce additional constructs of the integrase c-terminal domain. With the use of restriction enzymes, DNA fragments are generated that can be inserted into any vector of choice with complementary ends. Three main steps are required. First the gene of interest (insert) is amplified by PCR, flanked by restriction sites that can be cut by specific restriction enzymes. The vector and the insert both contain complementary restriction enzyme sites that are digested in the second step. To avoid self-ligation after digestion, the vector is typically treated with alkaline phosphatase. Subsequently, the insert is ligated into the linearized vector by DNA ligase, which catalyzes the formation of phosphodiester bonds between adjacent 5'-phosphate and 3'-hydroxyl residues. The success of the ligation is dependent on factors like the vector: insert ratio, and DNA concentration.

Protocol

Constructs with P3C cleavage site

For constructs with P3C cleavage sites, two rounds of PCR were performed with Phusion DNA polymerase (New England BioLabs) with the PE603 plasmid as the template.

Round 1 PCR: The first round added the P3C sequence on the N-terminus and the BamHI cleavage site on the C-terminus, generating P3C_220-288_StopBamHI and P3C_220-270_StopBamHI constructs respectively. Each reaction mix contained 40.75µl water, 5µl

buffer, 0.75µl polymerase, 0.5µl dNTPs (dATP, dCTP, dGTP, dTTP) mix, 1µl of PE603 template DNA (10ng/µl) and 1µl each of forward and reverse primers for a total reaction volume of 50µl. The PCR reaction cycle consisted of a denaturation step at 94° for 2 minutes, followed by 10 cycles consisting of thermal denaturation at 94° for 15 seconds, followed by primer annealing at 52°C for 30 seconds, and an elongation step for 45 seconds at 72°C. This was followed by 20 cycles consisting of consisting of thermal denaturation at 94° for 15 seconds, followed by primer annealing at 52°C for 30 seconds, and an elongation step for 45 seconds at 72°C, and a final elongation step at 72°C for 7 minutes. To confirm the success of the PCR, agarose gel electrophoresis with 10µl of PCR product was loaded on a 2% (w/v) agarose gel in 100ml TAE buffer containing 5µl Ethidium Bromide (10mg/ml) was performed to analyze products.

PCR clean up: PCR products obtained from the first round of PCR were purified using the PCR clean-up kit from Machery Nagel. 1 volume of PCR product was mixed with 2 volumes of buffer, which contains a chaotropic salt such as guanidinium thiocyanate, allowing the DNA to bind to the silica membrane during centrifugation at 11000xg for 30 seconds. 700µl of a buffer containing ethanol was used to wash the silica membrane by centrifugation at 11000xg for 1 minute. Another step of centrifugation was performed to dry the silica membrane and ensure removal of leftover ethanol. 30µl of ultra- pure water was added to the silica membrane, incubated at room temperature for 1 minute, and centrifuged for 1minute at 11000xg to elute the DNA. The concentration of the PCR product was quantified using the Nanodrop spectrophotometer.

Round 2 PCR: Using the products from the first round as template DNA, a second round of PCR was performed to add the NcoI cleavage site to both sequences, generating Nco-ATGglyHis6-P3C_220-288_StopBamHI and Nco-ATGglyHis6-P3C_220-270_StopBamHI constructs respectively. Primer sequences are shown in the table below. The protocol used in the first round (PCR mix, agarose gel electrophoresis, PCR clean up) was repeated. However, the DNA was eluted in buffer NE (5mM Tris-HCl, pH 8.5)

Constructs without P3C cleavage site

For constructs without P3C cleavage sites, only round of PCR was performed with the Phusion polymerase and PE603 as template DNA. At the end, NcoATGglyHis6_220-

270_StopBamHI and NcoATGglyHis6_220-288_StopBamHI were generated. The protocol used above (PCR mix, agarose gel electrophoresis, PCR clean up) was repeated. All primer sequences are described in the table below.

Final Construct	**Round 1 PCR**	**Round 2 PCR**
HIS-P3C-220-288	F – CTTGAAGTCCTCTTTCAGGGACCC ATTCAAAATTTTCGGGTTTATTACAGG R – CGGGATCCTTA ATCCTCATCCTGTCTACTTGCC	F – GCGCATGCCATGGGCCATCATCATCATCATCAC CTTGAAGTCCTCTTTCAGGGACCC R – CGGGATCCTTA ATCCTCATCCTGTCTACTTGCC
HIS-P3C-220-270	F – CTTGAAGTCCTCTTTCAGGGACCC ATTCAAAATTTTCGGGTTTATTACAGG R – CGGGATCCTTA ATCCCTGATGATCTTTGC	F – GCGCATGCCATGGGCCATCATCATCATCATCAC CTTGAAGTCCTCTTTCAGGGACCC R – CGGGATCCTTA ATCCCTGATGATCTTTGC
HIS-220-288	F – GCGCATGCCATGGGCCATCATCAT CATCATCAC ATTCAAAATTTTCGGGTTTATT ACAGG R – CGGGATCCTTA ATCCTCATCCTGTCTACTTGCC	
HIS-220-270	F – GCGCATGCCATGGGCCATCATCATCAT CATCAC ATTCAAAATTTTCGGGTTTATT ACAGG R – CGGGATCCTTA ATCCCTGATGATCTTTGC	

Table 1: Primers used in cloning pET15b constructs

Digestion: Subsequently, the vector (pET15b) and all four inserts described above were double digested. 2µg of pre-digested vector .To digest the insert, 20µl of insert (10ng/µl) was mixed with 5µl 10X Tango buffer (ThermoFisher), 2.5µl of NcoI and BamHI respectively, and 20µl of nuclease free water for a total reaction volume of 50µl. The mixture was left to incubate at 37°C for 2 hours. After digest, a PCR clean up was performed using the kit from Machery Nagel as described above.

Ligation: To ligate, a vector to insert ratio of 1:5 was used. The formula $I_{bp} \times V_{ng}/V_{bp} \times 5$ was used to calculate the amount of insert, where I_{bp} was the number of base pairs of the insert, V_{ng} was the amount of vector to be used, and V_{bp} was the number of base pairs of the vector. The ligation mix contained 2µl of buffer, 1µl of DNA ligase, 2µl of vector (20ng/µl), the appropriate amount of insert, and water up to 20µl. The reaction was allowed to incubate for 2 hours at room temperature. The ligation mix was transformed into TOP10 cells and incubated at 37°C overnight on LB-Agar supplemented with ampicillin.

Transformation: 50ng of plasmid was added in a sterile manner to 100µl of chemically competent TOP10 cells previously thawed on ice. The bacteria were left thirty minutes on ice, and then heat shocked at 42°C for one minute. The heat shock process allows the opening of the bacterial pores and plasmid entry into some of the bacteria. Subsequently, incubating on ice for one minute closed the bacterial pores. 900ml of 2X LB medium was added to the bacteria and incubated at 37°C for 1 hour with shaking at 250rpm. This step gives the bacteria with the plasmid time to generate antibiotic resistance proteins on the plasmid backbone. 150µl of the bacterial mixture was spread on a petri dish containing LB-agar medium and the appropriate antibiotic and incubated overnight at 37°C for selection.

Colony PCR: To screen for recombinant clones, colony PCR was performed. Six colonies per construct were suspended and cultured overnight in 2ml LB supplemented with ampicillin. 1ml of bacteria was boiled for 10 minutes. After centrifugation at 13,000rpm for 2 minutes, the pellet was re-suspended in 50µl of water, and 10µl was used to perform the PCR reaction using Taq polymerase, T7 universal primer (TAATACGACTCACTATAGGG) as the forward primer, and the pET15b primer as the reverse primer. Each reaction mix contained 10µl of template DNA, 5µl of 10X buffer, 0.25µl of Taq polymerase, 0.5µl of dNTPs, 1µl each of forward and reverse primers, and 32.25µl of water. The PCR reaction cycle consisted of a denaturation step at 94° for 2 minutes, followed by 25 cycles consisting of thermal denaturation at 94° for 15 seconds, followed by primer annealing at 50°C for 30 seconds, and an elongation step for 45 seconds at 72°C, and a final elongation step at 72°C for 7 minutes. Agarose gel electrophoresis with 10µl of PCR product was loaded on a 2% (w/v) agarose gel in 100ml TAE buffer containing 5µl Ethidium Bromide (10mg/ml) was performed to check for the presence of positive constructs.

DNA preparation: Using the leftover bacteria from the previous step, a DNA mini preparation was performed on colonies confirmed to contain the positive construct using NucleoSpin Plasmid kit from Machery Nagel. 1ml of bacteria was pelleted by centrifugation at 11000xg for 30 seconds and the supernatant was discarded. The pellet was completely resuspended in 250µl buffer A1 containing RNAse1 by pipetting up and down. 250µl of buffer A2 containing Sodium Dodecyl Sulfate (SDS) and sodium hydroxide was added and incubated at room temperature until lysate appeared clear (~5 minutes) to release plasmid DNA from the cells. 300µl of Buffer A3 containing guanidine hydrochloride was added and mixed by inverting 8 times to neutralize SDS and precipitate protein and genomic DNA. Buffer A3 also ensure

proper conditions for plasmid DNA to bind to the silica membrane. The mixture was centrifuged for 10 minutes at 11000xg and the supernatant was loaded on the silica membrane column. The column was centrifuged for 1 minute at 11000xg and the flow through was discarded. 500µl of buffer AW containing guanidine hydrochloride and 2-propanol was used to perform a first wash and discarded by centrifugation for 1 minute at 11000xg. A second wash with 600µl buffer A4 containing ethanol was performed followed by an extra drying step to completely remove traces of ethanol by centrifugation for 2 minute at 11000xg. 30µl of buffer AE (5 mM Tris/HCl, pH 8.5) was added to the silica membrane, incubated at room temperature for 1 minute, and centrifuged for 1minute at 11000xg to elute the DNA. The concentration of the PCR product was quantified using the Nanodrop spectrophotometer.

Sequencing: 20µl of each extracted plasmid DNA (50ng/µl) was sent to GATC for sequencing, with T7 universal primer and pET15b reverse primer.

Protein Production

Protein Engineering

All constructs of recombinant protein were expressed in (*E. coli* BL21 (DE3) (Novagen): F^- *ompT* *hsd*S_B(r_B^- m_B^-) *gal dcm* (DE3). BL21 is deficient in the *lon* protease and lacks the *ompT* outer membrane protease that can degrade proteins during purification. The bacteria also contain the lysogen encoding the T7 phage RNA polymerase, under the control of IPTG inducible lacUV5 operon. In the expression plasmid, the gene of interest is under the control of the T7 bacteriophage transcription promoter. Therefore, transcription can only be performed by T7 RNA polymerase in the presence of lactose or an analogue- Isopropyl β-D-1-ThioGalactopyranoside (IPTG). In the presence of an inducer, a conformational change is induced in the LacI repressor, preventing it from binding to the Lac operator site. This allows the expression of the T7 RNA polymerase that will bind to its promoter on the vector, and causes subsequent expression of the target gene. However, when an inducer is absent, the LacI repressor binds the Lac operator site, preventing the expression of T7 RNA polymerase, and causing repression of the target gene

Transformation

50ng of plasmid is added in a sterile manner to 100 µl of chemically competent BL21DE3 cells previously thawed on ice. The bacteria are left thirty minutes on ice, and then heat

shocked at 42°C for one minute. Subsequently, incubating on ice for one minute closes the bacterial pores. 900µl of 2X LB medium is added to the bacteria and incubated at 37°C for 1 hour with shaking at 250rpm. 150µl of the bacterial mixture is spread on a petri dish containing LB-agar medium and the appropriate antibiotic (ampicillin) and incubated overnight at 37°C for selection.

Pre-Culture

After transformation into BL21DE3 cells, 2ml of LB supplemented with ampicillin is inoculated with an isolated colony and incubated at 37°C for 8 hours with shaking. After 8 hours, the bacterial cells are spread on large petri dishes and placed at 37°C overnight. The day after, the plates are examined for phage contamination.

Culture (non-labeled protein)

The large petri dishes are scrapped with 10ml LB to recover the bacteria, and the OD600 is measured. For a 1L culture, bacteria are inoculated to an OD600 of 0.1 in a 5L Erlenmeyer flask. LB medium was supplemented with the appropriate antibiotic (ampicillin) and 10% (w/v) sucrose (50% (w/v) sucrose). Flasks are incubated at 37°C with shaking at 220rpm and growth is observed by measuring OD600. At an OD600 of 0.5, the temperature is reduced to 25°C, and shaking is reduced to 190rpm, till the cells reach an OD600 of 0.8. IPTG is added to a final concentration of 0.5mM to induce the target gene. Cells are incubated overnight at 25° for protein expression. The bacteria are then centrifuged at 4000rpm for 20 minutes at 4°C to pellet the cells, and resuspended in 10g/L NaCl to wash. They are subsequently pelleted, and frozen at -20°C. Final yield is about 4-6 grams per liter of culture.

Culture (Isotopically Labeled Protein)

The large petri dishes are scrapped with 10ml LB to recover the bacteria, and the OD600 is measured. For a 1L culture, bacteria are inoculated to an OD600 of 0.1 in a 5L Erlenmeyer flask. Minimal medium (10X M9 solution -Na_2HPO_4, KH_2PO4, NaCl, $^{15}NH_4Cl$), 100X Trace elements ($ZnSO_4$, $FeCl_3$, $CuCl_2$, $CoCl_2$, H_3BO_3, $MnCl_2$, 10ml ^{13}C glucose, $CaCl_2$, $MgSO_4$, Biotin, Thamin) is supplemented with the appropriate antibiotic (ampicillin). Flasks are incubated at 37°C with shaking at 220rpm and growth is observed by measuring OD600. At an OD600 of 0.5 the temperature was reduced to 25°C, and shaking is reduced to 190rpm, till the cells reach an OD600 of 0.8. IPTG was added to a final concentration of 0.5mM to induce the target gene. Cells were incubated for 4 hours at 25°C for protein expression. The bacteria

are the centrifuged at 4000rpm for 20 minutes at 4°C to pellet the cells, and resuspended in 10g/L NaCl to wash. They were subsequently pelleted, and frozen at -20°C. Final yield is about 2-3 grams per liter of culture

Protein purification

Lysis

Cells are resuspended in appropriate lysis buffer. Generally, they are resuspended and homogenized in a ratio of 10ml of buffer/gram of bacterial pellet. Roche Complete Inhibitor Cocktail tablets are added at the beginning of lysis to avoid protease degradation. Cells are lysed using a sonicator, for 1min/g of cells with pulse every 2 seconds at 40% amplitude at 4°C. To avoid overheating, the sample is immersed in ice, and the sonicator stops when a temperature of 10°C is reached.
The bacterial debris is cleared from soluble material by ultracentrifugation at 100000xg for 1hr at 4°C. The supernatant is recovered for immediate purification.

Purification

Generally, proteins of interest are purified using a 2-step process, first with affinity chromatography, and then with size exclusion chromatography. Each step is usually carried out using the AKTA purifier (GE Healthcare).

Affinity Chromatography

Affinity Chromatography is method of separation based on a specific but reversible interaction between a ligand immobilized on a matrix and its binding partner (protein). This specific interaction makes it possible to separate recombinant protein of interest from a mix of bacterial proteins. In this case, the protein of interest is labeled with an affinity tag – hexahistidine (6X HIS) or Glutathione-S-Transferase (GST), and bacterial lysate is passed over a matrix that selectively binds the 6X HIS or GST tag.

GST (Glutathione-S-Transferase) is a 26kDa protein attached to the N-terminus of the target protein. GSTrap (GE Healthcare) columns are utilized for purification, where Glutathione is immobilized on a sepharose matrix. GST-tagged protein binds to the Glutathione on the matrix, allowing for separation of the IN-CTD from the protein mix. To elute the IN-CTD from the resin, lysis buffer with 20mM reduced glutathione is utilized.

Immobilized Metal Affinity Chromatography allows for purification of proteins tagged with a histidine tag using immobilized divalent cations ($Ni^{2+,}$ Co^{2+}, Zn^{2+}, Cu^{2+}). A 6X HIS tag is attached to the N-terminus of target protein. For purification, HisTrap FF Crude (GE Healthcare) columns were used, where Ni^{2+} are immobilized on a sepharose matrix. After binding, increasing concentrations of imidazole are used to elute proteins. At low concentrations of imidazole, impurities or non-specifically bound endogenous proteins are washed. At higher concentrations, the protein of interest is eluted from the column.

Following Affinity Chromatography, samples are analyzed by SDS-PAGE. Pure fractions are pooled and concentrated using Amicon Ultra Centrifugal Filters for the subsequent purification step.

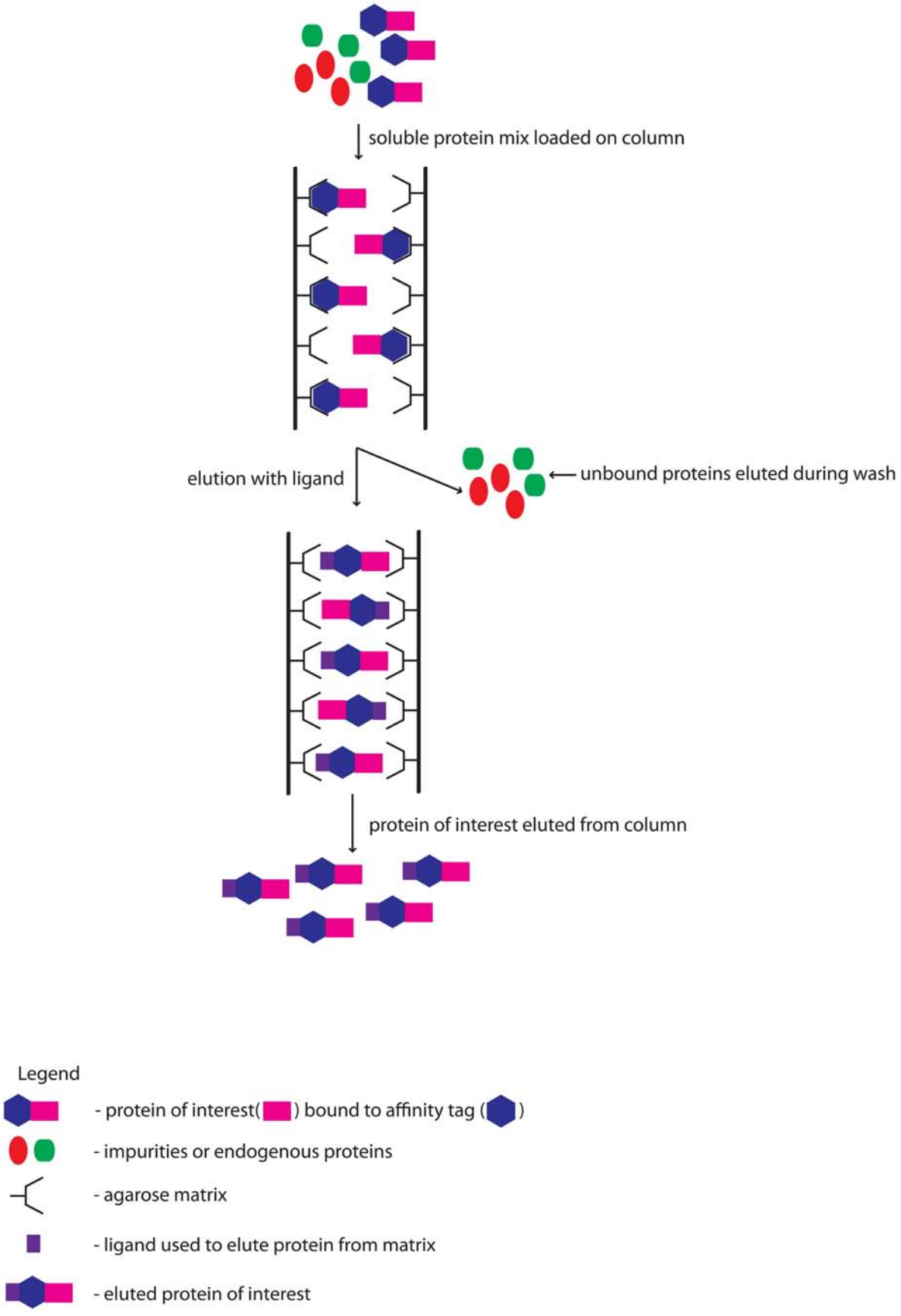


Figure 17: Theory of Affinity Chromatography

Size Exclusion Chromatography

Size Exclusion Chromatography separates proteins as a function of size and shape by filtering them through a gel. The gel consists of molecules of dextran covalently bound to highly cross-linked agarose that form porous granules. Proteins are separated based on their molecular weight in relation to the pore size. Proteins with diameters larger than the pore size are excluded from the pores and eluted in the dead volume. Proteins with diameters smaller than the pore size migrate through the pores depending on their size and shape in descending order in the elution volume. The proteins are eluted linearly as a function of the logarithm of their molecular mass, with larger proteins eluted first, and smaller proteins last.

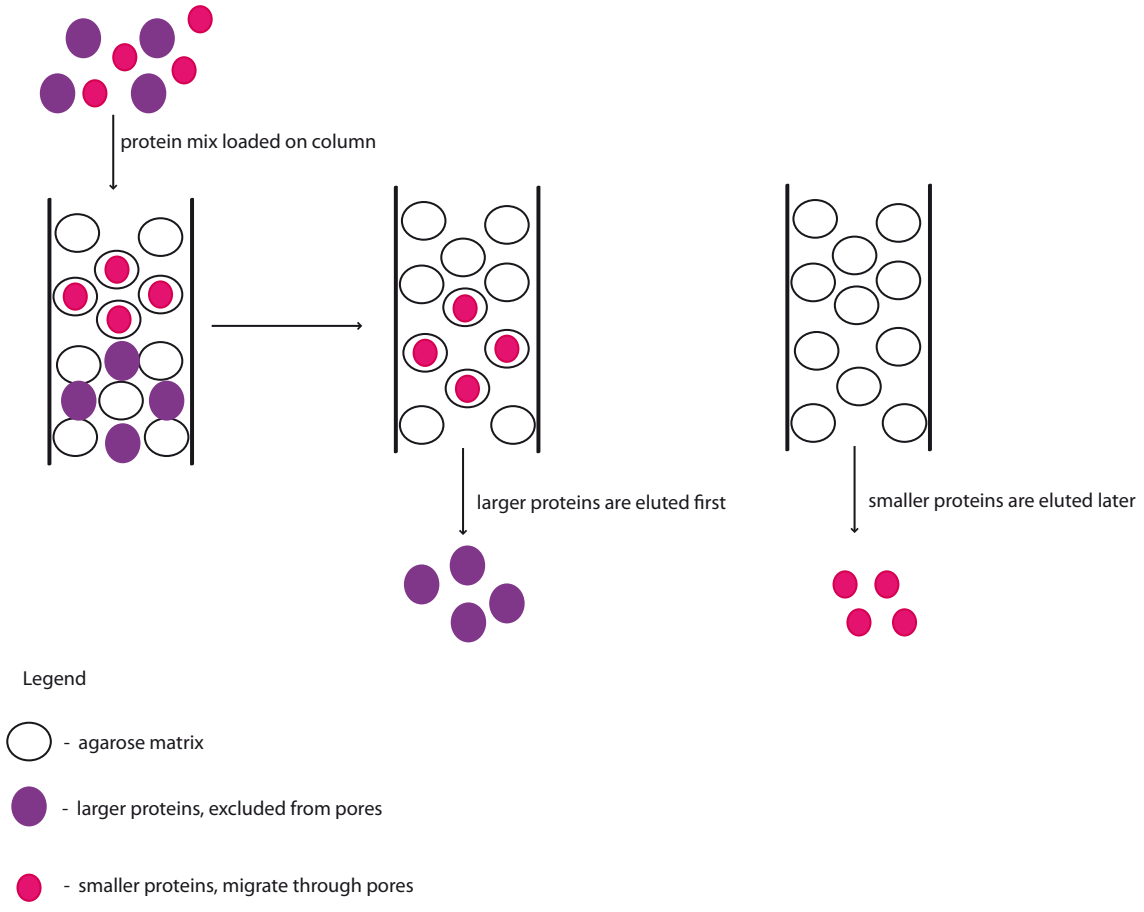


Figure 18: Theory of Size exclusion chromatography

Samples are analyzed by SDS-PAGE. Subsequently, purified fractions and pooled and stored for downstream applications.

Peptide Production and Quantification

Peptides representing the N-terminal tail of H4K20 were synthesized by the peptide synthesis platform at the IGBMC. Each synthetic peptide represented a fluorescent or non-fluorescent N-terminal tail of histone H4K20. Fluorescein was attached to the N-terminal tail to produce fluorescent peptides.

Peptide	Description	Sequence
H4K20me0	non-fluorescent non-methylated peptide	RHRKVLR
H4K20me1	non-fluorescent mono-methylated peptide	RHRK(me1)VLR
Fluo-H4K20me0	fluorescent non-methylated peptide	Fluo-KGG-RHRKVLR
Fluo-H4K20me1	fluorescent mono-methylated peptide	Fluo-KGG-RHRK(me1)VLR
Fluo-H4K20me2	fluorescent di-methylated peptide	Fluo-KGG-RHRK(me2)VLR
Fluo-H4K20me3	fluorescent tri-methylated peptide	Fluo-KGG-RHRK(me3)VLR

Table 2: List of peptides produced by the platform

Precise concentration of each peptide was calculated using NMR (Köhler, Recht et al. 2015). Briefly, a known concentration of pure L-tryptophan (6.3mM) was prepared and the concentration was determined by UV absorption at 280nm. A small volume of tryptophan was added to a small volume of peptide and D2O was added to the mixture; and a 1D proton NMR spectrum was recorded. To calculate the concentration, the signals of the tryptophan and peptide are integrated. The ratio between the integrated signals provided the concentration of the peptide.

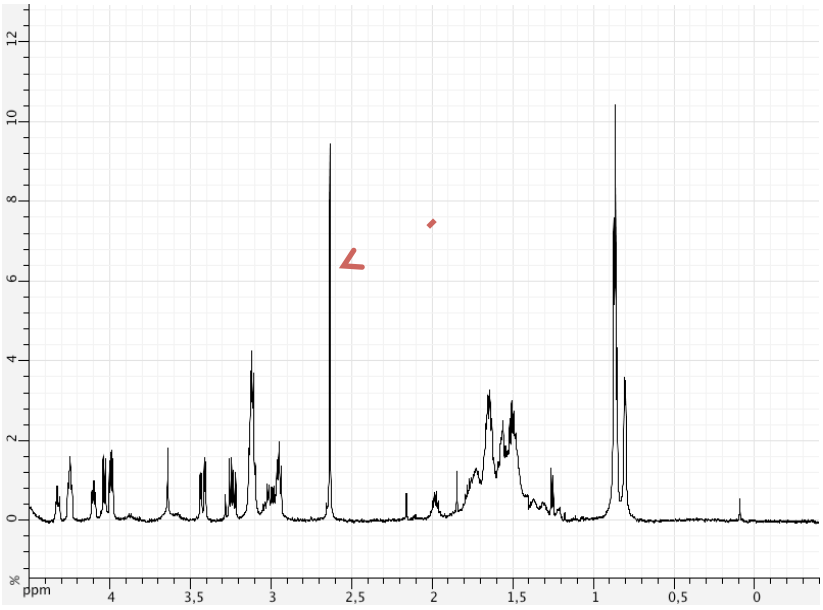


Figure 19: 1D spectrum of the Trp/H4K20Me1 peptide mixture . The peak at 2.63 ppm is characteristic of the N-methyl on the peptide

Biochemical Studies

These experiments were carried out with the aim of confirming the interaction between GST-IN-CTD and H4K20Me1 peptide. MST was used to determine the binding constant of the interaction. Although the principle and sensitivity of these techniques differ, they all assess changes in biophysical characteristics of the peptide.

Microscale Thermophoresis

Background

Microscale Thermophoresis (MST) is a technique that measure changes hydration shell, charge or size of fluorescent molecules as they move along a microscopic temperature gradient. This makes it a sensitive technique, allowing it to overcome limitations of previous fluorescence based techniques that were heavily dependent on parameters such as molecule size, or proximity of fluorescent molecules to each other (Jerabek-Willemsen, Wienken et al. 2011, Seidel, Dijkman et al. 2013, Jerabek-Willemsen, André et al. 2014).

In the Nanotemper Monolith NT.115 instrument, three fluorescence channels are available: blue (excitation 460–480 nm, emission 515–530 nm), green (excitation 515– 525 nm, emission 560–585 nm) and red (excitation 605–645 nm, emission 680–685 nm). Capillaries are filled through capillary action with about 4µl of sample. The capillary tray in the Monolith NT.115 instrument holds 16 capillaries, making it possible to perform 16 serial dilutions per experiment. Each capillary has diameter variations of less than 1µm, allowing results to be highly reproducible. As shown in the figure below, while samples are homogenously distributed, a fluorescence signal is measured initially. An infrared laser with a wavelength of 1480nm is used to focus on a precise spot (2nl) on the sample causing a temperature jump of 2-6K (T-jump). T-jump is dependent on inherent properties of the fluorophore, induces thermophoresis and causes a concentration gradient in the sample. This thermophoretic movement is measured for 30s. Once the infrared laser is turned off, an inverse T-jump is observed, followed by a back-diffusion of molecules, which is dependent on molecule size.

Change in thermophoresis is measured as change in normalized fluorescence (ΔF_{norm}), which is defined as F_{hot}/F_{cold}. (Figure 20). Titration of the non-fluorescent ligand results in a gradual change in thermophoresis, which is plotted to yield a binding curve, and can be fitted to derive binding constants.

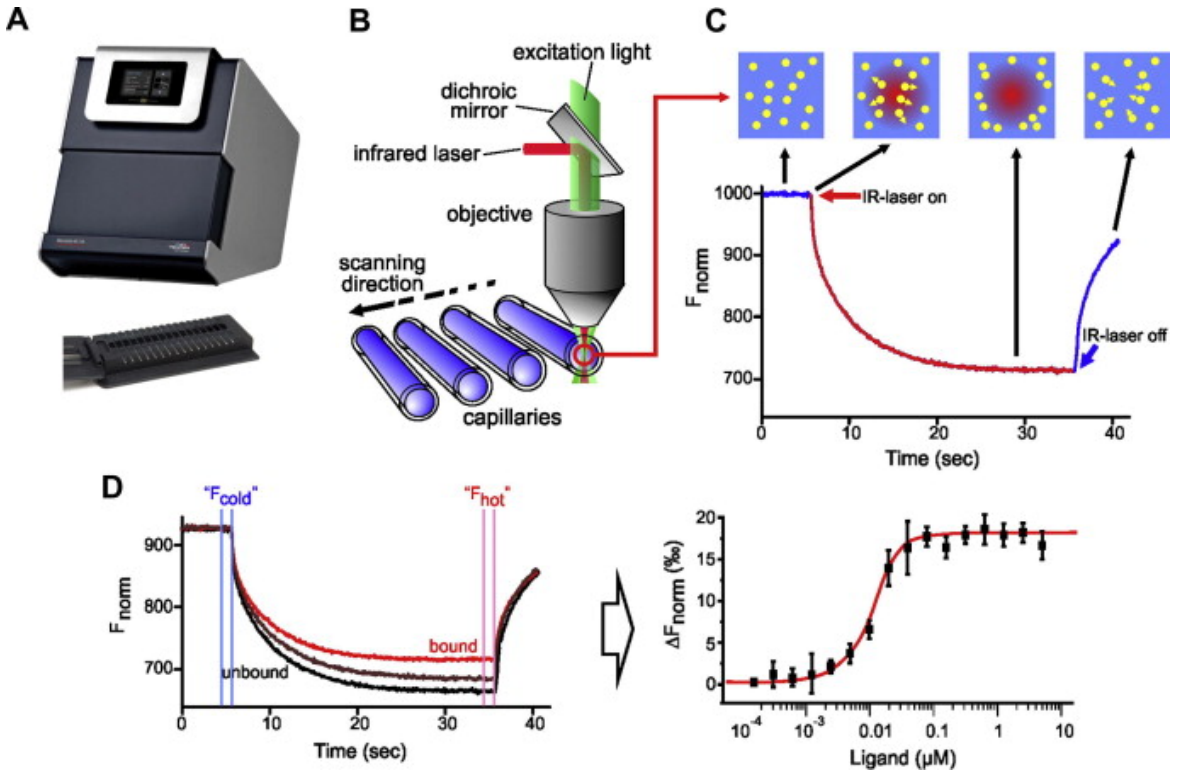


Figure 20: Schematic representation of MST A) Monolith NT.115 showing capillary tray. B) MST is measured in capillaries, which hold ~ 4ul of sample. Initial fluorescence is excited and measured in the sample through the objective. Infrared laser is used to heat a small amount of sample, which induces thermophoresis. C) While the samples are homogenously distributed, initial fluorescence is measured, followed by the activation of the infrared laser. T-jump is observed due to rapid change in the fluorophore following a fast temperature increase. Thermophoresis of the fluorescently labeled partner is measured for 30s. After the IR-laser is turned off, an inverse T-jump occurs, followed by a back-diffusion of molecules, driven by mass diffusion. D) Black trace – thermophoretic movement of an unbound fluorescent molecule, red trace- thermophoretic movement of a bound fluorescent molecule. Change in thermophoresis is defined by change in normalized fluorescence (ΔF_{norm}) - F_{hot}/F_{cold} . *Jerabek-Willemsen, Journal of Molecular Structure, 2014*

Protocol

1:1 serial dilutions of GST-IN-CTD are performed, ranging from 185uM to 5nM. Each peptide was mixed with diluted protein to a final concentration of 0.5µM. After addition, the protein and peptide mixture were allowed to equilibrate for 15 minutes at room temperature. All measurements were made using the Nanotemper Monolith NT.015 instrument with laser-on time 30seconds and laser-off time 5sec at 20% LED and 40% MST IR-Laser power.

For the competition experiments, 1:1 serial dilutions of GST-IN-CTD are performed, ranging from 185µM to 5nM. Experiments were performed at 0.5µM of each peptide. 0.5µM of the non-fluorescent peptide (H4K20Me0) was added to each protein sample. Subsequently, 0.5µM of the fluorescent monomethylated peptide was added to the mixture and incubated for 15 minutes at room temperature. The experiments were carried out at 20% LED and 40% MST IR-Laser.

Transferred Nuclear Overhauser effect Spectroscopy

Background

Transferred Nuclear Overhauser Effect Spectroscopy (trNOESY) is a two-dimensional proton NMR experiment allowing the detection of a binding event for the ligand when there is a change in sign and the intensity buildup rate of its intramoleculars NOEs (Bothner and Gassend 1973). This technique relies on the difference in correlation times between free and bound ligand. In a sample with free ligand or no interaction, samples exhibit a positive NOE due to short correlation times and slow NOE accumulation. In the presence of a complex, the ligand exhibits large correlation times, fast NOE accumulation, leading to negative NOEs. This technique makes it really easy to observe binding with the sign and size of NOEs.

Protocol

100µM GST-IN-CTD in 25mM HEPES pH 8, 2mM $MgCl_2$, 2mM βME, 150mM NaCl was mixed with 750µM of H4K20me1 peptide. Experiments were conducted at 298K. 3mm tubes were filled with 150µl of sample plus 10% D_2O and inserted in a 600 MHz Bruker spectrometer equipped with a TXI cryo-probe. Data was also collected on a control sample containing 750µM of H4K20me1 only.

To confirm that the interaction was occurring with the CTD of Integrase and not GST, experiments were recorded in the presence of GST only. 75µM of GST in 25mM HEPES pH 8, 2mM $MgCl_2$, 2mM βME, 150mM NaCl was mixed with 750µM H4K20me1. 3mm tubes were filled with 150µl of sample plus 10% D_2O and inserted in a 600 MHz Bruker spectrometer equipped with a TXI cryo-probe. The NOESY experiment was set up with a mixing time of 400 milliseconds and the acquisition lasted 12 hours (32 scans and a FID of 4096 * 512 points).

Water-Ligand Observed via Gradient Spectroscopy

Background

Water-Ligand Observed via Gradient Spectroscopy (Water-LOGSY) is a one-dimensional proton NMR experiment using bulk water to detect ligand binding (Dalvit, Fogliatto et al. 2001). As in NOESY based experiments, magnetization is transferred through intermolecular NOE and spin diffusion. However, in Water-LOGSY, bulk water magnetization is excited and transferred during the NOESY mixing time to the bound ligand. In a sample with free ligand

or no interaction, molecules interact with bulk water only, resulting in faster tumbling and positive NOEs. In the presence of an interaction, the rotational correlation times yield negative NOEs due to negative cross relaxation rates.

Protocol

100μM GST-IN-CTD in 25mM HEPES pH 8, 2mM $MgCl_2$, 2mM ®ME, 150mM NaCl was mixed with 750μM of H4K20me1 peptide. Experiments were conducted at 298K. 3mm tubes were filled with 150μl of sample plus 10% D_2O and inserted in a 600 MHz Bruker spectrometer equipped with a TXI cryo-probe. Data was also collected on a control sample containing 750μM of H4K20me1 peptide only.

Structural Studies

These experiments were carried out with the aim of determining the structure of the IN-CTD in complex with H4K20Me1 peptide, in order to identify the residues in the protein that are responsible for interacting with the peptide

Nuclear Magnetic Resonance

Nuclear Magnetic Resonance is a method based on studying the chemical properties of individual nuclei for the determination of the three-dimensional structure of protein in solution phase. Typically, this method is optimum for proteins between 5kDa and 25kDa.

Background

Every atom nuclei rotates around a given axis, giving them a spin property. Each spin is a vector quantity, possessing a spin angular momentum. Each charged spinning nucleus produces a magnetic field in a closed circuit. When magnetic nuclei are placed in an external magnetic environment, they produce an alternative field, which oppose the external magnetic field, causing a shielding effect, or a chemical shift. The chemical shift is very dependent on the strength of the external magnetic field, the atomic structure and geometry of the molecule of interest. Each nuclei spin experiences the magnetic field difference, producing a resonance frequency that is very sensitive to its environment (Poulsen 2002).
The chemical shift is measured in Hertz, shifted to a reference signal. To convert to ppm, the chemical shift difference is divided by the strength of the magnetic field of the spectrometer in MHz.

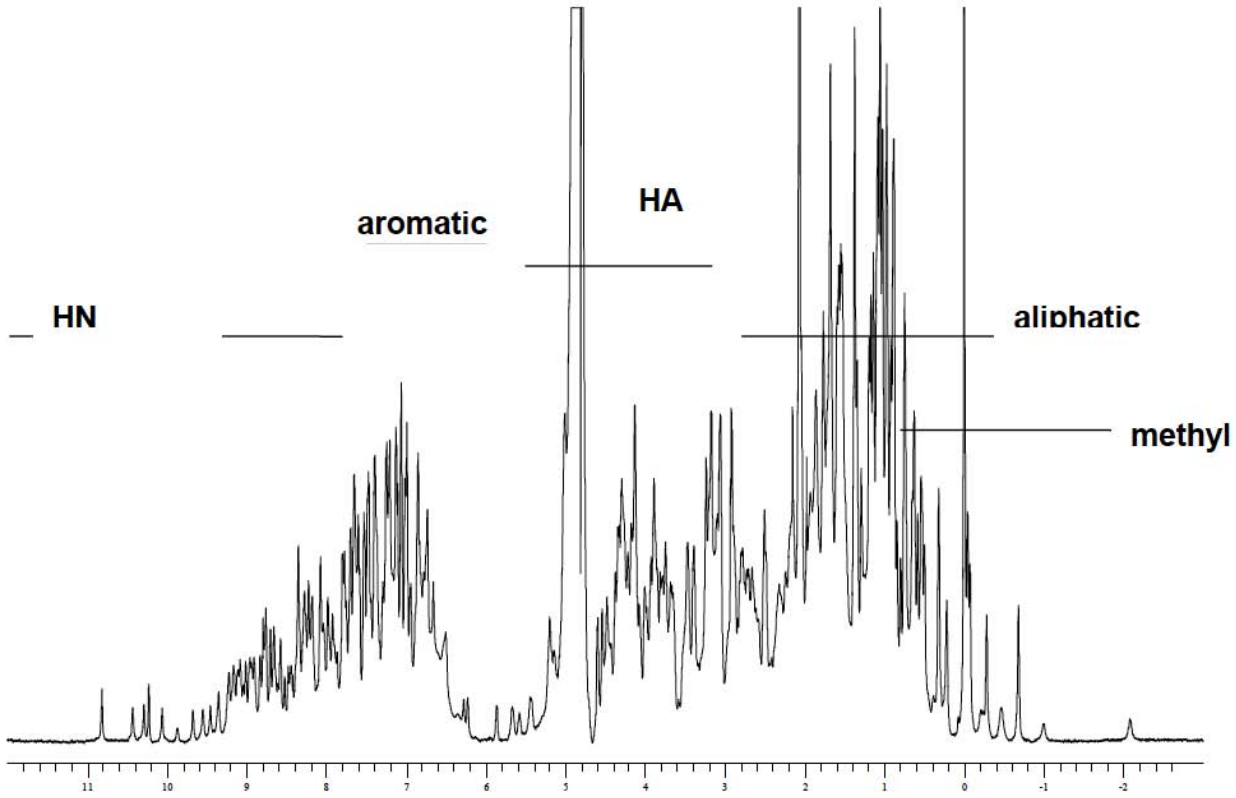


Figure 21: 1D spectrum of well-folded Hen egg white lysozyme showing the chemical shift dispersion of chemical groups. On the ppm axis, the numbers increase towards the left.

Thus, in a spectrum, the amide protons are between 7 and 10 ppm, the amide protons of the side chain chains (Asn and Gln) and the protons of the aromatics (Trp, Tyr, Phe, His) between 5 and 7 ppm, the Hα protons between 4 and 5 ppm, and the aliphatic protons below 4 ppm. These characteristics make it easy to differentiate between a well folded and an unfolded protein.

Multidimensional Spectra

Additional dimensions (2D and 3D) can be established to observe additional nuclei (^{13}C and ^{15}N) by introducing additional pulses. To determine the secondary structure of a protein, chemical shifts of ^{1}Hα, ^{13}Cα, ^{13}Cβ, ^{15}N, and ^{13}CO must be analyzed. Chemical shifts give a lot of information about the structure of the protein. It is possible to derive information about dihedral psi and phi angles from chemical shift analysis. ^{13}Cα in alpha helices tend to be positive, while ^{13}Cα in β strands tend to be negative, also providing preliminary information

about the secondary structure of the protein. In order to calculate tridimensional structures using NMR, NOEs are also assigned to obtain long-range hydrogen correlations.

Isotopically Labeled protein

To carry out multi-dimensional experiments, the protein of interest must be isotopically labeled (McIntosh and Dahlquist 2009). This is typically achieved by protein expression in isotopically enriched minimal medium. Minimal medium contains the most basic salts and trace elements needed for bacteria growth without any carbon or nitrogen sources. Carbon or Nitrogen is introduced by using isotopically labeled carbon or nitrogen sources. To label with ^{15}N, bacteria are grown in minimal medium and nitrogen is introduced by using $^{15}NH_4Cl$ and unlabeled glucose. ^{15}N labeled protein is used for ^{15}N HSQC experiments. To label with ^{13}C-^{15}N, bacteria are grown in minimal medium, with $^{15}NH_4Cl$ and ^{13}C-glucose as the carbon source.

Description of Multidimensional Experiments

^{1}H-^{15}N HSQC

The ^{15}N HSQC is the most basic experiment performed in protein NMR. HSQC (Heteronuclear Single Quantum Spectroscopy) is a two-dimensional experiment that involves the transfer of magnetization from the ^{15}NH group to the backbone ^{15}N via J-coupling (Cavanagh, Fairbrother et al. 2007). This transfer induces a chemical shift on the nitrogen, followed by transfer of magnetization back to the hydrogen for detection. Since each amino acid except proline has amide hydrogen attached to its backbone nitrogen, each amide peak theoretically represents a residue in the protein. Therefore, these experiments provide a fingerprint of the protein. Asparagine/Glutamine side-chain Nδ-Hδ2/Nε-Hε2 groups are visible in the top right corner as doublets. Tryptophan side-chain Nε-Hε groups appear in the bottom left corner.

Peak intensity and peak dispersion data obtained from ^{15}N HSQC experiments provide information on protein quality and are used to determine if further experiments can be carried out on the sample.

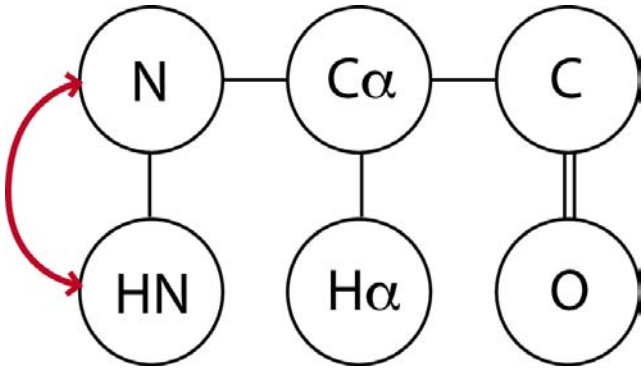


Figure 22: Transfer of magnetization during ^{15}N HSQC *Adapted from (Higman 2012)*

^{1}H-^{13}C HSQC

The ^{13}C HSQC is the carbon equivalent of the ^{15}N HSQC. It shows all H-C correlations, thereby providing a fingerprint of all carbon atoms in the protein (Cavanagh, Fairbrother et al. 2007). Magnetization is transferred from ^{1}H to all ^{13}C atoms, and back to ^{1}H for detection.

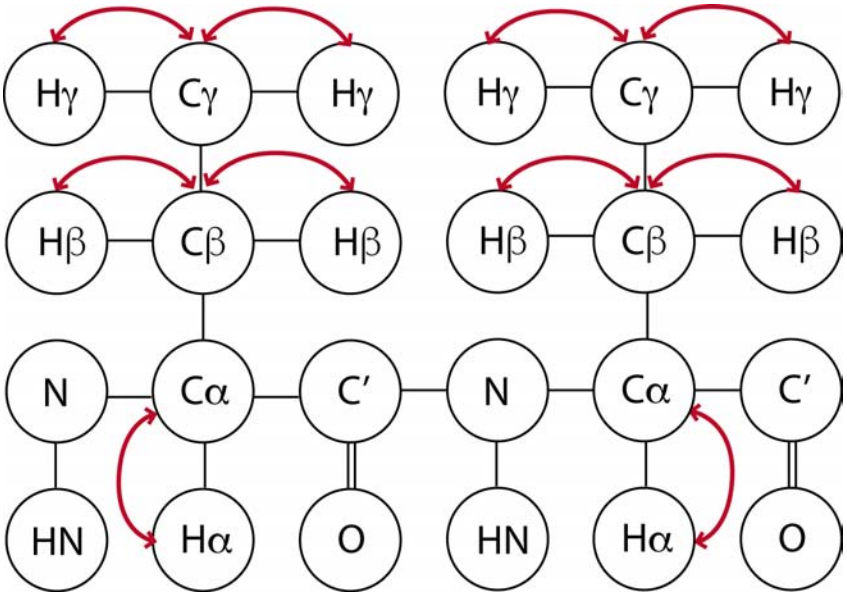


Figure 23: Red arrow indicate transfer of magnetization for ^{13}C HSQC Adapted from (Higman 2012)

HNCA

The HNCA is a three-dimensional experiment that correlates ^{15}N and ^{15}NH shifts with the chemical shift of Cα of the same residue (Kay, Ikura et al. 1990, Farmer, Venters et al. 1992). Since the amide nitrogen is also coupled to the Cα of the preceding residue, it also provides information for sequential assignment by yielding weak correlations between ^{15}NH and ^{15}N

shifts with chemical shifts of Cα of the previous residue. Magnetization is transferred from ^{15}NH to ^{15}N, and then to Cα via N-Cα J-coupling, and then back again to ^{15}N and ^{15}NH for detection. It is used to obtain Cα and Cα-1 chemical shifts for sequential backbone assignment.

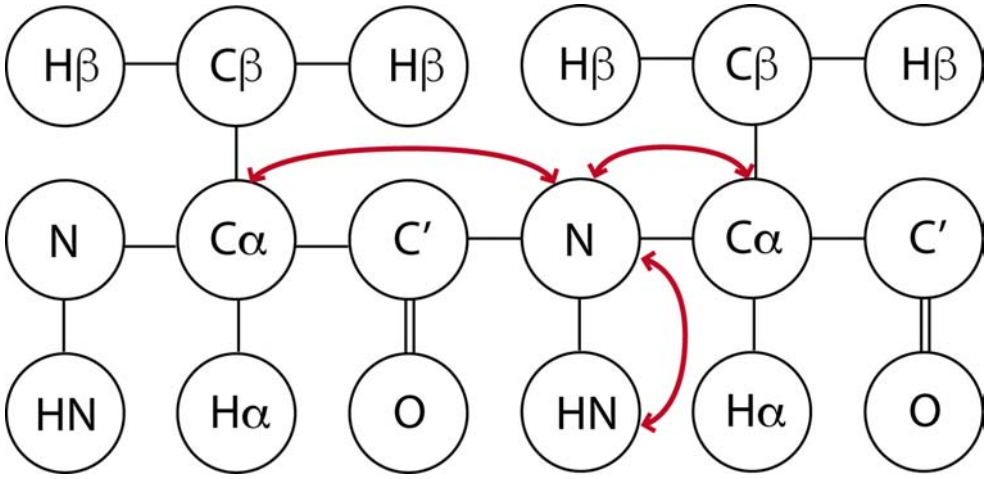


Figure 24: Transfer of magnetization for HNCA. *Adapted from (Higman 2012)*

CBCACONH

CBCACONH is a three dimensional experiment that correlates amide chemical shifts with Cα and Cβ of the previous residue through the ^{13}CO group, providing information about the amino acid preceding the amide group (Grzesiek and Bax 1992). Magnetization is transferred from Hα and Hβ to Cα and Cβ respectively. Subsequently, it is transferred to ^{13}CO, and the ^{15}N and ^{15}NH for detection. The CBCACONH is used along with HNCA and HNCACB for sequential backbone assignment.

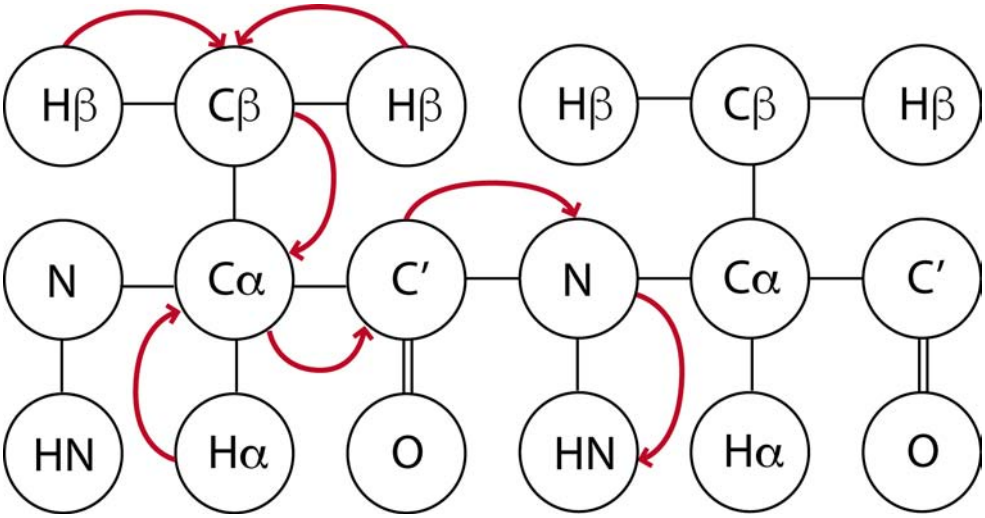


Figure 25: Transfer of magnetization for CBCACONH. *Adapted from (Higman 2012)*

HNCO

The HNCO is a three-dimensional experiment that correlates the ^{15}NH and ^{15}N chemical shifts with the shift of the carbonyl group of the preceding residue (Kay, Ikura et al. 1990). Magnetization is transferred from ^{15}NH protons to the attached ^{15}N, and then from the to ^{13}CO via ^{15}NH-^{13}CO J-coupling. Magnetization is passed back to the ^{1}H via ^{15}N for detection. It is used to obtain CO chemical shifts, which are important for sequential backbone assignment, and secondary structure prediction.

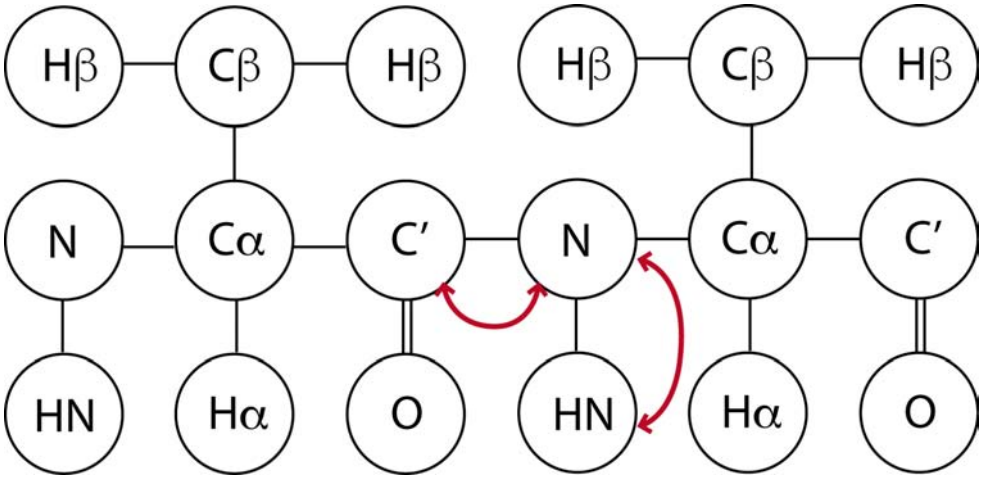


Figure 26: Transfer of magnetization for HNCO. *Adapted from (Higman 2012)*

HNCACB

The HNCACB is a three-dimensional experiment that correlates $^{13}C\alpha$ and $^{13}C\beta$ chemical shifts with ^{15}N and ^{15}NH (Grzesiek and Bax 1992). Magnetization is transferred from $^{1}H\alpha$ and $^{1}H\beta$ to 13Cα and 13Cβ. When a pulse is applied, Cβ magnetization is transferred to Cα. Subsequently, magnetization is transferred to intra-residue ^{15}N to ^{15}NH, and ^{15}N and ^{15}NH of the next residue via Cα-N J coupling for detection. Magnetization is transferred to ^{15}N and ^{15}NH from both Cα and Cα-1. Therefore for each NH group, its Cα, Cβ, Cα-1 and Cβ-1 are visible. The HNCACB is used alongside the HNCA for sequential backbone assignment.

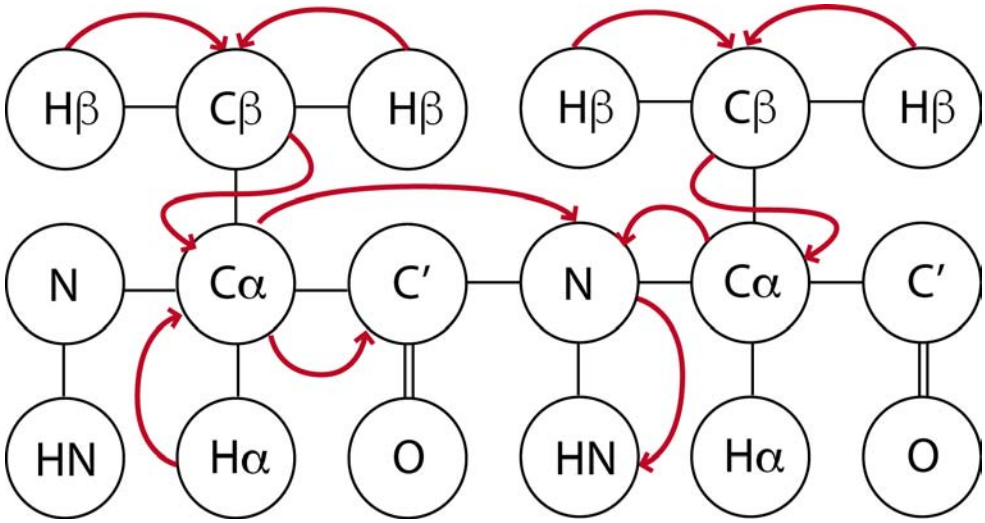


Figure 27: Transfer of magnetization for HNCACB. *Adapted from (Higman 2012)*

^{15}N NOESY HSQC

The ^{15}N NOESY (Nuclear Overhauser Effect Spectroscopy) is a three dimensional experiment that provides through space correlations for neighboring hydrogens (Marion, Driscoll et al. 1989, Marion, Kay et al. 1989). It is based on obtaining a large number of internuclear distances through Nuclear Overhauser effects (NOEs). Magnetization is transferred between all hydrogen atoms through NOEs. This magnetization is then transferred to nearby ^{15}N and then to ^{15}NH for detection. This method is used to obtain distance restraints for structure calculations. The ^{15}N NOESY can also be combined with other 2D and 3D experiments for sequential assignment.

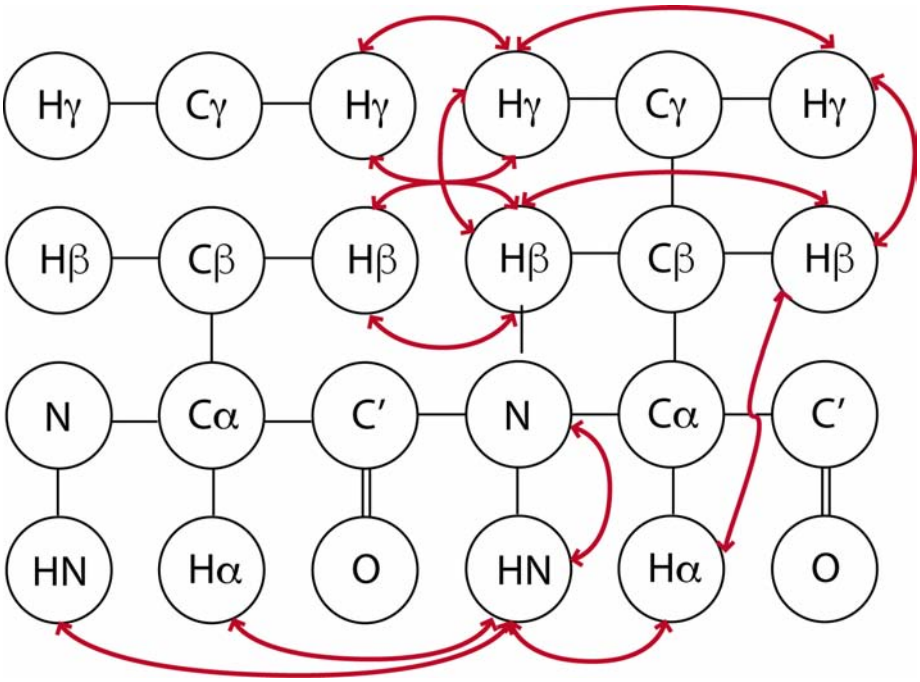


Figure 28: Transfer of magnetization for ^{15}N NOES. *Adapted from (Higman 2012)*

Chemical Shift Mapping

The chemical shift is a very sensitive tool for observing ligand binding to protein due to its sensitivity to its environment. Isotopically labeled protein, typically ^{15}N, is titrated with increasing concentration of ligand, and 2D HSQC spectra are recorded after each addition. Due to the sensitivity of chemical shifts, any binding event can be easily detected by following movement of the peaks. The peaks that are affected in the presence of ligand are more likely to be the binding site of the ligand. Changes in ^{15}NH chemical shifts are likely to be induced by hydrogen bonding interactions to amide protons. A change in environment due to interaction with binding partner and the variation of chemical shifts can also induce changes in ^{15}NH.

Chemical shift mapping can also be used to understand the kinetics of the interaction (Williamson 2013). For a protein P binding to a ligand L at a single site, the reaction P+L $\rightleftharpoons$ PL has two rate constants, k_{on} and k_{off} for the forward and backward reactions. Strong interactions undergo slow exchange where the k_{off} rate is slower than the difference in chemical shift between free and bound protein. Therefore, as the ligand is added, new resonance peaks that represent the complex will appear and grow in intensity while the peaks representing the free protein slowly disappear. Put simply, the interaction is so strong during

the experiment that both the free and bound protein can be observed simultaneously. On the contrary, weak interactions undergo fast exchange where the k_{off} rate is faster than the chemical shift difference. Therefore, as the ligand is added, resonance peaks move slowly move from the free protein to the complex form.

Protocol

In order to obtain residue assignments, 3D experiments were recorded using 250μM of ^{15}N ^{13}C labeled HIS-IN-CTD 220-270 in 25mM HEPES pH 8, 150mM NaCl. All experiments were recorded in a Bruker 700MHz spectrometer. Additionally, titration experiments were recorded with H4K20Me1 peptide using ^{15}N labeled protein in order to map chemical shift changes that occur upon interaction.

Circular Dichroism

Background

CD spectroscopy is based on the difference of absorbance of left circularly polarized light and right circularly polarized light by a chiral molecule. All amino acids, except glycine are chiral molecules, in that they lack a plane of symmetry, and are not superimposable with their mirror images. A solution of chiral molecules absorbs circularly polarized light at different levels at far-UV wavelengths due to differences in extinction coefficient for the two polarized rays. Far-UV spectra (190nm – 250nm) are derived from absorption by peptide bonds, and are representative of the secondary structure features in the protein. Different secondary structure features have characteristic CD spectra. While α-helices have negative bands at 222nm and 208nm and a positive band at 193nm; β-sheets have negative bands at 218nm and positive bands at 195nm. Disordered proteins display negative bands near 195nm. Therefore, CD is widely used to study protein conformation and observe changes in secondary structure content. (Martin and Bayley 2002, Kelly, Jess et al. 2005, Greenfield 2006)

Protocol

In order to observe if there were any changes in secondary structure composition at each pH, 200μM of HIS-IN-CTD 220-270 in 25mM HEPES pH 7, 150mM NaCl was mixed with 2mM of H4K20Me1 peptide. Likewise, 190μM of protein in 25mM HEPES pH 8, 150mM NaCl was mixed with 2mM of H4K20Me1 peptide. Control experiments were recorded with protein without peptide at each pH. Reference experiments were recorded with buffer only at

each pH. Far UV CD spectra was recorded using 65µl of each sample at 22°C in a JASCO J-815 Spectropolarimeter.

X-ray Crystallography

Background

X-ray crystallography is based on the use of X-rays - electromagnetic waves with a wavelength of 1 Angstrom (10^{-10}m) to obtain 3D structures from protein crystals at atomic resolution.

Crystallogenesis requires the sample needs to be in a state of super-saturation. There are two main steps involved in crystal formation: nucleation and growth. In a crystallization phase diagram, there are four main regions. In the region of under-saturation, the protein or precipitant concentration is too low for crystal formation. In the region of high super saturation, protein precipitation occurs. In the nucleation zone, crystal nucleation occurs, and several microcrystals may appear. The metastable zone is the best zone for the growth of large stable crystals.

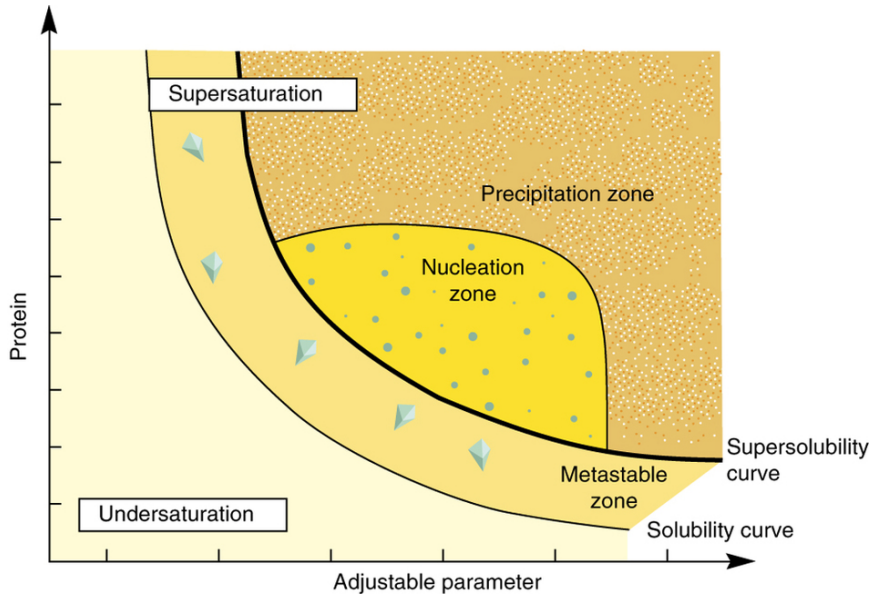


Figure 29: Vapor Diffusion Crystallization phase diagram . *Khurshid et al, Nature Protocols, 2014*

The most commonly used method for obtaining protein crystals is by vapor diffusion. Vapor diffusion can be achieved with the use of hanging drops or sitting drops. The principle of both methods is the same: a drop of protein is mixed with a drop of precipitating agent, and allowed to equilibrate with a reservoir of higher concentration of precipitating agent. As

equilibration occurs, the concentration of precipitant and protein in the drop increase, crystal formation can occur. Batch crystallization or counter-diffusion methods can also be used.

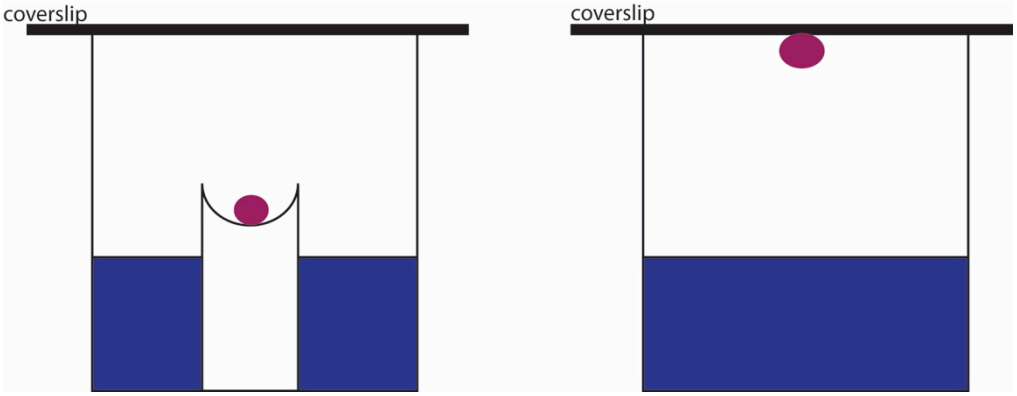


Figure 30: Vapor diffusion methods. Sitting drop vapor diffusion (left) and hanging drop vapor diffusion (right)

Important factors for crystal formation include protein concentration, pH, buffer concentration, temperature and salt. In order to find crystallization conditions, initial tests are typically done using sparse-matrix commercial screens. These high throughput screens are usually done using the mosquito robot. Conditions found in initial hits can be refined using manual drops to grow larger crystals (Asherie 2004, Chayen and Saridakis 2008, Wlodawer, Minor et al. 2008, Khurshid, Saridakis et al. 2014).

Protocol

All initial crystallization conditions were determined using the Sparse Matrix strategy using the TPP Labtech mosquito crystal. Generally, 200nL of protein was mixed with 200nL reservoir in 2 or 3 well 96 well MRC Crystallization plates. Plates are stored in Formulatrix RockImager at 20°C that regularly took a snapshot of each drop.

GST-IN-CTD crystallization screens

Several attempts were made to obtain crystals of the GST-IN-CTD. GST-IN-CTD was tested at several concentrations, and plus/minus peptide and DNA. Screens tested included JCSG, MIDAS, MPD, CLASSICS, NUCLEIX, and WIZARDS, TOP96.

Construct	Protein Conditions	Buffer conditions	Conditions Screened
GST-IN-CTD	17mg/ml + /- peptide	150mM NaCl, 100mM Arginine	JSCG+, MIDAS, MPD, WIZARDS
	+/- peptide	150mM NaCl	JCSG+, MPD, PACT, TOP96
	5mg/ml +/- pep	150mM NaCl	MPD, CLASSICS, JCSG, WIZARDS
	peptide+ seeds	150mM NaCl, 100mM Arginine	JSCG+, PEGS, PEGIONPH
	+ peptide+seeds	150mM NaCl	Additive
	+ DNA+peptide	pH 9, 150mM NaCl	Classics, JSCG+, MPD, NUCLEIX, WIZARDS

Table 3: Summary of crystallization conditions tested for GST-IN-CTD

HIS-IN-CTD 220-270 crystallization screens

Using the TTP Labtech's mosquito crystal, the Sparse Matrix strategy was used to determine initial crystallization conditions. HIS-IN-CTD 220-270 in 25mM HEPES pH 8, 150mM NaCl was tested at 5.3mg/ml, and plus peptide and DNA, at a ratio at 1:1.2, and 1:1.2:1.2 respectively. Screens tested included JCSG, PEGS, MPD, CLASSICS, NUCLEIX, WIZARDS, ANION and CATION. Subsequently, Additive and SlicepH screens were used, in an attempt to improve the quality of DNA crystals.

Once initial conditions were obtained, manual drops were set up in Hampton Research 24 well VDX plates to optimize crystallization conditions, and to improve crystal size and quality.

RESULTS

Purification Results

Expression Vector	Purification Tag	Cloning Strategy	Sequence limit	Molecular Weight	Theoretical Isoelectric Point	Ext. Coefficient
pHGWA	6X HIS-P3C	Gateway	220-288	11.6kDa	9.3	18450
pDEST 15	GST-HIS-P3C	Gateway	220-288	36.63kDa	7.59	65570
pET 15	6X HIS-P3C	Restriction	220-270	7.89kDa	9.99	13980
pET 15	6X HIS-P3C	Restriction	220-288	9.86kDa	8.82	15470
pET 15	6X HIS	Restriction	220-270	7.01kDa	10.20	13980
pET 15	6X HIS	Restriction	220-288	8.98kDa	9.21	15470

Table 4: Summary of constructs produced for this study

Purification of GST-IN-CTD (Gateway Construct)

9.8g of cells from a 2L production was resuspended in 98ml of lysis buffer - 25mM HEPES pH 8, 2mM $MgCl_2$, 2mM βME, 150mM NaCl, 100mM Arginine. Cells were lysed using a sonicator with a 13mm probe for 10 minutes with 2 sec on/off pulse at 40% amplitude at 4°C. Following ultracentrifugation at 185000xg for 1 hour, the supernatant was loaded on a 5ml GSTrap FF column (GE Healthcare) with a flow rate of 1ml/min using the AKTA purifier. Protein was eluted in one step of 20mM reduced glutathione. The amount of protein after affinity purification was 20mg. Protein concentration/quality was analyzed using the Nanodrop. Subsequently, the protein sample was concentrated using the Amicon Ultra 15ml with a MWCO of 10kDa for the next purification step. A second step of purification was carried out using the S200 16/60 column (GE Healthcare) in 25mM HEPES pH 8, 2mM $MgCl_2$, 2mM βME, 150mM NaCl, with a final amount of 17mg protein for 2L of culture. The GST tagged IN-CTD was soluble in 150mM NaCl, reaching a final concentration of 10mg/ml without precipitation in the absence of arginine

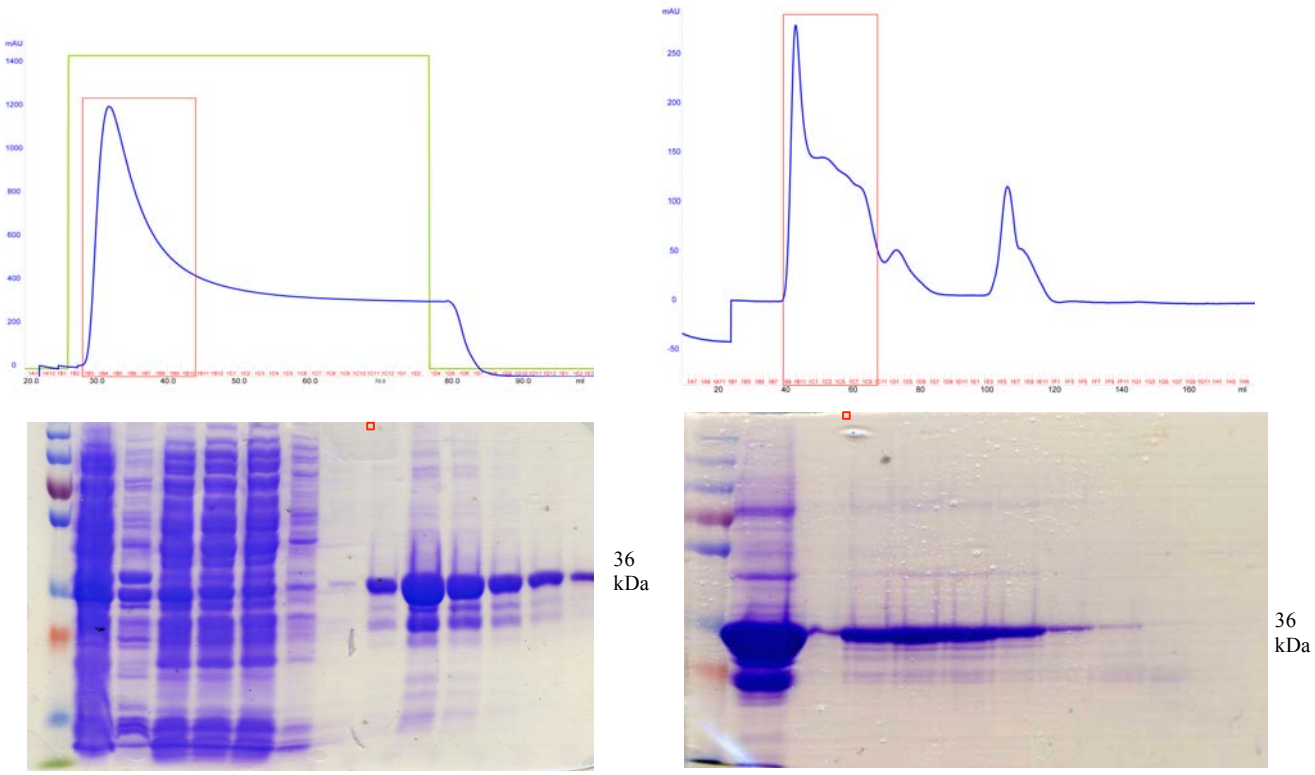


Figure 31: 2L purification of GST-IN-CTD . Left: Gel and Chromatogram following affinity purification with GSTrap column. Right: Gel and Chromatogram following purification with S200 16/60 column. Red boxes highlight pooled fractions.

Purification of ^{15}N HIS-P3C-220-288 (Gateway Construct)

24g of cells from a 6L production was resuspended in 240ml of lysis buffer -25mM HEPES pH 7, 1M NaCl, 10mM imidazole. Cells were lysed using a microfluidizer, passing 3 times through the microfluidizer, at a pressure of 1500psi. Following ultracentrifugation at 185000xg for 1 hour, the supernatant was loaded on a 5ml HisTrap FF Crude column with a flow rate of 3ml/min using the AKTA purifier. Protein was eluted using a gradient up to 500mM Imidazole. The amount of protein after affinity purification was 8mg. Protein concentration/quality was analyzed using the Nanodrop. Subsequently, the protein sample was concentrated using the Amicon Ultra 15ml with a MWCO of 3kDa for the next purification step. A second step of purification was carried out using the S7516/60 column in 25mM HEPES pH 8, 1M NaCl gave a final amount of 1mg protein per 6 liters of culture.

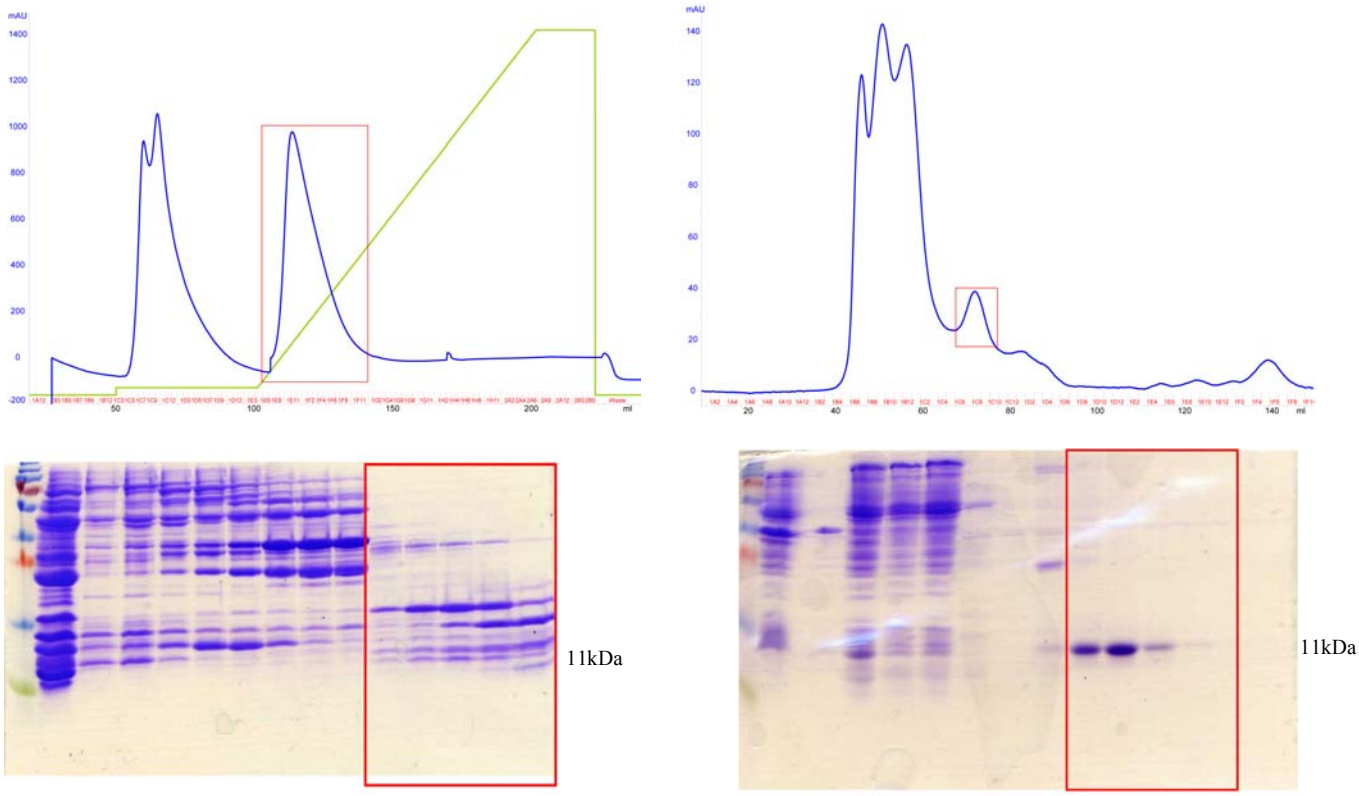


Figure 32: 6L Purification of HIS-IN-CTD (Gateway Construct) . Left: Gel and Chromatogram following affinity purification with HISTrap FF crude column. Right: Gel and Chromatogram following purification with S75 16/60 column. Red boxes highlight pooled fractions

Purification of HIS-220-270 (pET15b construct – pH 7)

7.3g of cells from a 2L production was resuspended in 73ml of lysis buffer -25mM HEPES pH 7, 1M NaCl, 10mM imidazole. Cells were lysed using a sonicator with a 13mm probe for 8 minutes with 2 sec on/off pulse at 40% amplitude at 4°C. Following ultracentrifugation at 185000xg for 1 hour, the supernatant was loaded on a 1ml HisTrap FF Crude column with a flow rate of 1ml/min using the AKTA purifier. Protein was eluted using a gradient up to 500mM Imidazole. The amount of protein after affinity purification was about 20mg – 10mg/L of culture. Subsequently, the protein sample was concentrated using the Amicon Ultra 15ml with a MWCO of 3kDa for the next purification step. A second step of purification was carried out using the S75 16/60 column in 25mM HEPES pH 8, 1M NaCl gave a final amount of 7.5mg of protein/L of culture.

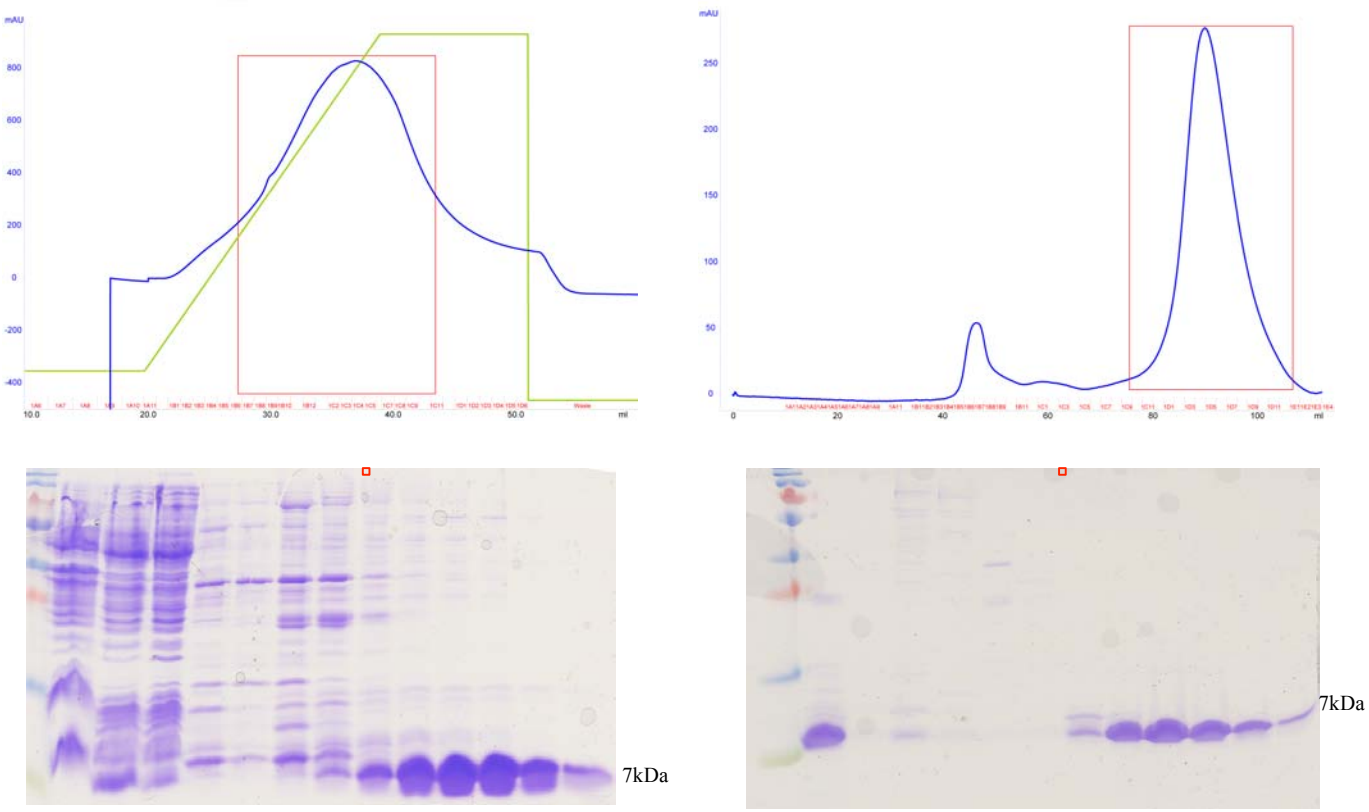


Figure 33: 2L Purification of HIS-IN-CTD 220-270 at pH 7 . Left: Gel and Chromatogram following affinity purification with HISTrap FF crude column. Right: Gel and Chromatogram following purification with S75 16/60 column. Red boxes highlight pooled fractions.

Purification of HIS-220-270 (pET15b construct – pH 8)

7.5g of cells from a 2L production was resuspended in 75ml of lysis buffer -25mM HEPES pH 7, 1M NaCl, 10mM imidazole. Cells were lysed using a sonicator with a 13mm probe for 8 minutes with 2 sec on/off pulse at 40% amplitude at 4°C. Following ultracentrifugation at 185000xg for 1 hour, the supernatant was loaded on a 1ml HisTrap FF Crude column with a flow rate of 1ml/min using the AKTA purifier. Protein was eluted using a gradient up to 500mM Imidazole. The amount of protein after affinity purification was about 24mg – 12mg/L culture. Subsequently, the protein sample was concentrated using the Amicon Ultra 15ml with a MWCO of 3kDa for the next purification step. A second step of purification was carried out using the S7516/60 column in 25mM HEPES pH 8, 1M NaCl gave a final amount of 8.5mg protein/L of culture.

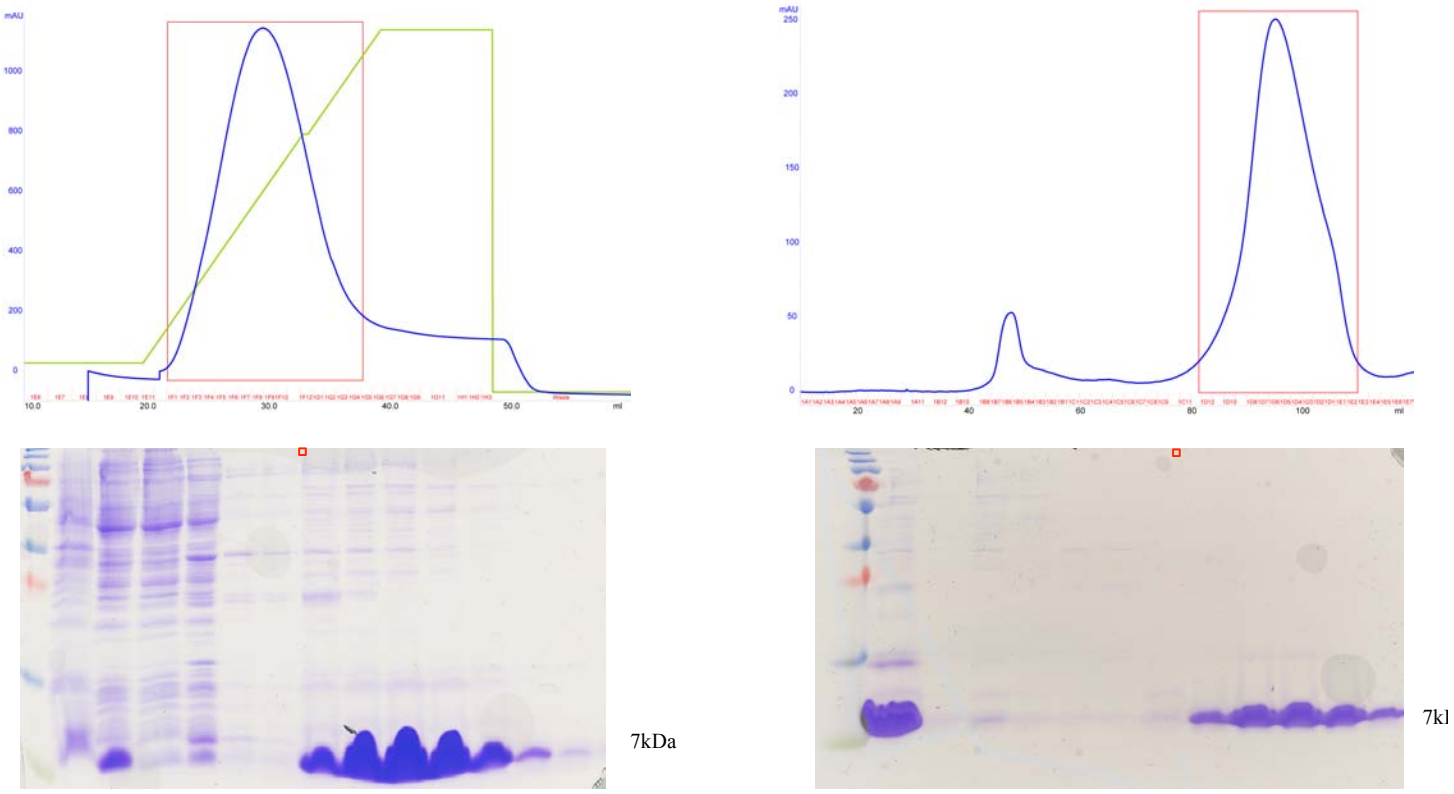


Figure 34: 2L Purification of HIS-IN-CTD 220-270 at pH 8. Left: Gel and Chromatogram following affinity purification with HISTrap FF crude column. Right: Gel and Chromatogram following purification with S75 16/60 column. Red boxes highlight pooled fractions.

Purification of ^{15}N ^{13}C HIS-220-270 (pET15b construct – pH 8)

7.3g of cells from a 2L production was resuspended in 73ml of lysis buffer -25mM HEPES pH 7, 1M NaCl, 10mM imidazole. Cells were lysed using a sonicator with a 13mm probe for 8 minutes with 2 sec on/off pulse at 40% amplitude at 4°C. Following ultracentrifugation at 185000xg for 1 hour, the supernatant was loaded on a 1ml HisTrap FF Crude column with a flow rate of 1ml/min using the AKTA purifier. Protein was eluted using a gradient up to 500mM Imidazole Protein was eluted using a gradient up to 500mM Imidazole. The amount of protein after affinity purification was about 5mg in 2L culture. Subsequently, the protein sample was concentrated using the Amicon Ultra 15ml with a MWCO of 3kDa for the next purification step. A second step of purification was carried out on half of the protein using the S7510/300 column in 25mM HEPES pH 8, 1M NaCl, giving a final amount of 1mg protein/L of culture. This is a significant improvement from the Gateway construct, where the yield was 1mg protein/6L culture.

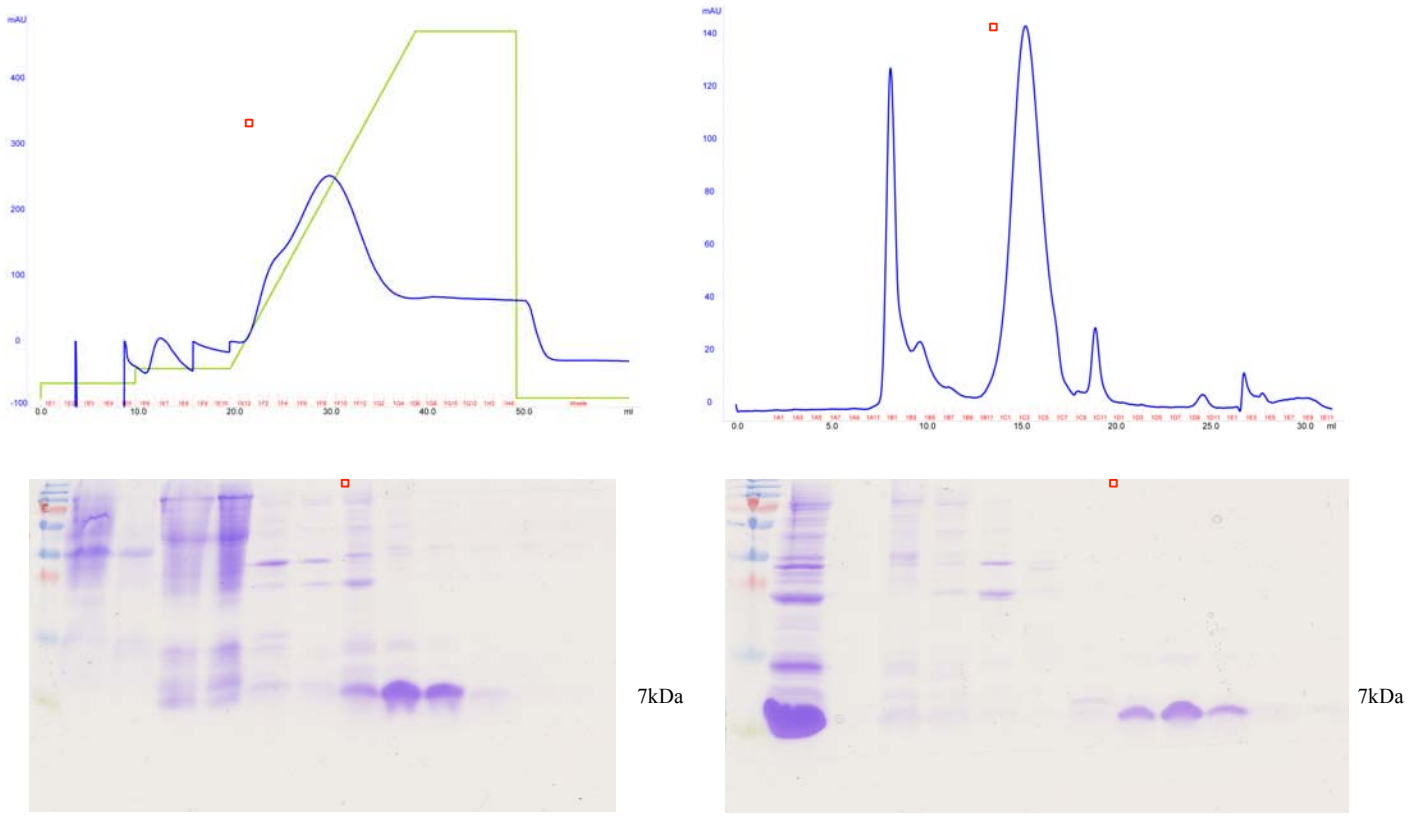


Figure 35: 2L Purification of double labelled HIS-IN-CTD 220-270 at pH 8. Left: Gel and Chromatogram following affinity purification with HISTrap FF crude column. Right: Gel and Chromatogram following purification with S75 16/60 column. Red boxes highlight pooled fractions.

Biochemical Results

MicroScale Thermophoresis

MST experiments were performed with the aim of obtaining binding constants for the interaction between IN-CTD and fluorescent peptides corresponding to mono-, di-, tri-, and non-methylated peptides.

Starting from a stock concentration of 370µM of GST-IN-CTD in 25mM HEPES pH 8, 2mM βME, 2mM $MgCl_2$ and 150mM NaCl, 1:1 serial dilutions of protein were performed, ranging from 185µM to 5nM. Each peptide was mixed with diluted protein to a final concentration of 0.5µM. After addition, the protein and peptide mixture were allowed to equilibrate for 15 minutes at room temperature. All measurements were made using the Nanotemper Monolith NT.015 instrument with laser-on time 30sec and laser-off time 5sec at 20% LED and 20% MST IR-Laser power. All experiments were performed in triplicates.

Similarly, for the competition experiments, 1:1 serial dilutions of protein were performed, ranging from 185µM to 5nM. For each protein sample, the non-fluorescent peptide (H4K20Me0) was added to a final concentration of 0.5µM Subsequently, 0.5µM of the fluorescent monomethylated peptide was added to the mixture and incubated for 15 minutes at room temperature. The experiments were carried out at 20% LED and 20% MST IR-Laser power with a laser-on time 30sec and laser-off time 5sec. This experiment was also performed in triplicates.

Results obtained from these experiments show that GST-IN-CTD is able to bind all peptides in these experimental conditions. However, GST-IN-CTD shows preferential binding to H4K20Me1 with a Kd of 0.8µM± 0.1. It binds to H4K20Me2 with a Kd of 3.75µM±0.6, H4K20Me3 with a Kd of 5.2µM±0.5, and H4K20Me0 with a Kd of 4.7µM±1.0. In the presence of competing non-fluorescent H4K20Me0, the fluorescent H4K20Me1 binds to GST-IN-CTD with a similar Kd 1.83µM, suggesting a preference for monomethylation. Put together, this data suggests that GST-IN-CTD preferentially binds with the monomethylated peptide, binding with the highest Kd of 0.8µM. The presence of a competing peptide doesn't change the Kd significantly, indicating that even when other methylation states are available, the IN-CTD preferentially binds the monomethylated peptide.

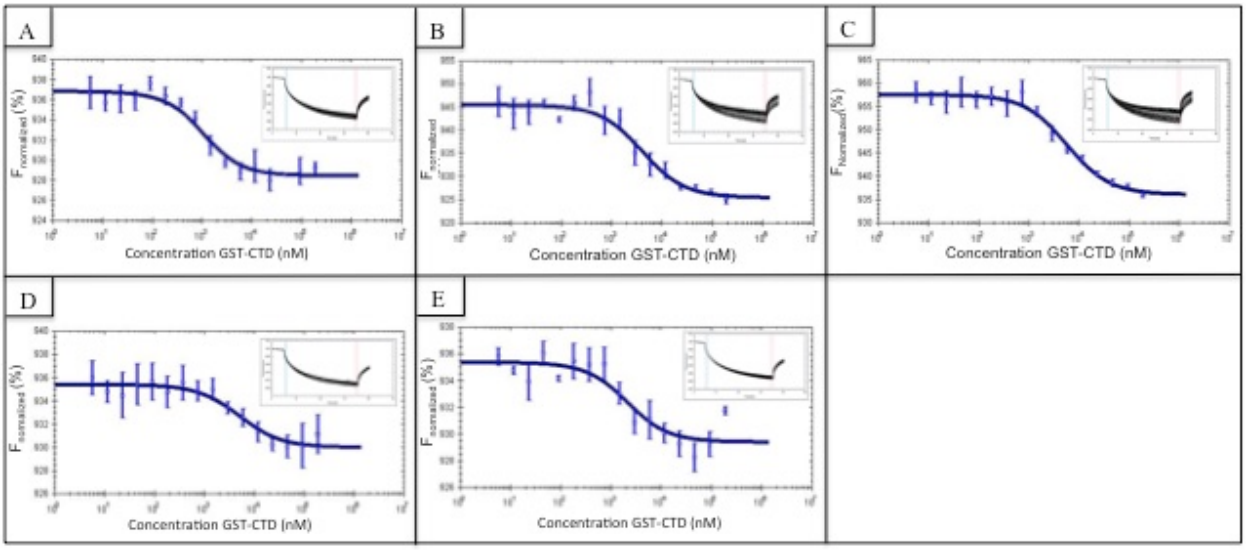


Figure 36: Data from MST experiments. GST-IN-CTD binds to A) the monomethylated peptide (H4K20Me1) with a Kd of 0.8µM± 0.13, B) the dimethylated peptide (H4K20Me2) with a Kd of 3.75µM±0.6, C) the trimethylated peptide (H4K20Me3) with a Kd of 5.2 µM±0.5, and D) the nonmethylated peptide (H4K20Me0) with a Kd of 4.7µM±1.0. E) In the presence of a equimolar concentration of non-fluorescent competing peptide (H4K20Me0), the monmethylated peptide binds with a Kd of 1.83µM±0.4. All the data was analyzed using the NTAnalysis software. Data are represented as mean ± SD at each point from three independent replicates using the signal from Thermopheresis.

trNOESY-NMR

trNOESY was used as a complementary method to confirm the interaction between GST-IN-CTD and the H4K20Me1 peptide, by observing NOEs changes on the peptide. In the absence of protein, 750µM of H4K20Me1 peptide in 25mM HEPES pH 8, 2mM $MgCl_2$, 2mM βME, 150mM NaCl exhibits positive NOEs only (spectrum on the left) due to short correlation times and slow NOE accumulation. In the presence of 100µM GST-IN-CTD, complex formation is observed, leading to faster correlation times and faster NOE accumulation, changing the NOEs signal on the peptide from positive (spectrum on the left) to negative (spectrum on the right).

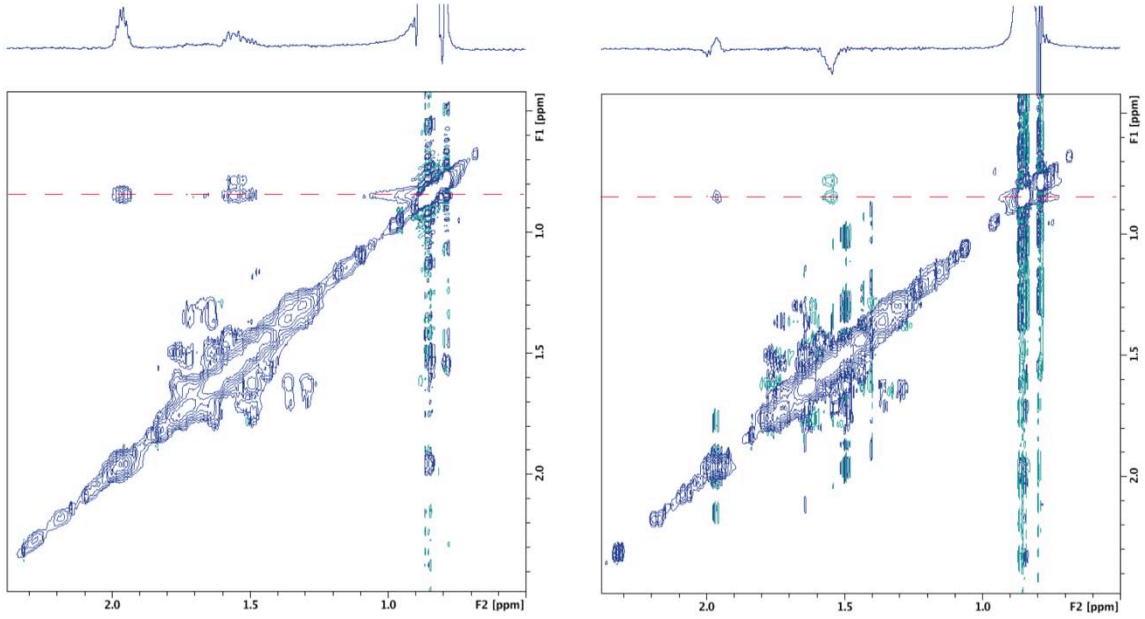


Figure 37: Data from trNOESY experiments. trNOESY spectra for the solution containing both the peptide and the protein (top spectra) and the control sample (bottom). A 1D projection of one row (signaled by a red dashed line) containing several peaks of the peptide is displayed above each spectrum, illustrating the change in sign.

To confirm the specificity of the reaction for the IN-CTD and not GST, the experiment was repeated with 75µM of GST only or 75µM GST-IN-CTD. Upon addition of 75µM GST alone to 750µM of H4K20me1 peptide, there is no change in the NOEs signal as there is no magnetization transferred to the peptide, indicating that there is no interaction between GST and H4K20Me1. In the presence of GST-IN-CTD, a change is NOEs signal is observed again.

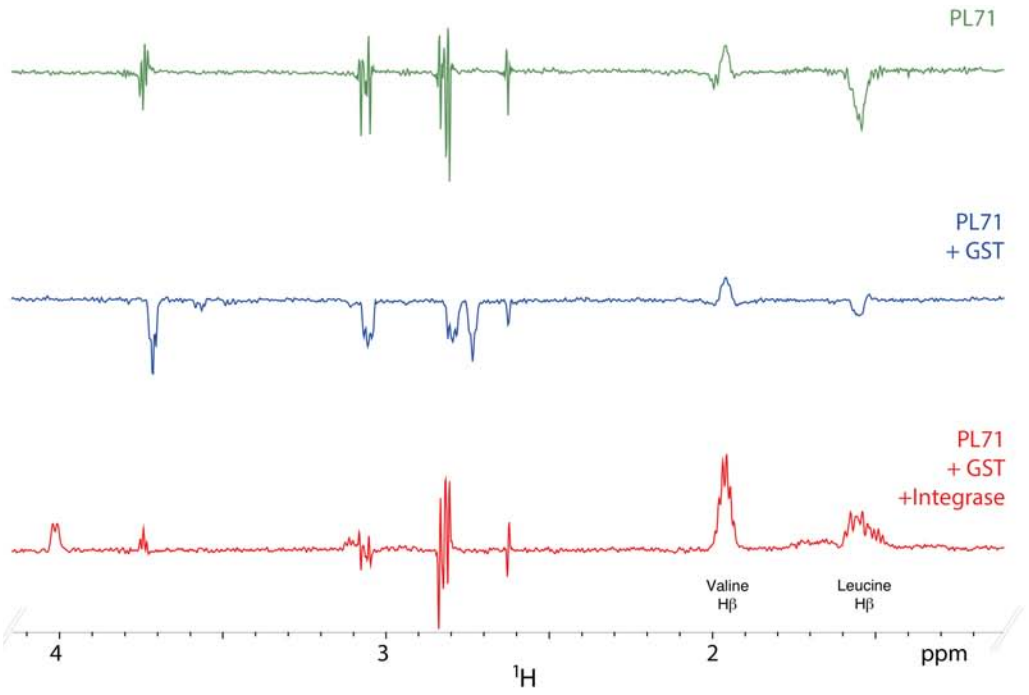


Figure 38: 1D slices extracted at 0.8 ppm from several trNOESY experiments:with free H4K20me1 peptide (green), upon addition of 75 µM of GST (blue) and with 75 µM of GST-IN-CTD (red)

WaterLOGSY

WaterLOGSY experiments were performed in parallel with trNOESY experiments, in order to further confirm the interaction between GST-IN-CTD and the H4K20Me1 peptide. In the absence of protein, 750µM of H4K20Me1 peptide in 25mM HEPES pH 8, 2mM $MgCl_2$, 2mM βME, 150mM NaCl has positive NOEs only (blue, lower spectrum), resulting from faster tumbling due to interaction with bulk water only. However, upon addition of 100µM GST-IN-CTD, there is a change in the NOEs signal on the peptide from positive to negative (red, upper spectrum). This change is indicative of the presence of protein-peptide complex.

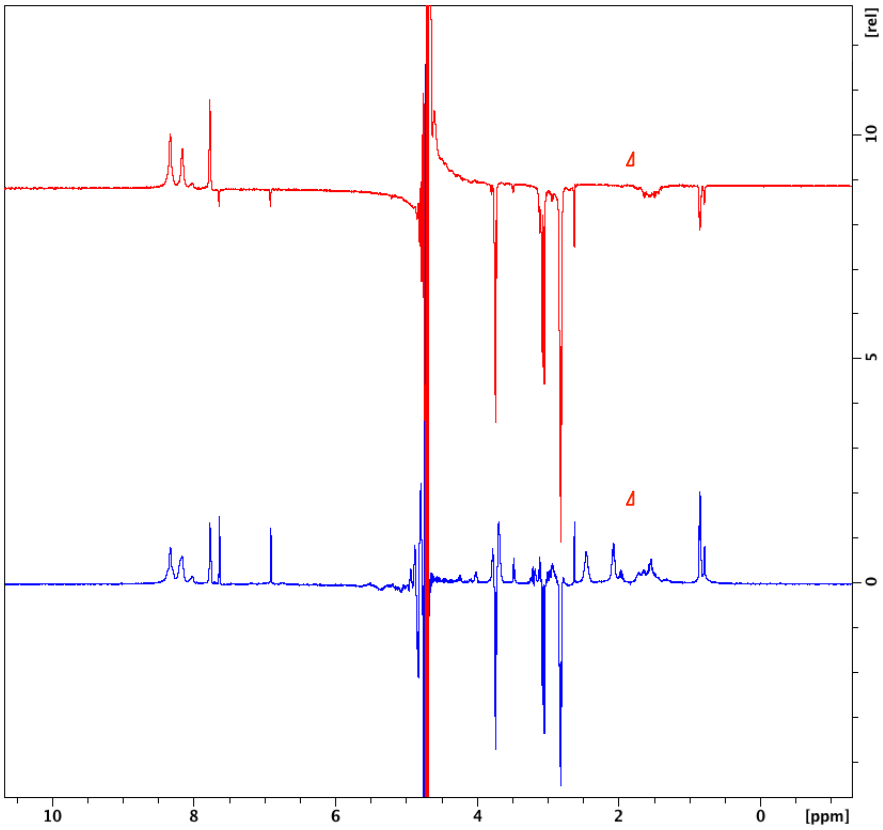


Figure 39: Data from WaterLOGSY spectra for the solution of protein and peptide (top spectrum, red) and the control tube, containing only free peptide in buffer (bottom spectrum, blue). Change of NOEs on the peptide is highlighted by red circle

Preparation of Isotopically labeled protein (Gateway construct)

In order to better understand the interactions between IN-CTD and H4K20Me1 on the molecular level, it was important to identify the residues on the protein that interact with the peptide. One of the limitations of the GST-IN-CTD construct is its size. At 36.6kDa, the size is unfavorable for NMR, as NMR is better suited for proteins under 25kDa. Moreover, the high order oligomerization state of GST-IN-CTD would have made it difficult to characterize the ^{15}N HSQC. Therefore it was important to produce HIS tagged IN-CTD for NMR.

Initial attempts to produce His-tagged ^{15}N labeled protein (HIS-P3C-220-288) in varying conditions were largely unsuccessful. In order to determine the optimum conditions for protein production and purification for labeled protein, several tests were carried out on unlabeled protein to assess protein quality using 1D spectra. Several steps were taken to

troubleshoot and refine the protocol, and to determine the concentration needed to obtain a good signal for NMR studies. These steps would be described in the following paragraphs.

For the first attempt, bacteria were inoculated to an OD600 of 0.1 in a 5L Erlenmeyer flask in 1X M9 medium with unlabeled NH_4Cl supplemented with ampicillin. Flasks were incubated at 37°C with shaking at 220rpm. At an OD600 of 0.5 the temperature was reduced to 25°, and shaking was reduced to 190rpm, till the cells reach an OD600 of 0.8. IPTG was added to a final concentration of 0.5mM to induce the target gene. Cells were incubated overnight at 25°C for protein expression. For affinity purification, protein was batch purified using EMD Millipore PureProteome Nickel Magnetic beads in 50mM phosphate pH 7, 2mM βME, 1M NaCl, followed by gel filtration using the Superdex S75 10/300 column (GE Healthcare). 30μM of protein obtained from both GF elution peaks was assessed using NMR on the Bruker 700MHz spectrometer with 10% D2O (256 scans). Both samples were found to be largely unfolded due to the presence of broad peaks between 7 and 9ppm.

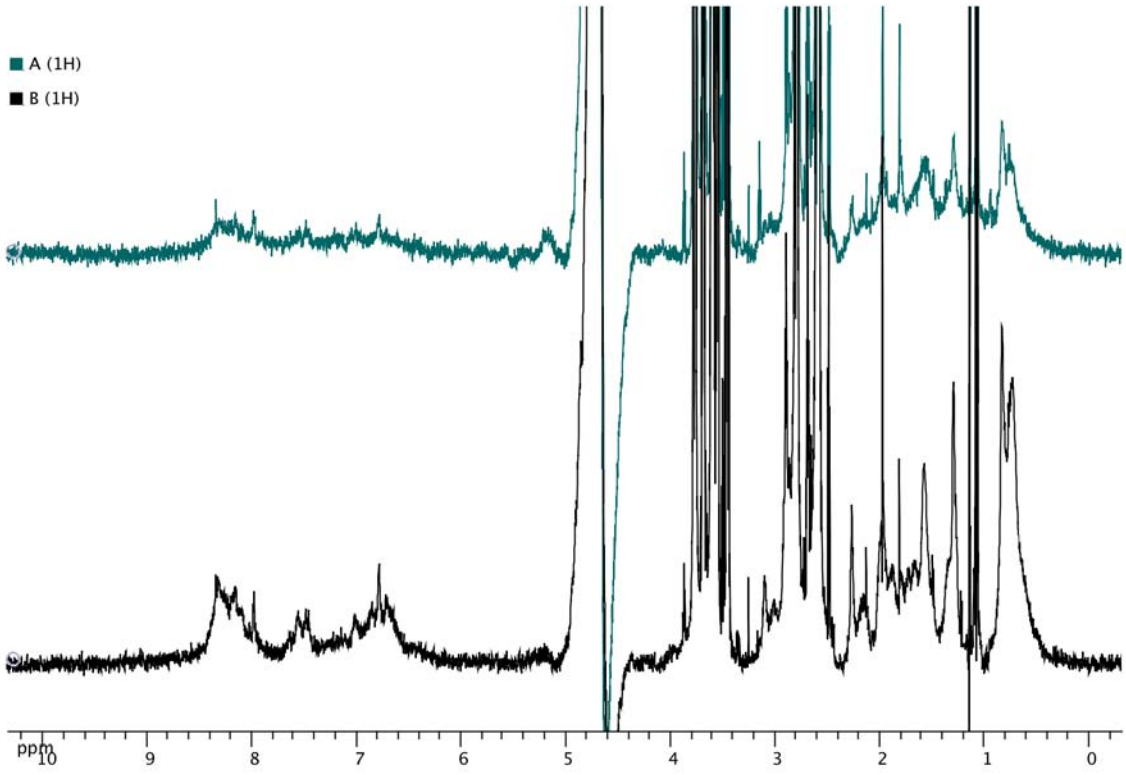


Figure 40: Superposition of 1D spectra obtained for both samples at 288K. The presence of broad peaks between 7 and 9 ppm indicate the protein was unfolded.

To investigate the effects of bacterial growth rate on protein expression and folding, another batch of protein production was performed +/- 10% sucrose added to the M9 medium. It was hypothesized that slower growth rate would allow more stable folding of the protein. Protein was batch purified, and samples analyzed on the 600MHz spectrometer with 10% D2O (512 scans), with concentrations ranging from 10μM to 50μM.

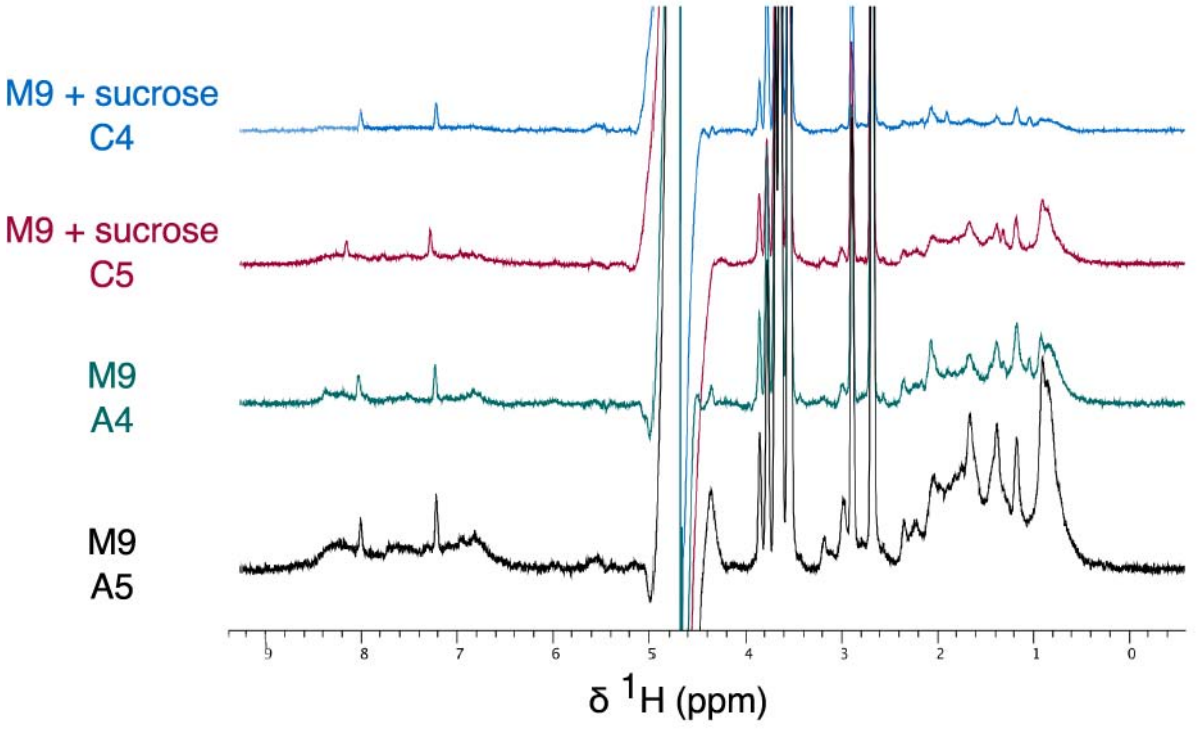


Figure 41: Superposition of 1D spectra obtained for all samples at 298K. Both protein peaks obtained from bacteria cultured with sucrose (blue and red), and without sucrose (teal and black). Protein samples were still unfolded.

Another attempt to produce well-folded protein was carried out, testing varying induction temperatures, testing 37°C and 25°C. 6M Urea was also added to the purification buffer for some samples, with the hypothesis that unfolding the protein during purification, and refolding slowly by dialysis would allow for better protein folding. Moreover, the purification buffer was changed to 25mM HEPES pH 8, 2mM βME, 2mM $MgCl_2$, 1M NaCl, the pH in which the GST-IN-CTD protein was stable. All samples were purified using the EMD Millipore PureProteome Nickel Magnetic beads, and the quality checked in 3mm tubes on the 600MHz spectrometer (256 scans). While there was a slight improvement in NH peak distribution (between 8 and 8.5ppm), the peaks were still quite broad, indicating they were

still unfolded. Moreover, traces of urea were still left over after several hours of dialysis (5.5ppm).

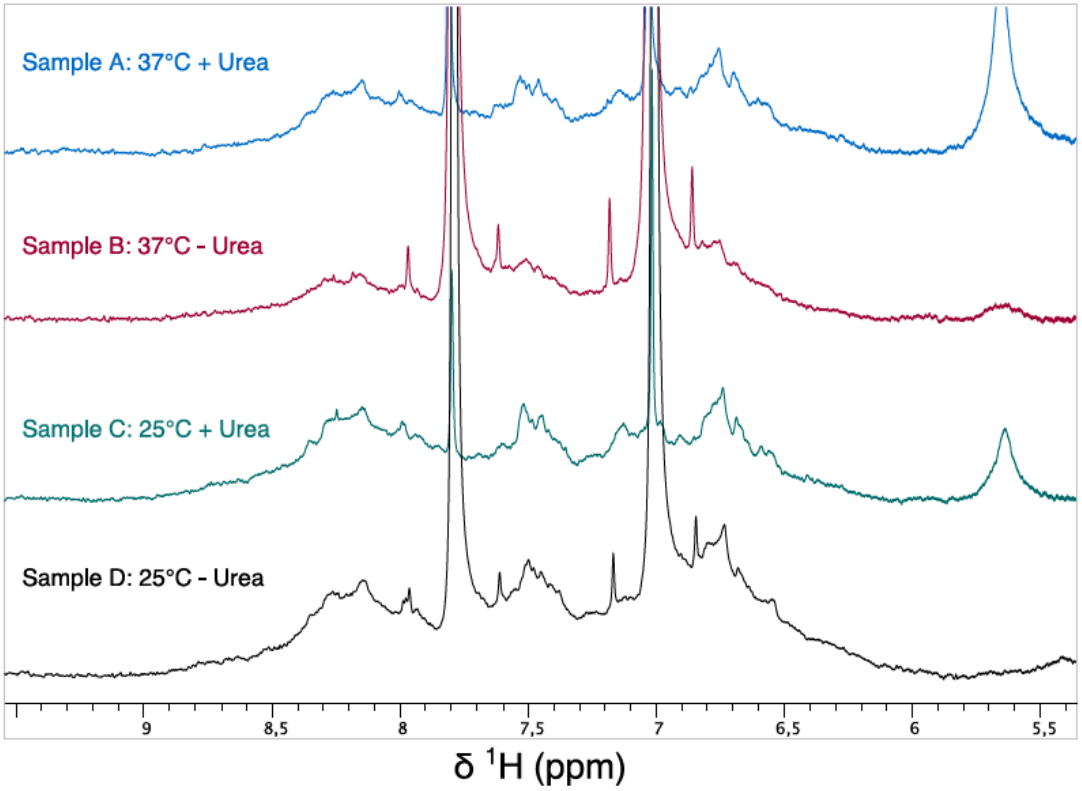


Figure 42: Superposition of 1D spectra obtained for all samples at 280K. Both protein peaks obtained from protein induced at 37°C, +/- urea (blue and red respectively), and 25°C, +/- urea (teal and black). Protein samples were still unfolded. Blue and teal spectra (37°C + urea, 25°C + urea) show the presence of urea at 5.5ppm.

Finally, the induction protocol was changed, with the induction time reduced from overnight to 4 hours, using the standard lab protocol of protein expression at 37°C, but inducing at 25°C. Bacteria were inoculated to an OD600 of 0.1 in a 5L Erlenmeyer flask in 1X M9 medium with $^{15}NH_4Cl$ supplemented with ampicillin. Flasks were incubated at 37°C with shaking at 220rpm. At an OD600 of 0.5 the temperature was reduced to 25°C, and shaking was reduced to 190rpm, till the cells reached an OD600 of 0.8. IPTG was added to a final concentration of 0.5mM to induce the target gene. Cells were harvested after 4 hours of protein expression at 25°C.

Following both steps of purification and subsequent dialysis, 10μM of protein was obtained in 25mM HEPES pH 8, 2mM $MgCl_2$, 2mM βME and 150mM NaCl, and analyzed on 700MHz spectrometer by ^{15}N HSQC at 280K, 293K and 303K. At this concentration, only 5 resonance

peaks could be detected from the protein, indicating that the concentration was too low to provide any information.

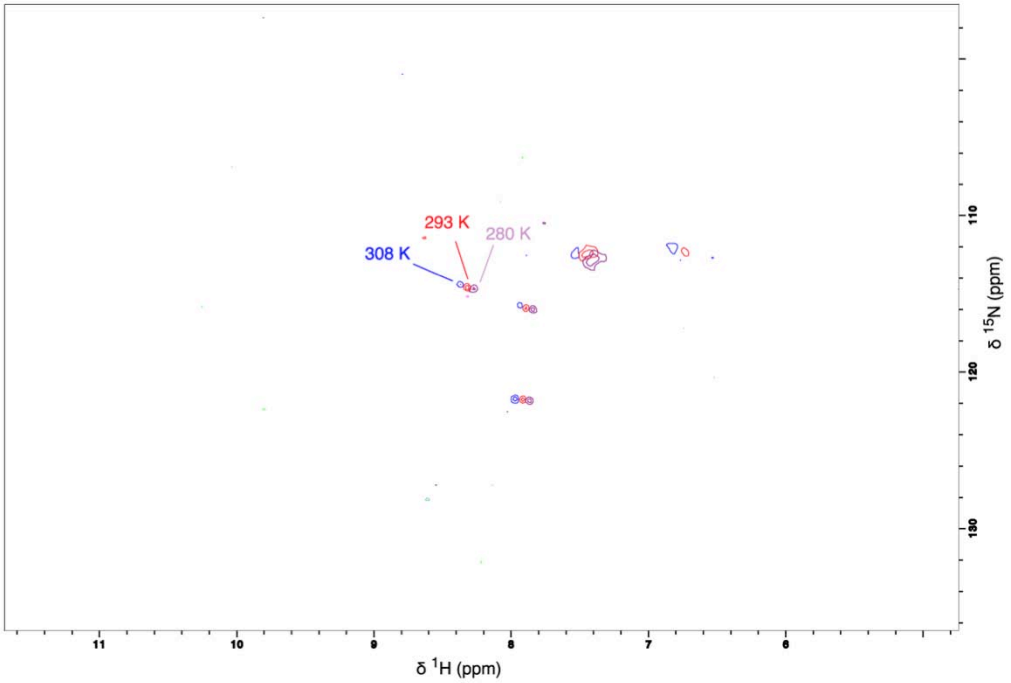


Figure 43: Superposition of ^{15}N HSQC performed at various temperatures :at 280K (purple), 293K (red) and 303K (blue). At all temperatures, the protein concentration was too low to provide any further information

Results from NMR experiments

Protein Quality Control (Gateway Construct)

A large-scale (6L) purification was performed in order to increase protein concentration and obtain adequate signal from the ^{15}N HSQC experiment. ^{15}N HSQC SOFAST experiments were recorded on 90μM of ^{15}N HIS IN-CTD in 25mM HEPES pH 7, 2mM βME, 1M NaCl with 10% D2O, on a 700 MHz spectrometer in a 3mm tube, with 416 scans at 280K, and a receiver gain of 912, lasting about 6 hours. The pH was changed to 7, in order to improve HSQC spectra, due to slower NH exchange. A higher protein concentration led to an increase in observed resonance peaks. Some peaks between 8.5ppm and 10ppm are well dispersed indicating that corresponding regions of the protein are well folded and globular. However, peaks that appear between 7.5ppm and 8.5ppm are not well dispersed and indicate that there are unfolded or disordered regions in the protein. The peak at 10ppm/130ppm corresponds to tryptophan residues that are in a homogenous environment that is well folded. Moreover, the number of resonance peaks observed did not correspond to the number of residues in the protein.

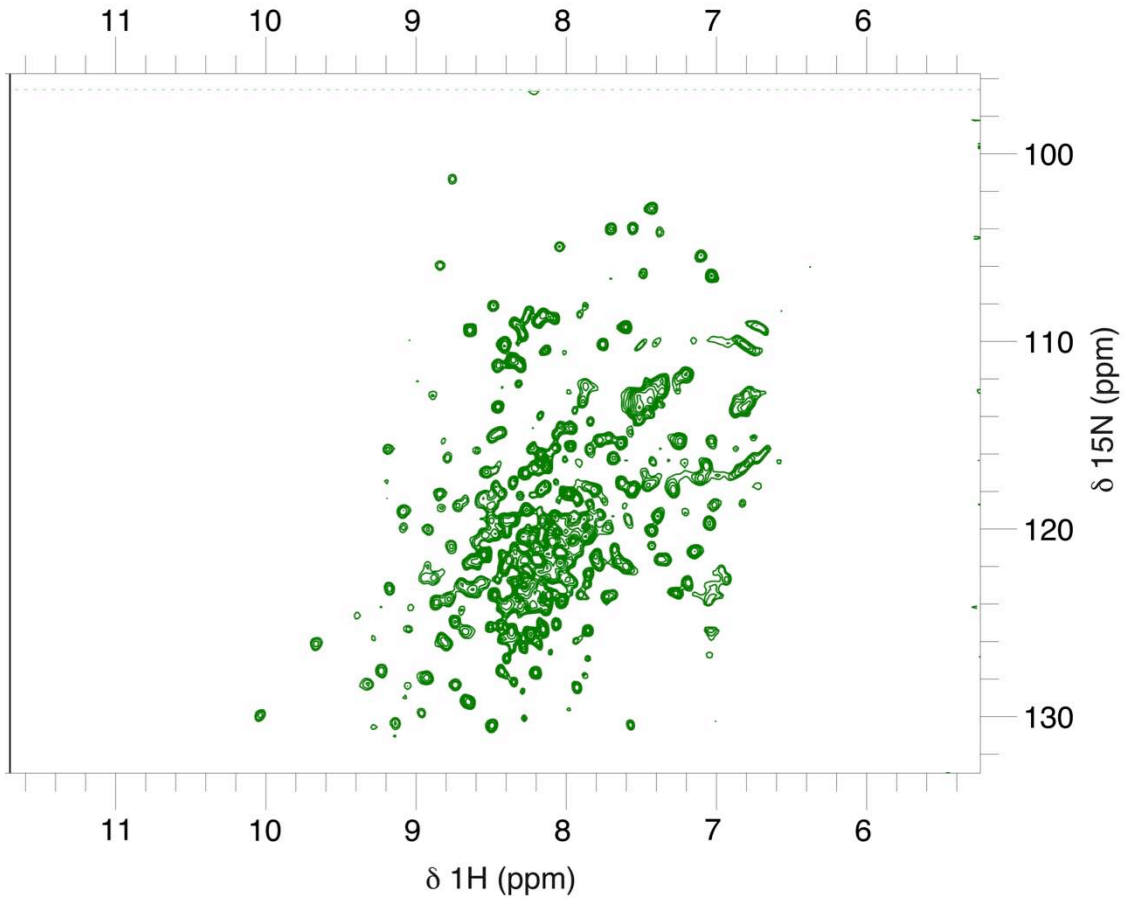


Figure 44: ^{15}N HSQC of HIS-IN CTD 220-288 in 25mM HEPES pH 7, 2mM βME, 1M NaCl. Some peaks of the protein are well dispersed (between 8.5ppm and 10ppm), while some regions are disordered with overlapping peaks (between 7.5ppm and 8.5ppm)

^{15}N HSQC data was recorded with protein in 25mM HEPES pH 7, 2mM βME, 500mM NaCl, using the same parameters stated above, in order to investigate if lower salt improves spectra quality by changing the conformation of the protein. Changing salt concentrations did not lead to significant changes in spectra quality.

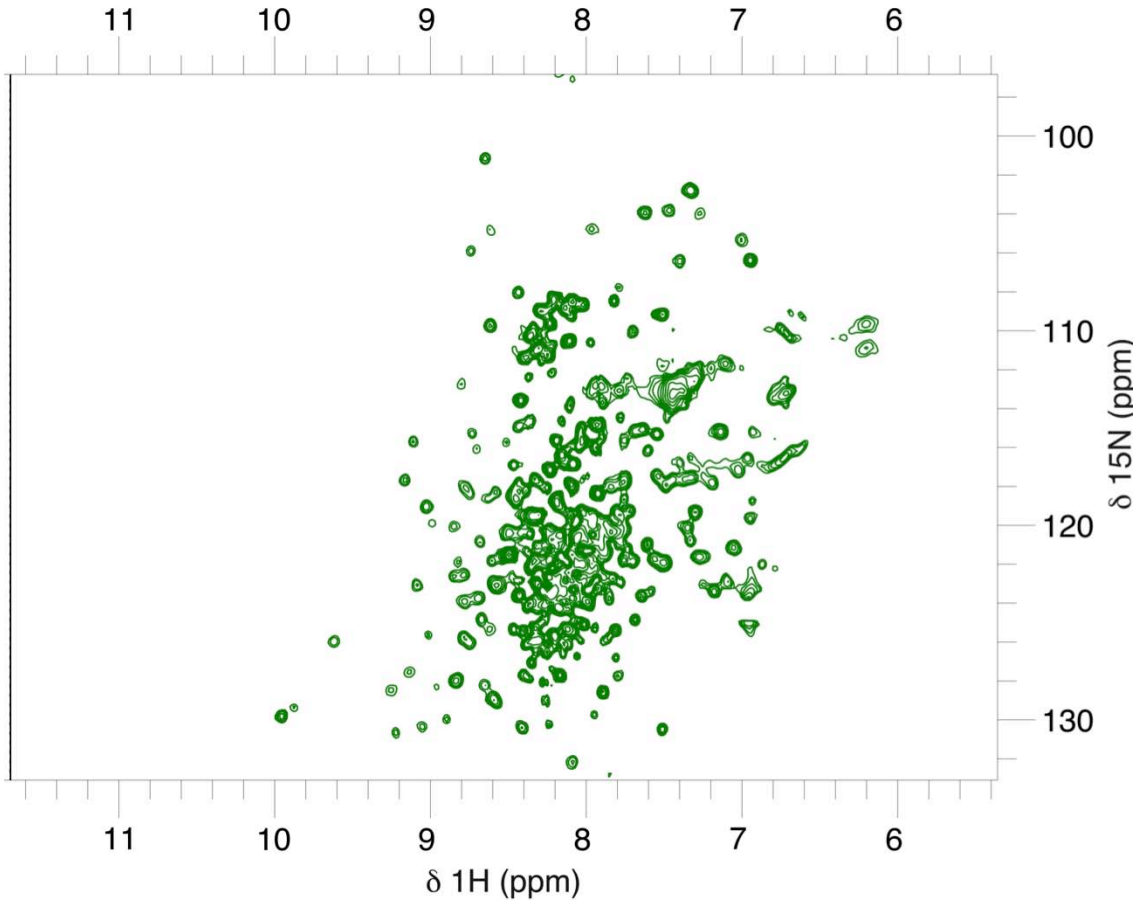


Figure 45: ^{15}N HSQC of HIS-IN-CTD 220-288 in 25mM HEPES pH 7, 2mM βME, 500mM NaCl.

Peptide Interactions

Although the spectra were of bad quality, ^{15}N HSQC SOFAST experiments were recorded in the presence of peptide in order to determine if interactions could be observed in these conditions. 90μM of ^{15}N HIS IN-CTD in 25mM HEPES pH 7, 2mM βME, 1M NaCl with 10% D2O, was mixed with 1mM and 2mM H4K20me1 peptide. Data was collected on a 700 MHz spectrometer in a 3mm tube, with 416 scans at 280K, and a receiver gain of 912, lasting about 6 hours each. Chemical shifts changes were upon the addition of increasing concentrations of peptide, suggesting the presence of an interaction. Moreover, the chemical shift changes were limited to certain resonance peaks, meaning that they were specific to residues that were affected by the presence of the peptide, or directly involved in the interaction with the peptide.

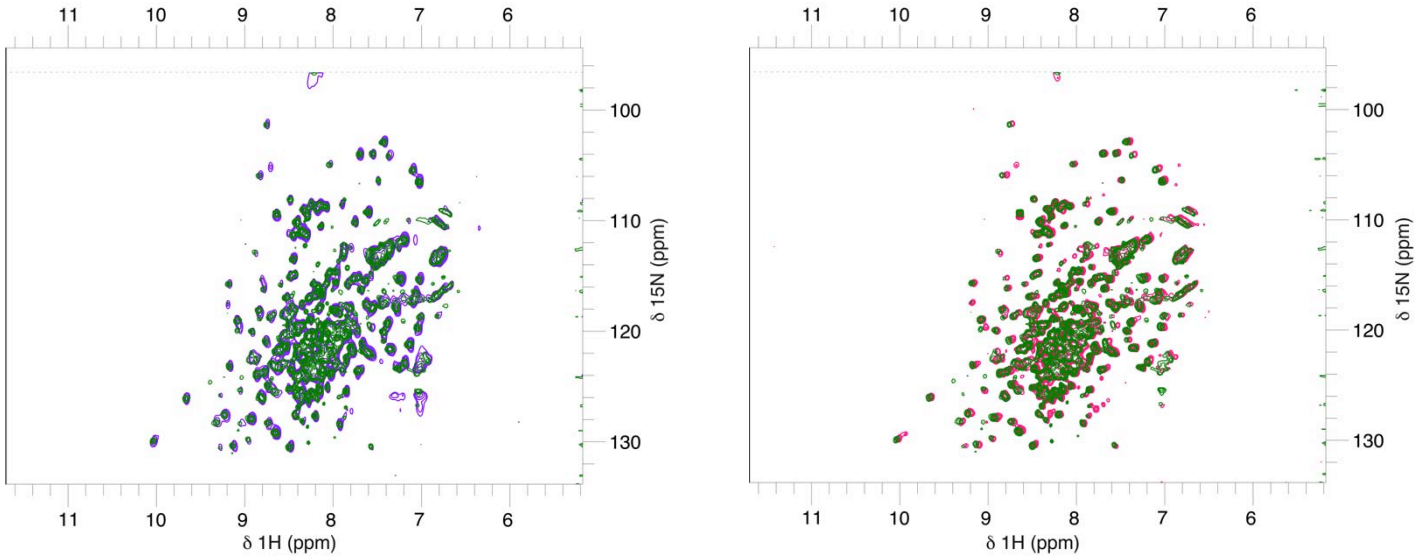


Figure 46: Peptide interactions with old construct. Left: ^{15}N HSQC of HIS IN-CTD in 25mM HEPES pH 7, 2mM βME, 1M NaCl with 1mM peptide: protein only (green), protein + 1mM peptide (purple). Right: ^{15}N HSQC of HIS IN-CTD in 25mM HEPES pH 7, 2mM βME, 1M NaCl with 2mM peptide: protein only (green), protein + 2mM peptide (pink).

Protein Quality Control (New Constructs)

A further inspection of the sequence of the Gateway construct revealed that it contained extra 32 residues at the N-terminus before the sequence of the IN-CTD. It was hypothesized that these residues were unfolded, and contributed to the bad quality of the spectrum. Therefore, the IN-CTD was re-cloned into pET15b vector, removing the extra residues. A full-length version (220-288) and a truncated version (220-270) were produced.

Protein Folding

Two versions of HIS IN-CTD were cloned into pET15b vectors: the full length CTD (HIS-IN-CTD 220-288) and a truncated version (HIS-IN-CTD 220-270). ^{15}N labeled protein was produced and purified as described previously, and the quality was tested using ^{15}N HSQC. 170μM of each protein in 25mM HEPES pH 7, 2mM βME, 1M NaCl with 10% D2O was analyzed on the 700 MHz spectrometer in a 3mm tube, with 128 scans at 298K, and a receiver gain of 912.

Results from ^{15}N HSQC experiments on both proteins indicate that the quality of the protein was significantly improved by changing the plasmid. In both cases, the resonance peaks were better dispersed, indicating that both proteins were both globular and better folded than the

protein in the gateway vector. Additionally, the number of peaks significantly reduced, and was closer to the number of residues in the protein, compared to the gateway vector.

In comparison to HIS-IN-CTD 220-270, HIS-IN-CTD 220-288 contains some unfolded regions, evident by overlapping peaks visible between 7.9ppm and 8.5ppm, indicating that residues 271-288 are flexible and lack a stable conformation under these conditions. The spectrum of HIS-IN-CTD 220-270 indicates that the protein quality is better than HIS-IN-CTD 220-288.

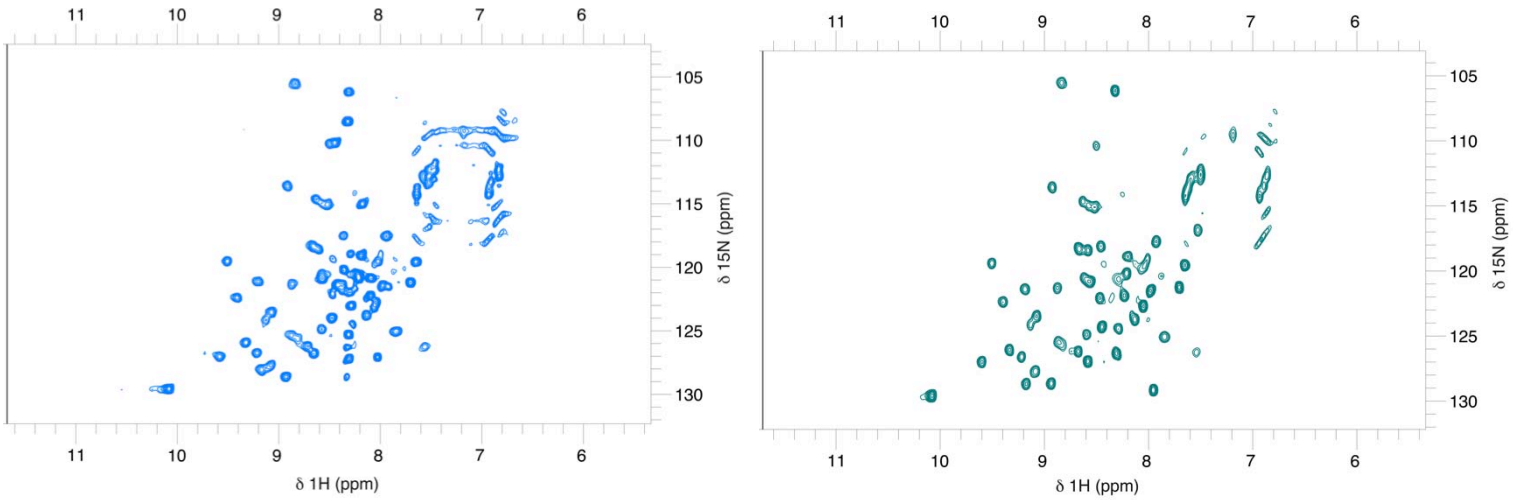


Figure 47: HSQC spectra of new constructs. . Left: ^{15}N HSQC of HIS-IN-CTD 220-288 in 25mM HEPES pH 7, 2mM βME, 1M NaCl (blue), Right: ^{15}N HSQC of HIS-IN-CTD 220-270 in 25mM HEPES pH 7, 2mM βME, 1M NaCl (teal). Overlapping peaks between 7.9ppm and 8.5ppm in the spectrum on the left are better dispersed in the spectrum on the right

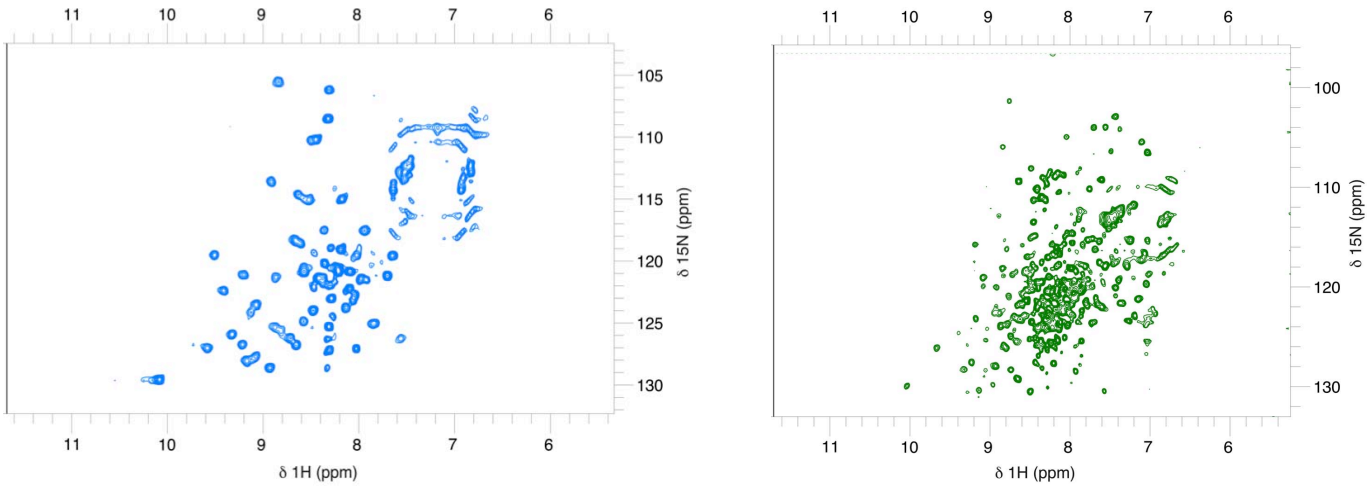


Figure 48: HSQC spectra comparing pET15b and Gateway constructs . Left: 15N HSQC spectra of HIS-IN-CTD 220-288 in pET15b vector. Right: 15N HSQC spectra of HIS-IN-CTD 220-288 in Gateway vector

Protein Stability

In order to determine the stability of the protein at higher temperature or over a longer period of time, ^{15}N HSQC spectra of HIS-IN-CTD 220-270 in 25mM HEPES pH 7, 2mM βME, 1M NaCl in a 3mm tube with 10% D2O on a 700 MHz spectrometer were collected at different temperatures. Initial data collection was at 283K with 128 scans and a receiver gain of 912. Subsequently, the temperature was increased to 293K, 303K, with a ^{15}N HSQC recording at each temperature. Finally, the temperature was reduced to 283K. The quality of the ^{15}N HSQC was not affected by the changes in temperature, as resonance peak shifts observed upon temperature change were expected. Moreover, the spectra of the first and last recordings at 283K completely overlap, showing that the protein quality was not adversely affected by changing the temperature.

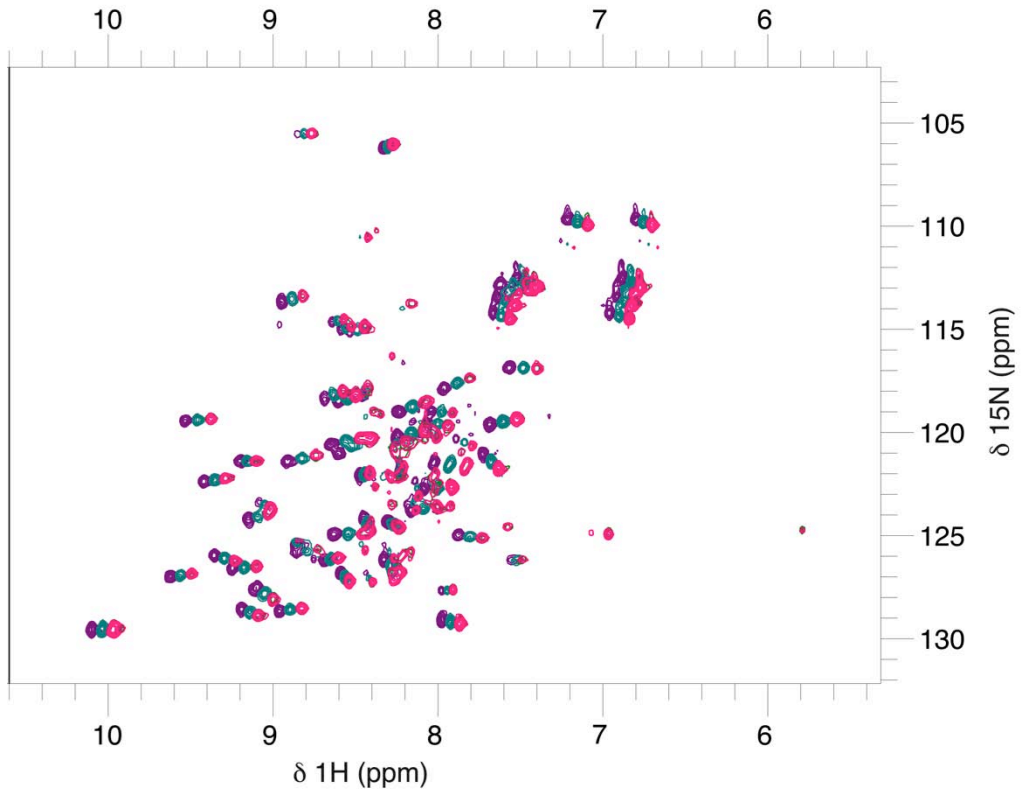


Figure 49: Protein Stabilty at varying temperatures.. Left: ^{15}N HSQC of HIS-IN-CTD 220-270 in 25mM HEPES pH 7, 2mM βME, 1M NaCl at different temperatures (pink- 283K, teal - 293K, purple – 303K). The spectrum in green (second 283K) is not visible as it completely overlaps with the pink spectra

Peptide Interactions with HIS-IN-CTD 220-288

^{15}N HSQC spectrum was recorded with 170µM HIS-IN-CTD 220-288 in 25mM HEPES pH 7, 2mM βME, 1M NaCl, plus 2mM H4K20Me1 peptide in a 3mm tube with 10% D2O on a 700 MHz spectrometer. Experiments were recorded at 298K with 128 scans and a receiver gain of 912.

In the presence of 2mM peptide, three effects are observed: disappearance of resonance peaks, new resonance peaks showing up, and chemical shift changes. These effects are specific and limited to certain peaks, indicating that these peaks are affected by peptide binding.

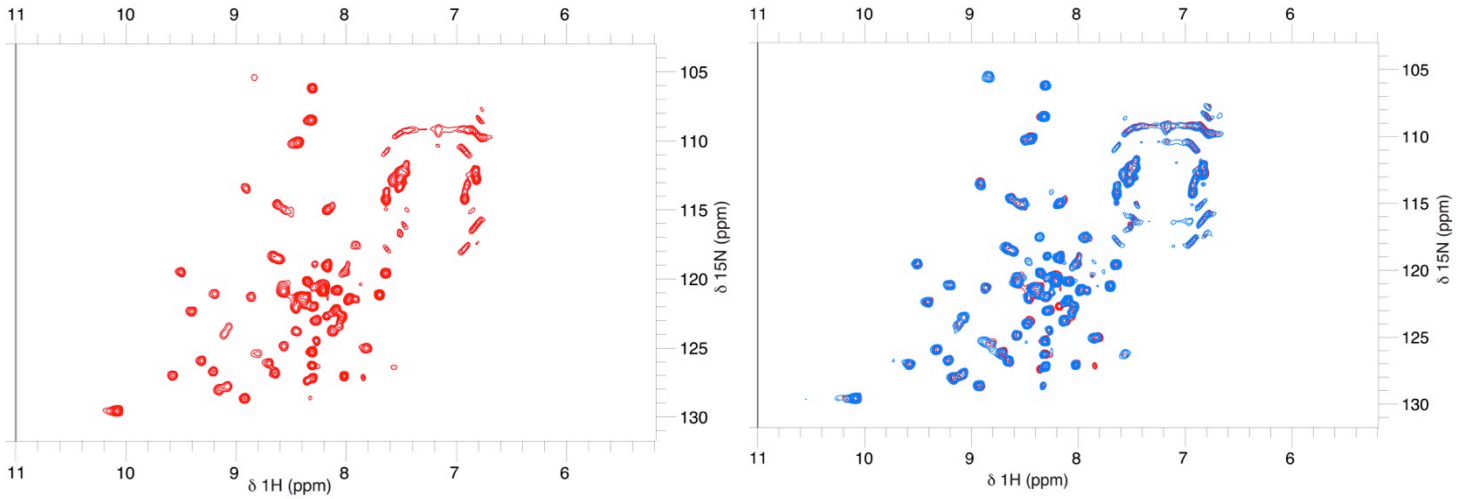


Figure 50: Peptide interactions in 1M NaCl . Left: 15N HSQC of HIS-IN-CTD 220-288 plus 2mM peptide in in 25mM HEPES pH 7, 2mM βME, 1M NaCl. Right: Superposition of spectra plus 2mM peptide (red) and without peptide (skyblue)

Peptide Interactions with HIS-IN-CTD 220-270

Titration experiments with H4K20Me1 peptide were performed on 170µM HIS-IN-CTD 220-270 in 25mM HEPES pH 7, 2mM βME, 1M NaCl in a 3mm tube with 10% D2O on a 700 MHz spectrometer. Experiments were recorded at 298K with 256 scans and a receiver gain of 912. ^{15}N HSQC spectrum was recorded on protein only, and then with increasing concentrations of peptide (1mM and 2mM).

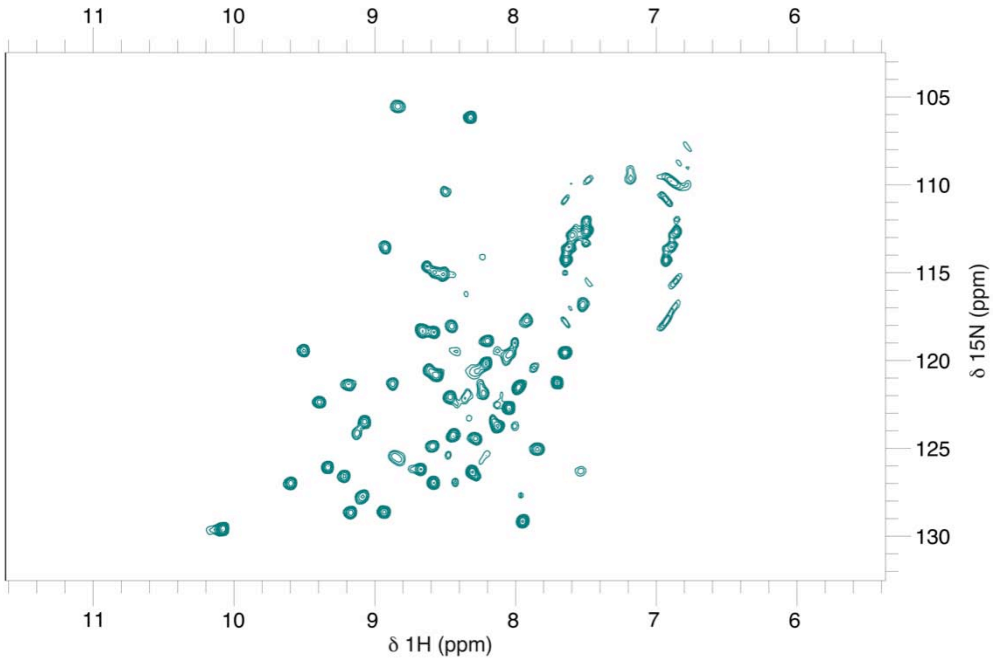


Figure 51: ^{15}N HSQC of HIS-IN-CTD 220-270 in 25mM HEPES pH 7, 2mM βME, 1M NaCl at 298K

In the presence of 1mM peptide, slight chemical shifts are observed on some resonance peaks. Additionally, new peaks appear at other parts of the spectrum. These changes in the spectrum are representative of interactions between the protein and peptide in 1M NaCl.

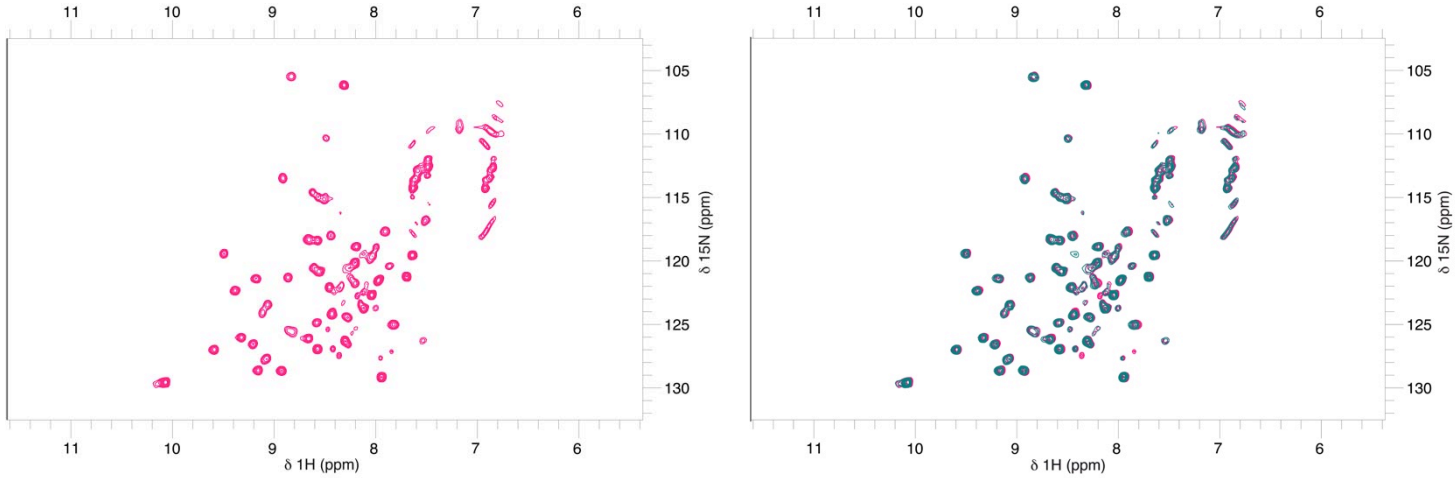


Figure 52: Interaction with 1mM peptide at pH 7 in 1M NaCl: ^{15}N HSQC of HIS-IN-CTD 220-270 in 25mM HEPES pH 7, 2mM βME, 1M NaCl plus 1mM peptide. Right: Superposition of ^{15}N HSQC spectrum without peptide (teal) and with 1mM peptide (pink)

Fewer changes are observed when the peptide concentration is increased to 2mM. There is a further chemical shift change on one peak. Additionally, there is an increase of intensity of

the new peaks observed at 1mM peptide. This increase in intensity is indicative of a residue becoming more solvent exposed.

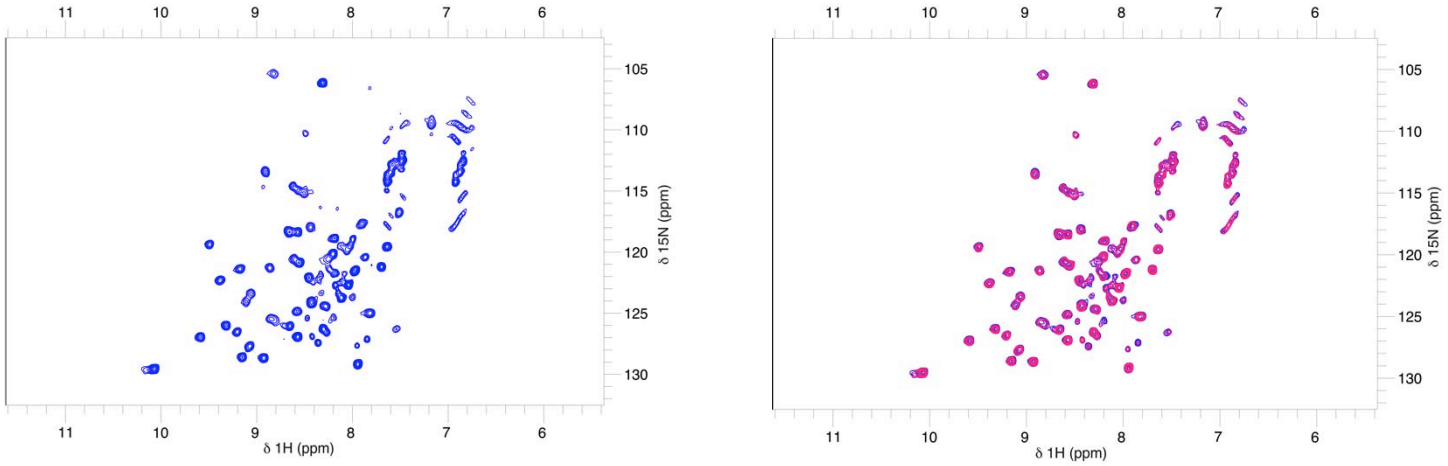


Figure 53: Interaction with 2mM peptide at pH 7 in 1M NaCl. Left: ^{15}N HSQC of HIS-IN-CTD 220-270 in 25mM HEPES pH 7, 2mM βME, 1M NaCl plus 2mM peptide. Right: Superposition of ^{15}N HSQC spectrum in 1mM peptide (pink) and with 2mM peptide (blue)

Dynamics and Interactions of HIS-IN-CTD 220-270 at pH 7

Effects of Salt Concentration on Protein Quality

In order to fully understand the interactions with the H4K20Me1 peptide, it was important to study the interactions at a lower ionic strength of 150mM NaCl that is closer to physiological conditions. HIS-IN-CTD 220-270 in 25mM HEPES pH 7, 2mM βME, 1M NaCl is purified as stated above, and the salt concentration is reduced to 150mM NaCl by dialysis.

^{15}N HSQC experiments were recorded on 170μM HIS-IN-CTD 220-270 in 25mM HEPES pH 7, 2mM βME, 150mM NaCl in a 3mm tube with 10% D2O on a 700 MHz spectrometer. Experiments were recorded at 298K with 256 scans and a receiver gain of 912.

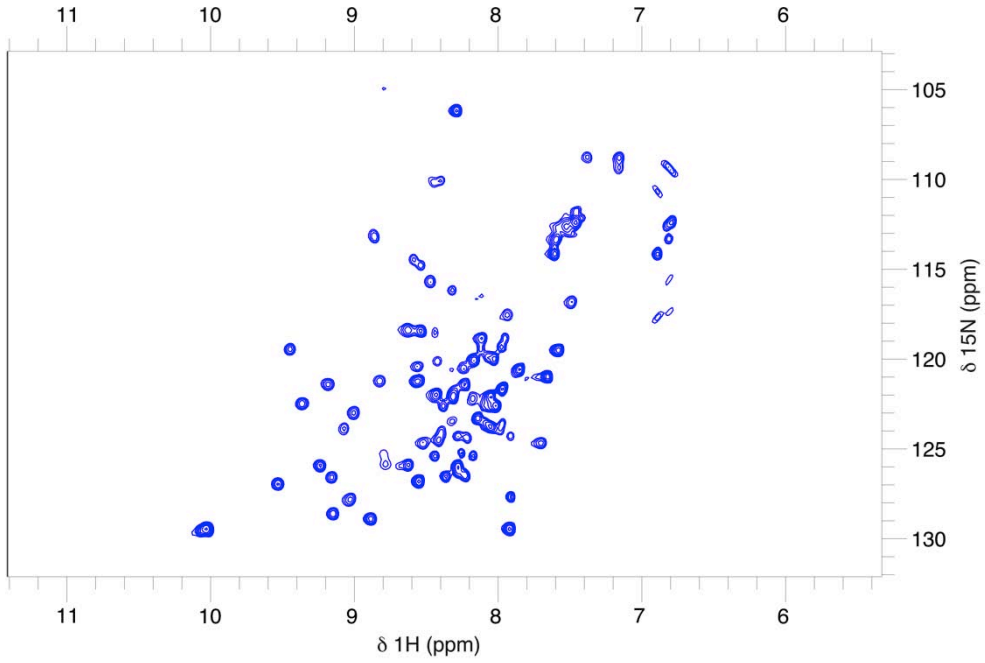


Figure 54: HIS-CTD-220-270 in 150mM NaCl at pH 7 .^{15}N HSQC of HIS-IN-CTD 220-270 in 25mM HEPES pH 7, 2mM βME, 150mM NaCl

A superposition of the spectra at 150mM NaCl and 1M NaCl reveals some differences in protein conformation or dynamics between HIS-IN-CTD 220-270 in both conditions. Generally, there is movement of all resonance peaks to the right. There are also more overlapping peaks between 7.9ppm and 8.5ppm, suggesting that the protein is less well folded in 150mM NaCl There is also a possibility that the protein is less soluble, and therefore more oligomerized in lower salt. Additionally, there is a change in intensity in some peaks, with some intensities getting stronger, while some become weaker, indicating that these residues are either more buried, or more exposed depending on salt concentration.

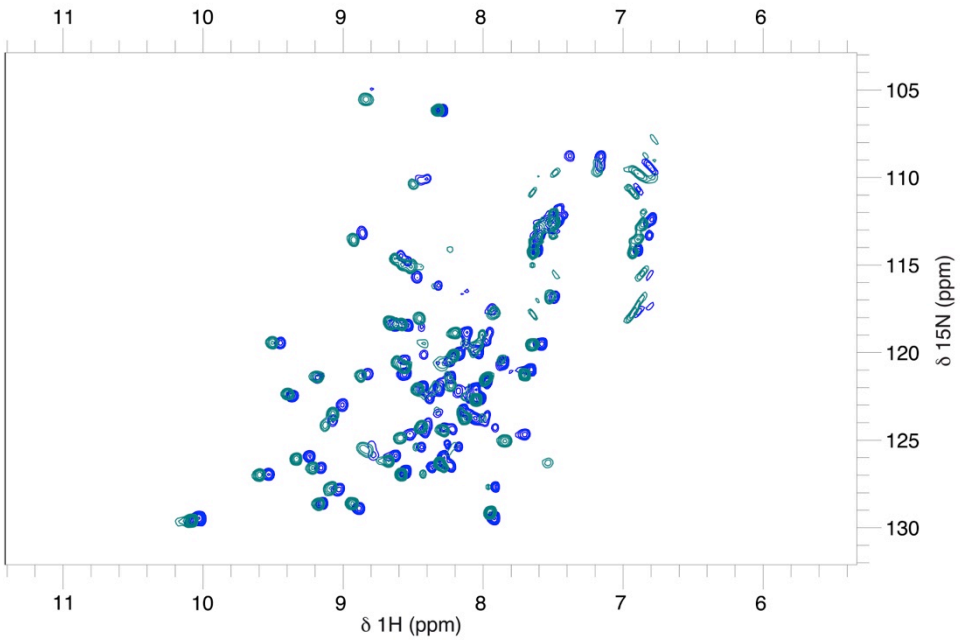


Figure 55: Comparison of HIS-CTD-220-270 in different salt concentration. Superposition of ^{15}N HSQC spectra of HIS-IN-CTD 220-270 1M NaCl (teal) and 150mM NaCl (blue)

Peptide Interactions at 150mM NaCl and pH 7

Titration experiments with H4K20Me1 peptide were performed on 170μM HIS-IN-CTD 220-270 in 25mM HEPES pH 7, 2mM βME, 150mM NaCl in a 3mm tube with 10% D2O on a 700 MHz spectrometer. Experiments were recorded at 298K with 256 scans and a receiver gain of 912. ^{15}N HSQC spectrum was recorded on protein only (Figure 52), and with increasing concentrations of peptide (1mM and 2mM).

In the presence of 1mM peptide, two effects are observed: chemical shift changes on some peaks, and the appearance of new peaks. These changes are not large, suggesting that these residues are in fast exchange.

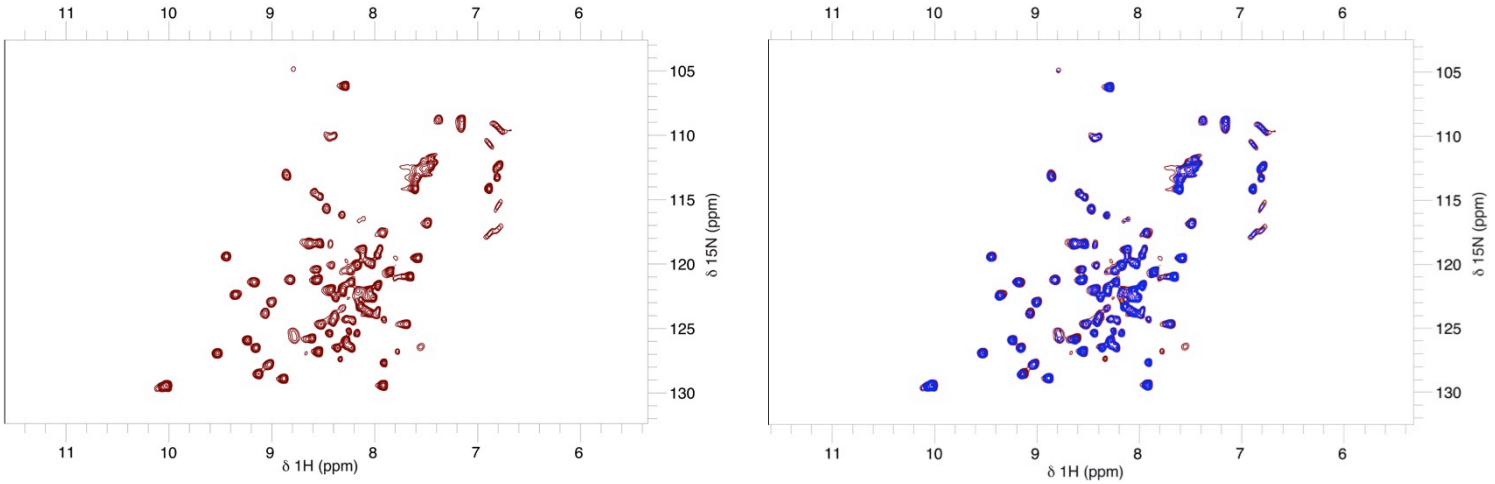


Figure 56: Interaction with 1mM peptide at pH 7 in 150mM NaCl Left: ^{15}N HSQC of HIS-IN-CTD 220-270 in 25mM HEPES pH 7, 2mM βME, 150mM NaCl plus 1mM peptide. Right: Superposition of ^{15}N HSQC spectrum without peptide (blue) and with 1mM peptide (maroon)

When the peptide concentration is increased to 2mM, further chemical shift changes were observed Additionally, there was an increase in the intensity of peaks observed, also suggesting the solvent exposure of previously buried residues. Furthermore, satellite peaks were seen emerging from 2 resonance peaks. Interestingly, the peaks show up around the same chemical shifts as those at 1M NaCl (Figure 51), suggesting identical dynamics of interactions in both salt concentrations at pH 7

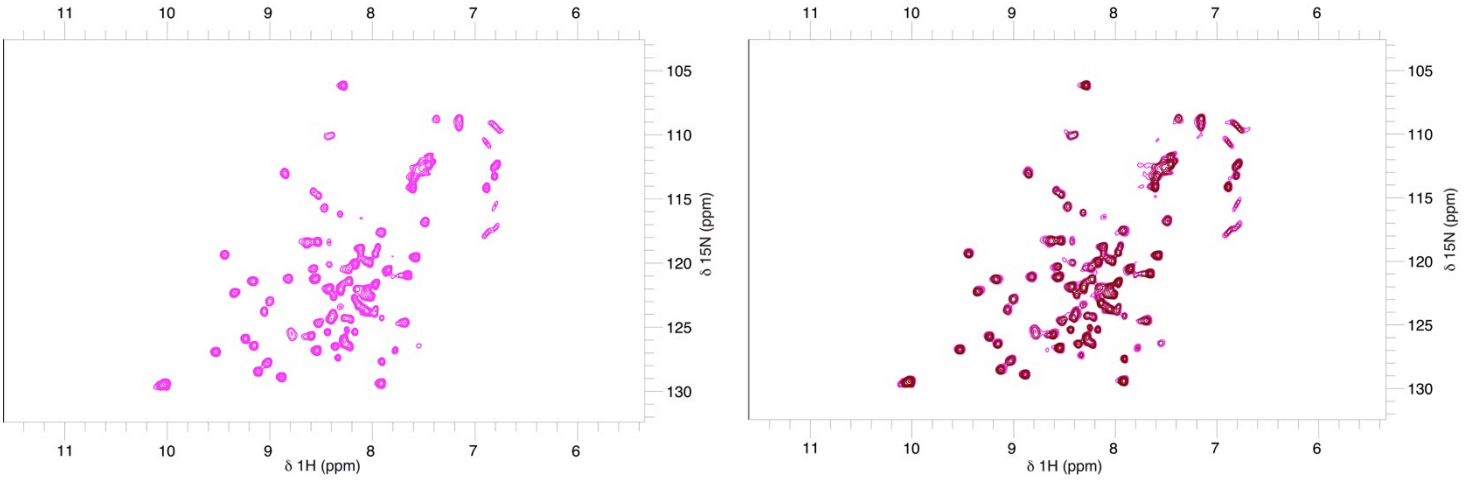


Figure 57: Interaction with 2mM peptide at pH 7 in 150mM NaCl . Left: ^{15}N HSQC of HIS-IN-CTD 220-270 in 25mM HEPES pH 7, 2mM βME, 150mM NaCl plus 2mM peptide. Right: Superposition of ^{15}N HSQC spectrum with 2mM peptide (magenta) and with 1mM peptide (maroon)

Dynamics and Interactions of HIS-IN-CTD 220-270 at pH 8

^{15}N HSQC of HIS-IN-CTD 220-270 at pH 8

In order to be more consistent with conditions used in MST and NOESY experiments, the experiments were repeated at pH 8. ^{15}N HSQC spectrum was recorded on 140μM of HIS-IN-CTD 220-270 in 25mM HEPES pH 8, 2mM βME, 150mM NaCl plus 10% D2O in a 3mm tube. Data was collected on a 700 MHz spectrometer were collected at 293K with 256 scans and a receiver gain of 912.

At pH 8, the ^{15}N HSQC spectrum appears different from the spectra at pH 7. At pH 8, there are significantly fewer peaks (56 peaks) that correspond more accurately to the number of residues in the construct. Additionally, several resonance peaks visible in the spectrum at pH 7 and are no longer visible at pH 8. This data suggests a pH dependent change in dynamics, conformation or population state.

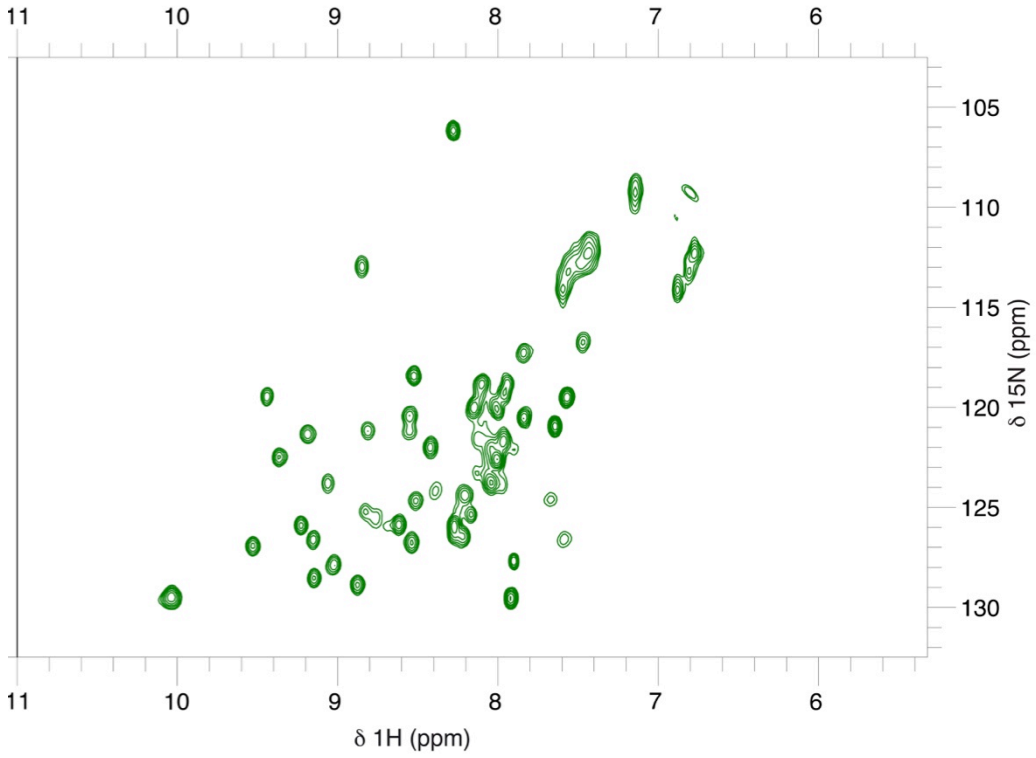


Figure 58: HIS-IN-CTD 220-270 at pH 8 . ^{15}N HSQC of HIS-IN-CTD 220-270 in 25mM HEPES pH 8, 2mM βME, 150mM NaCl

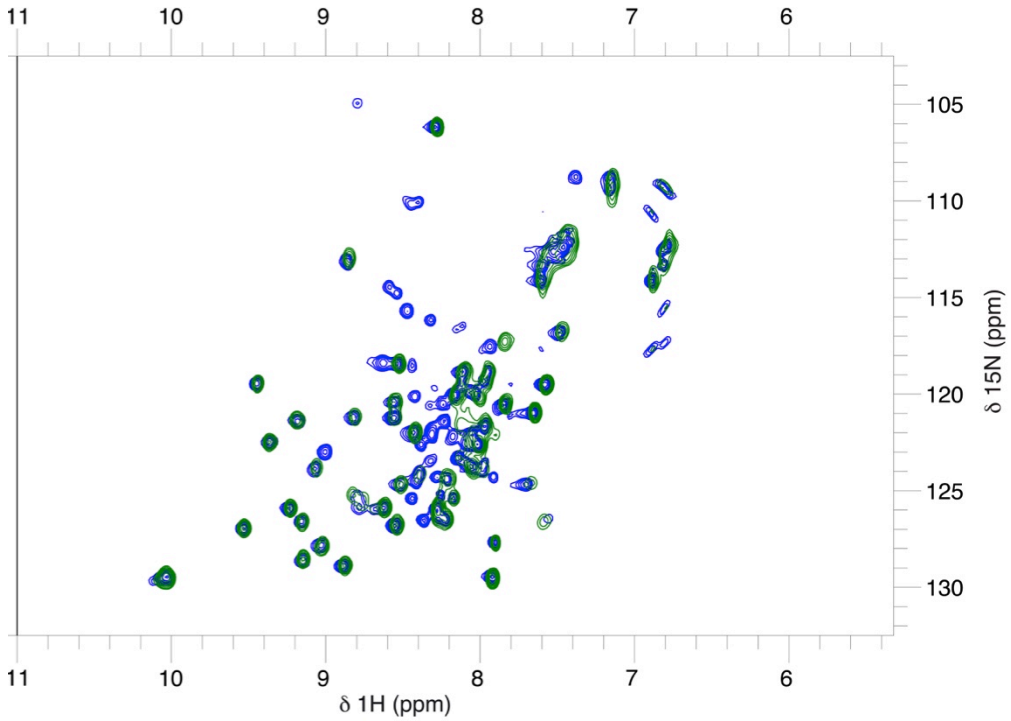


Figure 59: Comparison of spectra at pH 7 and pH 8. Superposition of ^{15}N HSQC spectra of HIS-IN-CTD 220-270 at pH 7 (blue) and pH 8 (green). Several resonance peaks at pH 7 are no longer visible at pH 8.

Peptide Interactions at pH 8

Titration experiments with H4K20Me1 peptide were performed on 140μM HIS-IN-CTD 220-270 in 25mM HEPES pH 8, 2mM βME, 150mM NaCl in a 3mm tube with 10% D2O on a 700 MHz spectrometer. Experiments were recorded at 298K with 256 scans and a receiver gain of 912. ^{15}N HSQC spectrum was recorded on protein only (Figure 56), and with increasing concentrations of peptide (1mM and 2mM).

Upon addition of 1mM peptide, chemical shift changes were observed. Consistent with pH 7, the changes observed were representative of interactions with fast exchange Additionally, 2 new resonance peaks were observed. One of these peaks appears in the same position upon addition of 1mM peptide at pH 7.

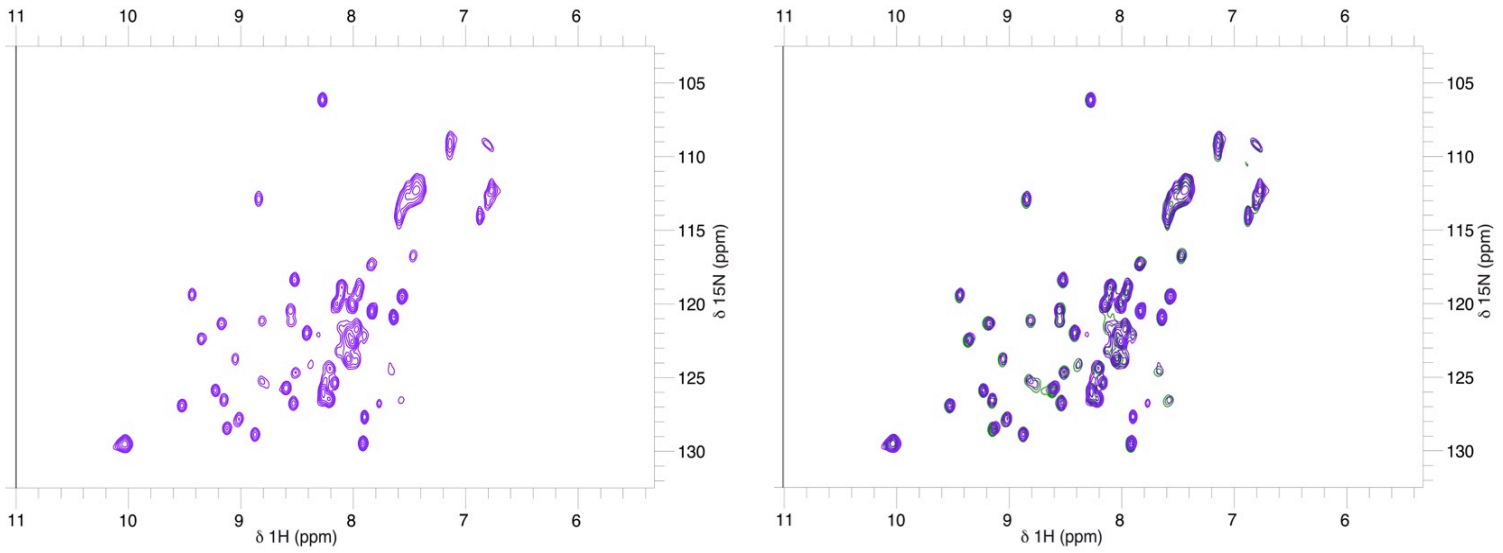


Figure 60: Interaction with 1mM peptide at pH 8 in 150mM NaCl Left: ^{15}N HSQC of HIS-IN-CTD 220-270 in 25mM HEPES pH 8, 2mM βME, 150mM NaCl plus 1mM peptide. Right: Superposition of ^{15}N HSQC spectrum without peptide (green) and with 1mM peptide (mauve)

Increasing the peptide concentration to 2mM results in several changes in the ^{15}N HSQC spectrum. Generally, there is a major rearrangement of peaks between 7.9ppm and 8.5ppm. More specifically, four major effects are observed: resonance peaks shifting, changes in peak intensity, and deletion of resonance peaks and appearance of new resonance peaks. Further chemical shifts are observed on peaks that appear to be in fast exchange. The disappearing peaks could be due to intermediate exchange, or peaking broadening due to oligomerization or protein unfolding. Additionally, changes in peak intensity could be due to residues being more buried, or exposed in the presence of peptide. It is also possible that new peaks correspond to residues that have become more stable in the presence of peptide, or residues that were previously involved in protein interactions and have become free upon peptide addition.

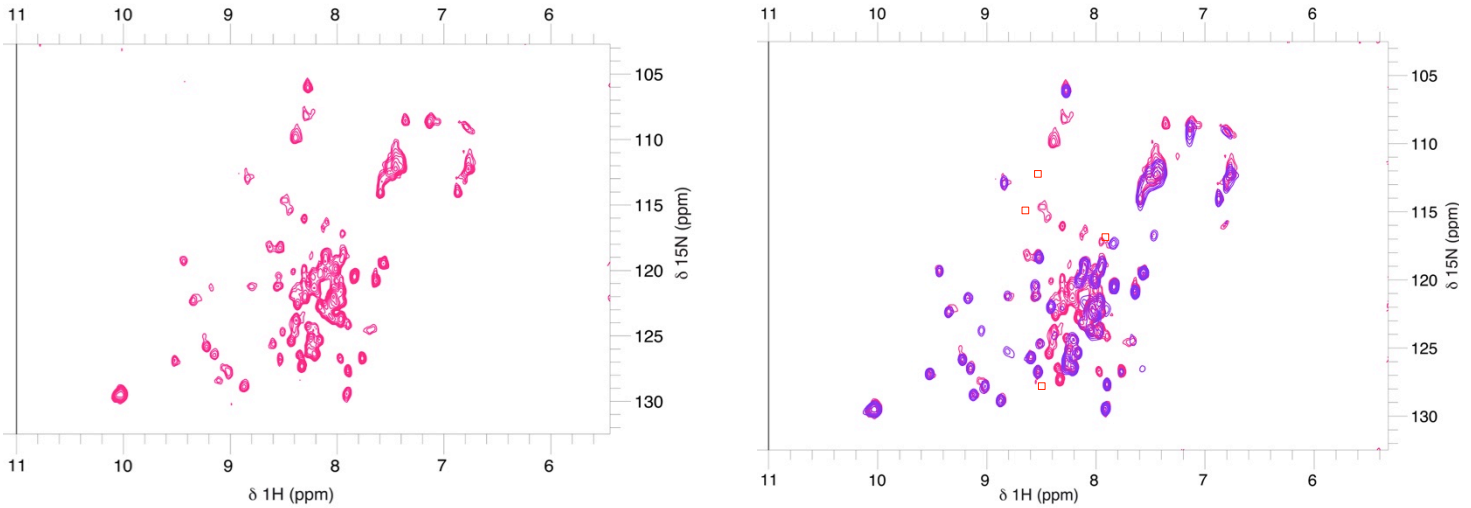


Figure 61: Interaction with 2mM peptide at pH 8 in 150mM NaCl Left: ^{15}N HSQC of HIS-IN-CTD 220-270 in 25mM HEPES pH 8, 2mM βME, 150mM NaCl plus 2mM peptide. Right: Superposition of ^{15}N HSQC spectrum with 2mM peptide (pink) and with 1mM peptide (purple) with red squares highlighting some new resonance peaks

Interestingly, some of the new resonance peaks observed in the ^{15}N HSQC of HIS-IN-CTD 220-270 observed in the presence of 2mM peptide at pH 8 were present in the spectrum of protein only at pH 7 (highlighted in red boxes below), indicating changes in conformation or dynamics in the presence of 2mM peptide that are similar to what is observed at pH 7. This suggests that the interaction with 2mM peptide at pH 8 exposes residues that were not visible (buried or involved in other interactions) without peptide at pH 8. These residues are visible at pH 7.

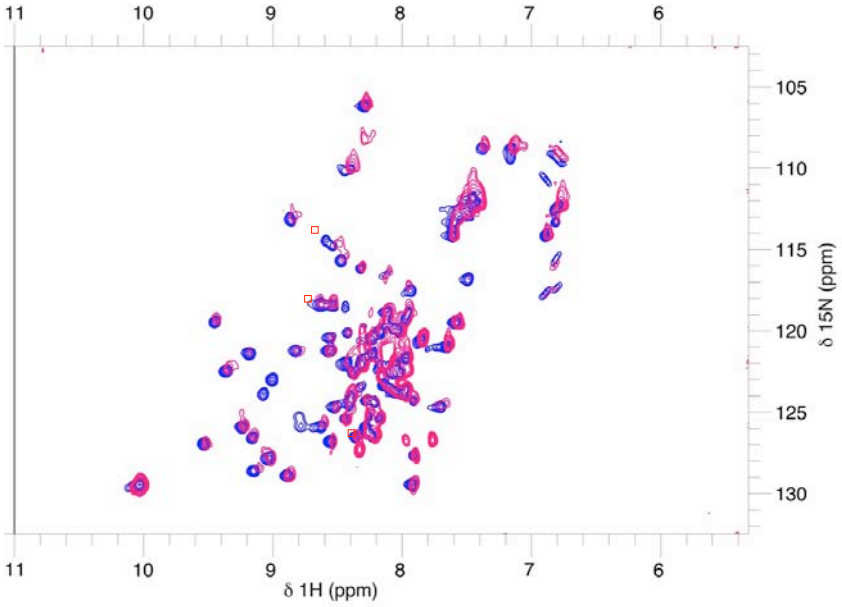


Figure 62: Comparison of complex at pH 8 and protein only at pH 7.. Superposition of spectra of ^{15}N HSQC of HIS-IN-CTD 220-270 at pH 8 plus 2mM peptide (pink) and at pH 7 without peptide (blue).

^{15}N HSQC experiments were performed with protein and water only, in order to confirm that changes observed in the spectra were due to peptide binding, and not a change in conformation due to protein dilution. An experiment with 10µl H_2O (equivalent of the volume added for 2mM peptide) was performed on 140µM HIS-IN-CTD 220-270 in 25mM HEPES pH 8, 2mM βME, 150mM NaCl in a 3mm tube with 10% D2O on a 700 MHz spectrometer. Experiments were recorded at 298K with 256 scans and a receiver gain of 912.

No significant chemical changes were observed in the presence of water only, suggesting that changes observed in the presence of peptide are due to peptide binding, and not changes in protein conformation from dilution.

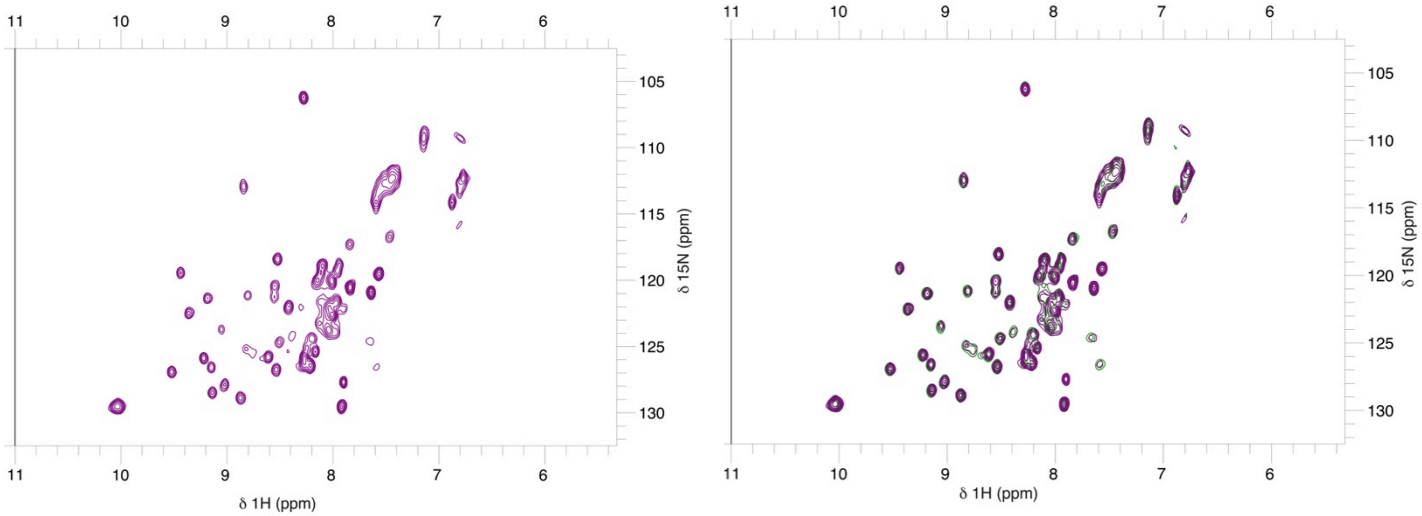


Figure 63: Control experiments with water. . Left: ^{15}N HSQC of HIS-IN-CTD 220-270 in 25mM HEPES pH 8, 2mM βME, 150mM NaCl plus 10µl water. Right: Superposition of protein plus water (purple) and protein only (green)

NMR Assignment

In order to identify the residues corresponding to the resonance peaks involved in the interaction with peptide, 3D experiments were carried out with the aim of assigning residues. In order to improve the quality of the spectra, 2M d5-glycine was added to the buffer as previously done (Eijkelenboom, Puras Lutzke et al. 1995). The addition of 2M d5-glycine improved the spectrum quality, probably by stabilizing the protein

2D spectra with Glycine

A ^{15}N HSQC experiment was recorded to assess the protein quality at pH 7, to observe if the addition of glycine would improve the quality of the spectra. The experiment was recorded on on 140µM HIS-IN-CTD 220-270 in 25mM HEPES pH 7, 2mM βME, 150mM NaCl, 2M glycine in a 3mm tube with 10% D2O on a 700 MHz spectrometer.

Although a loss of intensity was observed in some resonance peaks, the quality of the spectra was generally improved with the addition of glycine as there were less overlapping peaks between 7.9ppm and 8.5ppm.

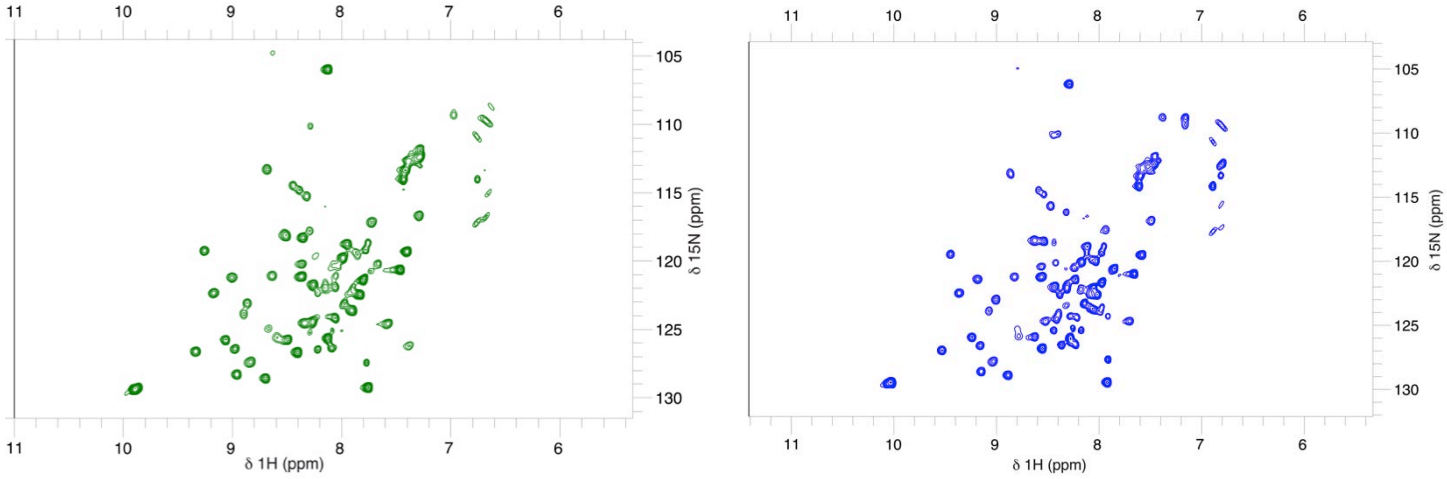


Figure 64: Effects of glycine on spectra quality. Side by side comparison of ^{15}N HSQC of HIS-IN-CTD 220-270 in buffer at pH 7 plus 2M glycine (left, green) and buffer minus glycine (right, blue)

Protein Assignment

The following experiments were performed in order to carry out sequential backbone assignment: ^{15}N HSQC, ^{13}C HSQC, HNCA, CBCACONH, HNCACB and HNCO. ^{15}N NOESY was used to obtain long-range distance information for structure calculation. All experiments were recorded on ^{15}N ^{13}C labeled HIS-IN-CTD 220-270 in 25mM HEPES pH 8, 2mM βME, 150mM NaCl, and 2M d-5 glycine. All data analysis was performed using CcpNmr Analysis software (Vranken, Boucher et al. 2005).

38 residues out of 51 residues in the IN-CTD 220-270 could be assigned from the ^{15}N HSQC. 3 of out the 13 unidentifiable residues are prolines (P233, P238, P261), which lack NH protons and therefore do not show up on ^{15}N HSQC spectra. Other missing residues include

Q221, N222, S230, R231, D232, L242, E246, G247, N254 and S255. Resonance peak 38 corresponds to the side-chain Nε-Hε groups of W235/W243. Data obtained from CBCACONH/HNCA experiments allowed the Cα and Cβ assignments for some peaks that could not be linked to the ^{15}N HSQC spectrum.

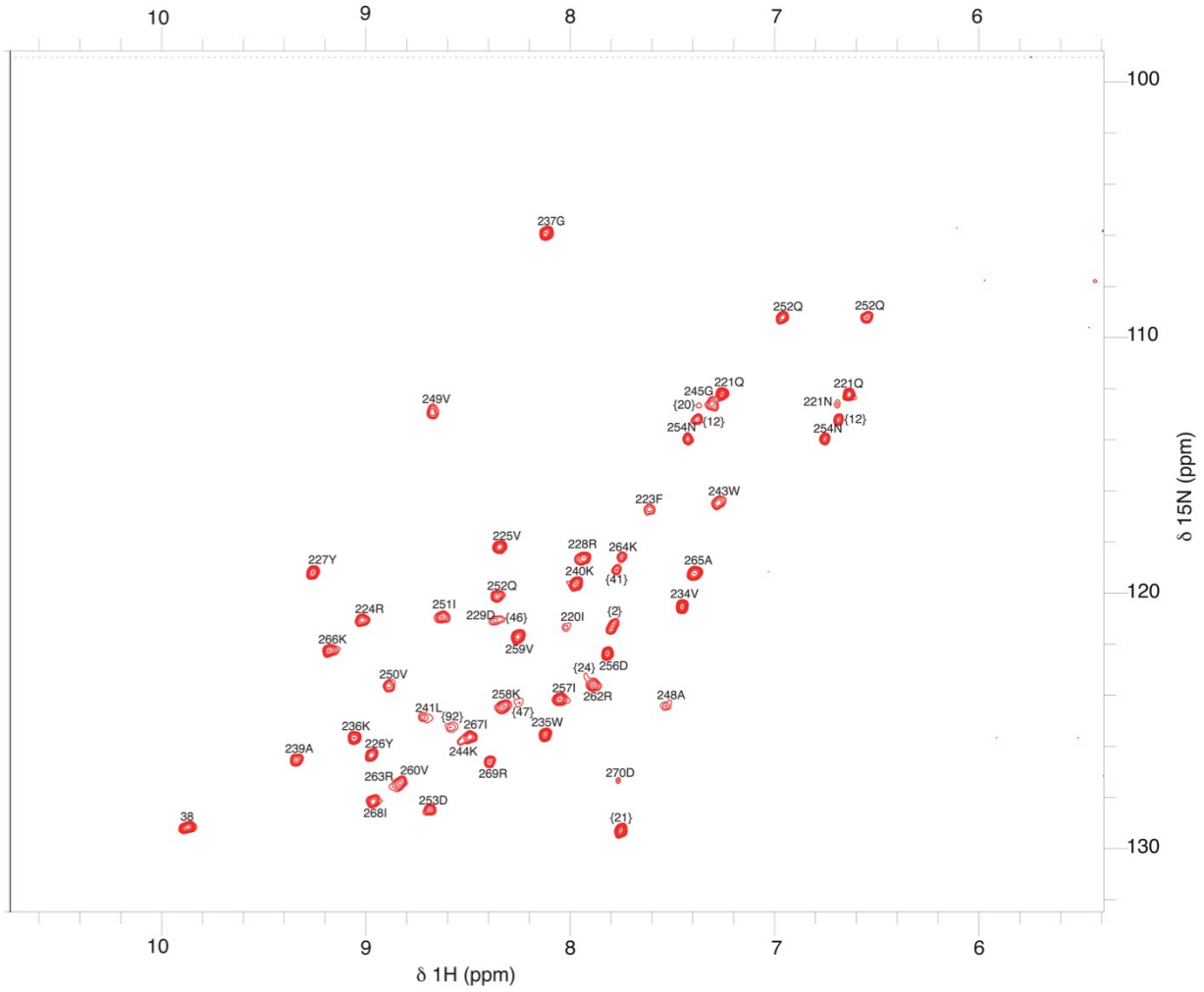


Figure 65 : Protein Assignment**.** Assignment of residues in HIS-IN-CTD 220-270 in 25mM HEPES, 2mM βME, 150mM NaCl, 2M d-5 glycine

Additionally, there were 7 resonance peaks that could not be sequentially linked to other residues. Surprisingly, some of these peaks display a strong signal on the HSQC and have large peak volumes. However, it was difficult to link them to neighboring residues using 3D experiments.

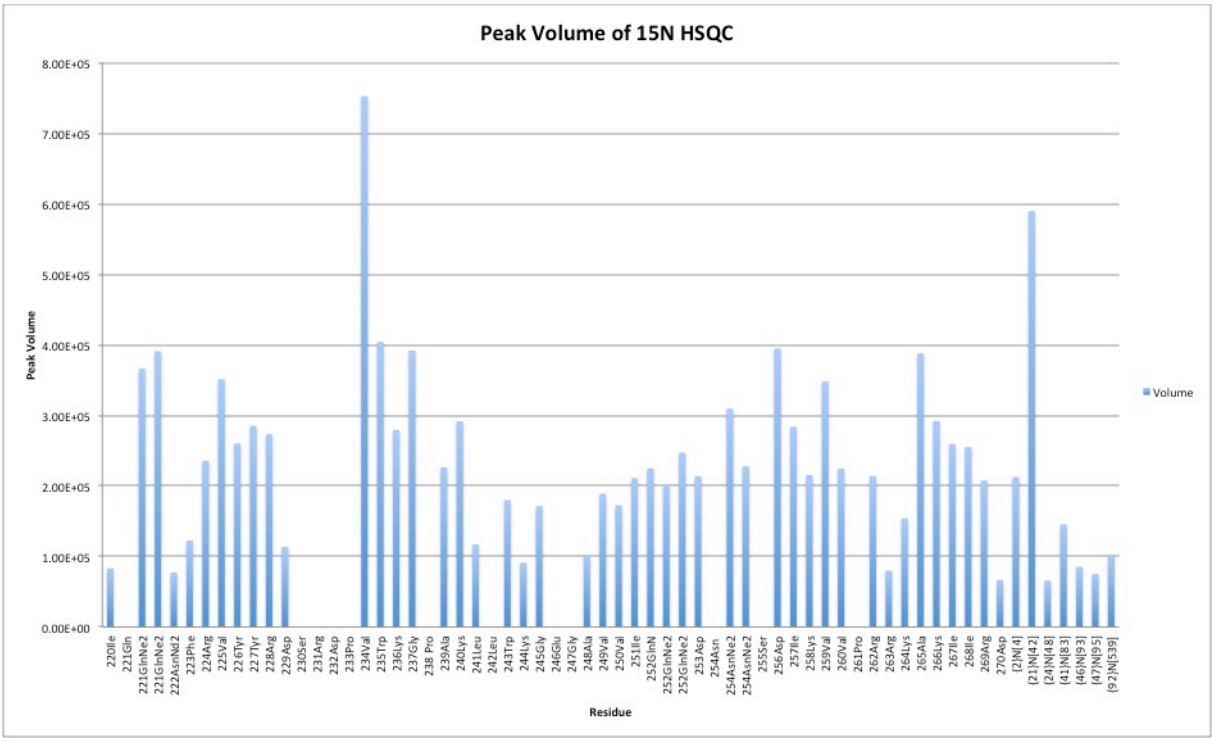


Figure 66: Peak Volumes of assigned residues. Residues without peaks are either prolines or other unidentified residues.

Secondary Structure Prediction

The embedded DANGLE (Dihedral ANgles from Global Likelihood Estimates) (Cheung, Maguire et al. 2010) program on CcpNmr Analysis Software was used to predict phi/psi angles, and secondary structure features using the amino-acid sequence and experimental chemical shift information for each residue. Overall, predicted phi/psi angles and secondary structure features were consistent with published results (Eijkelenboom, Puras Lutzke et al. 1995, Cherepanov, Sun et al. 2005), with 5 β-strands predicted. β sheets were predicted from residues F223-R228, A248-D253, K236-L241, D256-V261 and K264-I268, with a coil and α turn from P261-R263. We hypothesize that missing chemical shift data in unidentified regions correspond to residues in loops or coils, or residues involved in dimer formation.

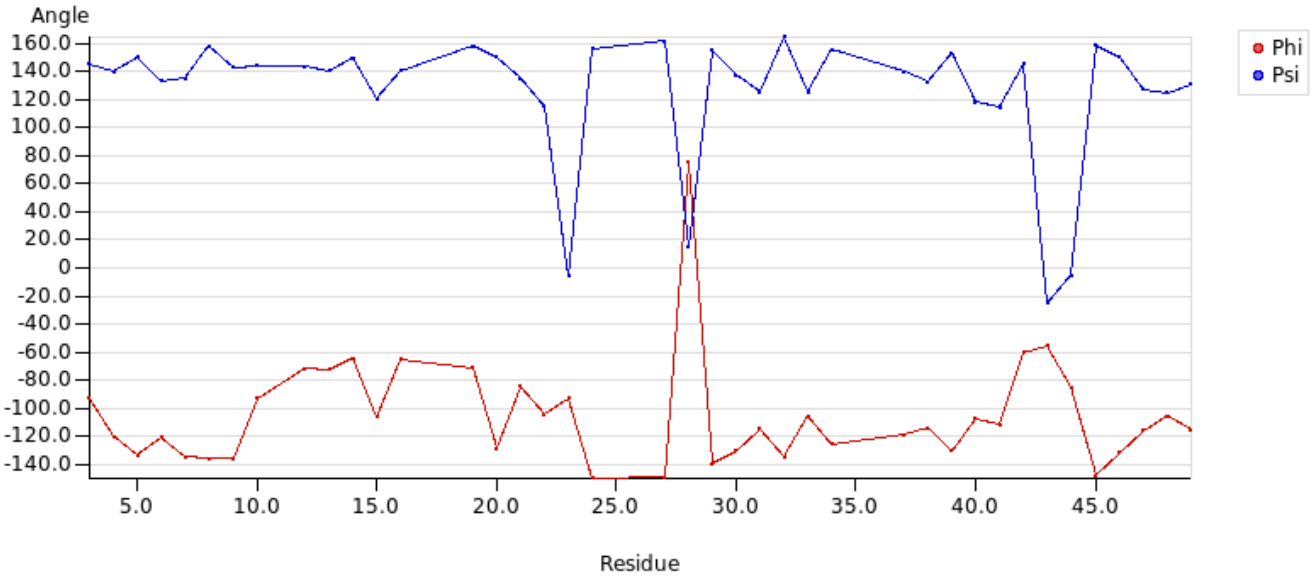


Figure 67: Dihedral Angle Prediction**.** Predicted phi (red) and psi (blue) angles, from residue 1 (I220) to 51 (D270)

Figure 68: Secondary structure chart. showing predicted regions of β-sheets in grey arrows, with the α helical turn from residues 42 to 44.

Chemical Shift Mapping Analysis

Assignment results were transposed to ^{15}N HSQC titration experiments. Most residues below 7.9ppm and 8.2ppm can be identified. However, without glycine, overlapping peaks between 7.9ppm and 8.2ppm were difficult to assign. Although 3D experiments were recorded at pH 8, many of the residues could be traced in the spectra at pH 7.

Analysis at pH 8

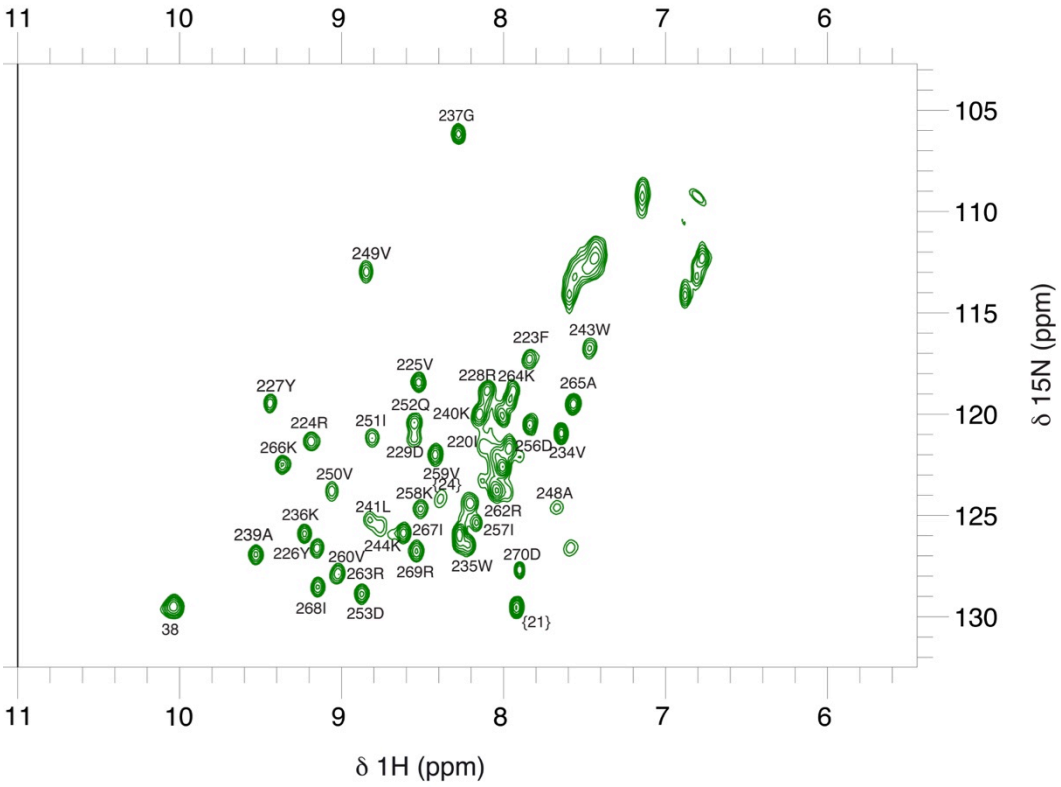


Figure 69: Assignment results at pH 8. . ^{15}N HSQC spectrum showing assignment results at pH 8

At pH 8, chemical shift changes can be easily observed upon the addition of 1mM peptide on I268, I267/K244, and K266, indicating that these residues are impacted by the presence of peptide and are involved in peptide interactions. New resonance peaks could not be assigned.

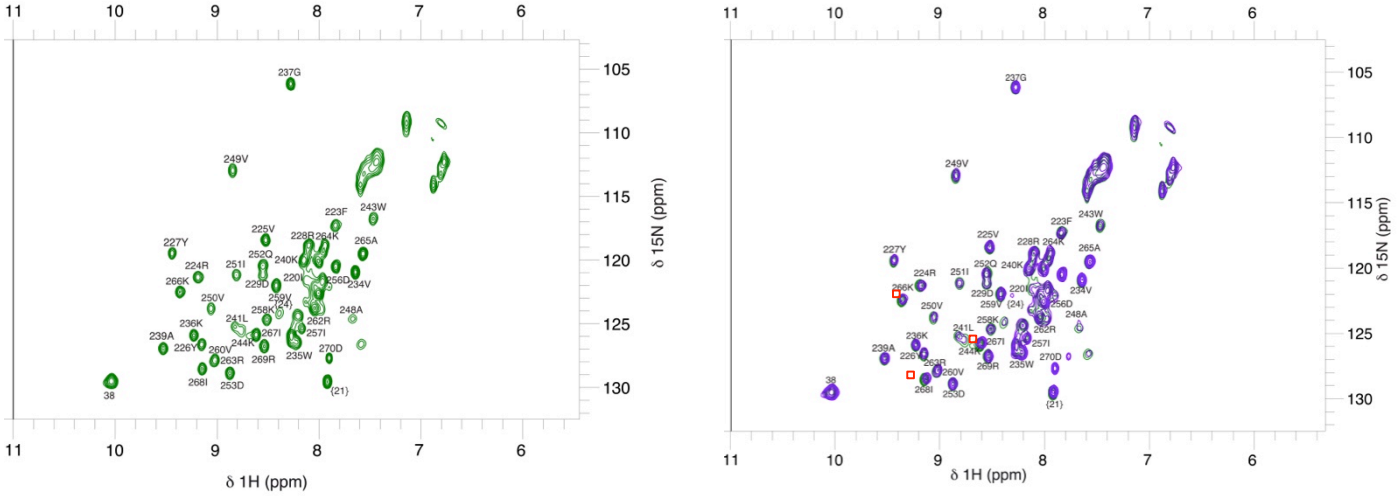


Figure 70: Residues affected by 1mM peptide at pH 8. Left: ^{15}N HSQC spectrum showing assignment of HIS-IN-CTD 220-270 results plus 1mM peptide at pH 8, Right: Superposition of spectra plus peptide (purple) and minus peptide (green). Affected residues are highlighted in red squares

At 2mM peptide, residues that are no longer visible on the ^{15}N HSQC spectrum are F223, W243, L241 and V250. When compared to the spectrum at pH 7, it appears that the resonance peak of F223 moved faded from its original position to where it appears in the spectrum at pH 7. In addition to I268, I267/K244, and K266, further chemical shift changes were observed G237 and V234. There was a secondary peak form on V225 and decrease in intensity of R224 and I268.

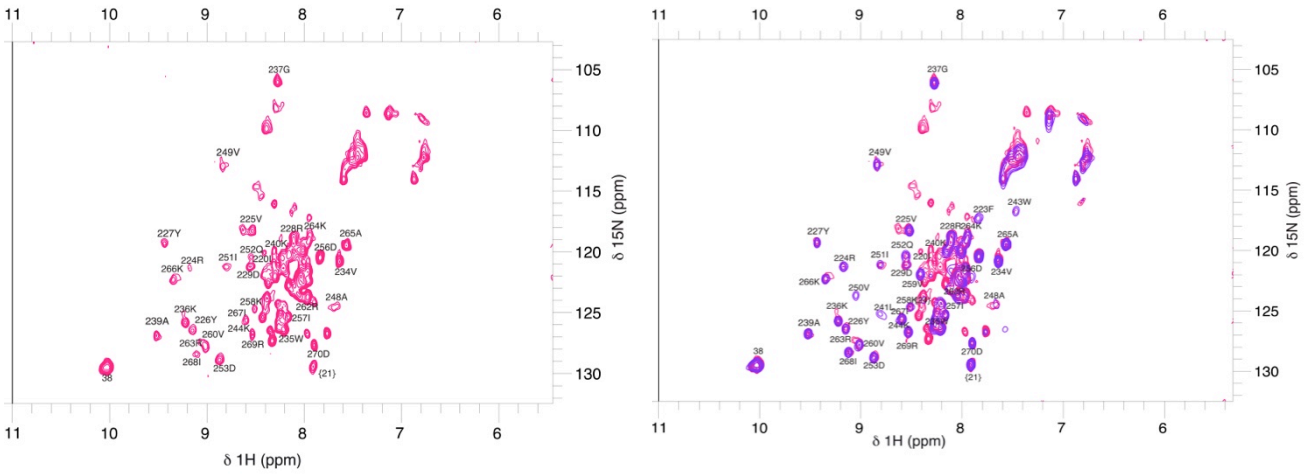


Figure 71: Residues affected by 2mM peptide at pH 8. Left: ^{15}N HSQC spectrum showing assignment of HIS-IN-CTD 220-270 results plus 2mM peptide at pH 8, Right: Superposition of spectra plus peptide (pink) and 1mM peptide (purple).

Analysis at pH 7

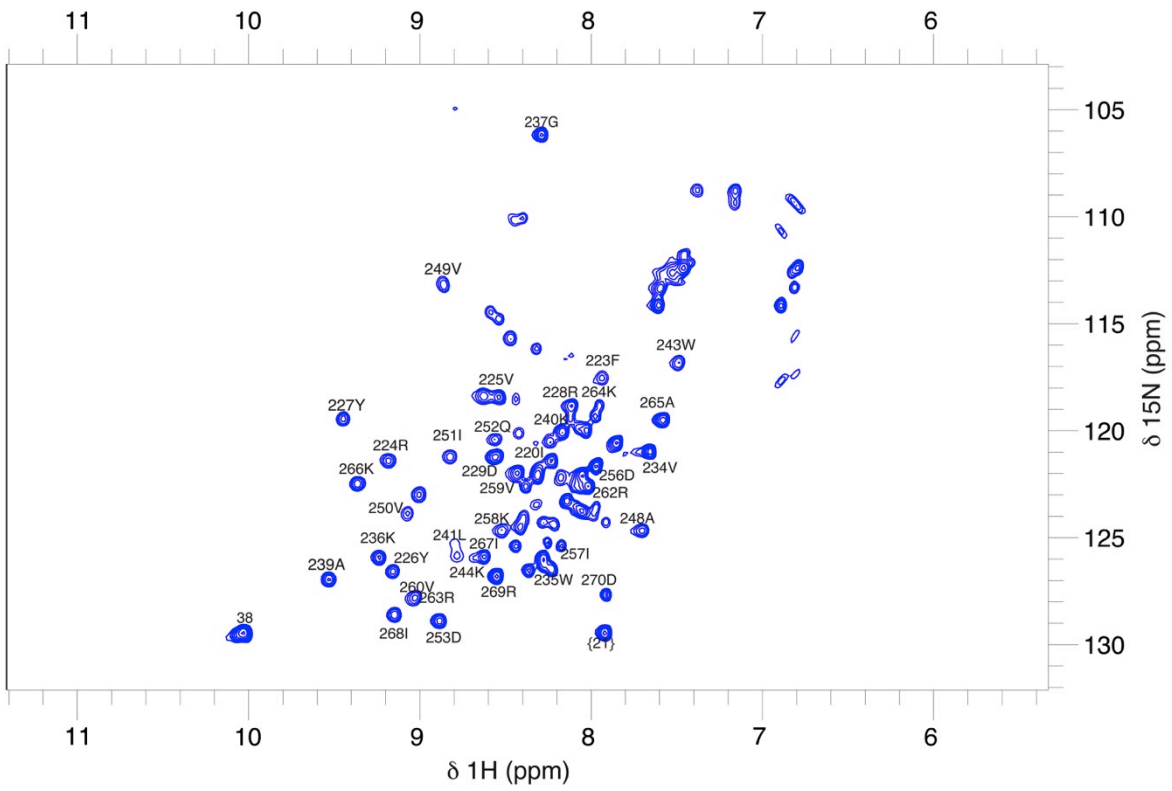


Figure 72: Assignment results at pH 7**.** ^{15}N HSQC spectrum showing assignment results at pH 7

In addition to new resonance peaks, chemical shift changes were observed on I268, I267/K244, K266 and F223, in the presence of 1mM peptide.

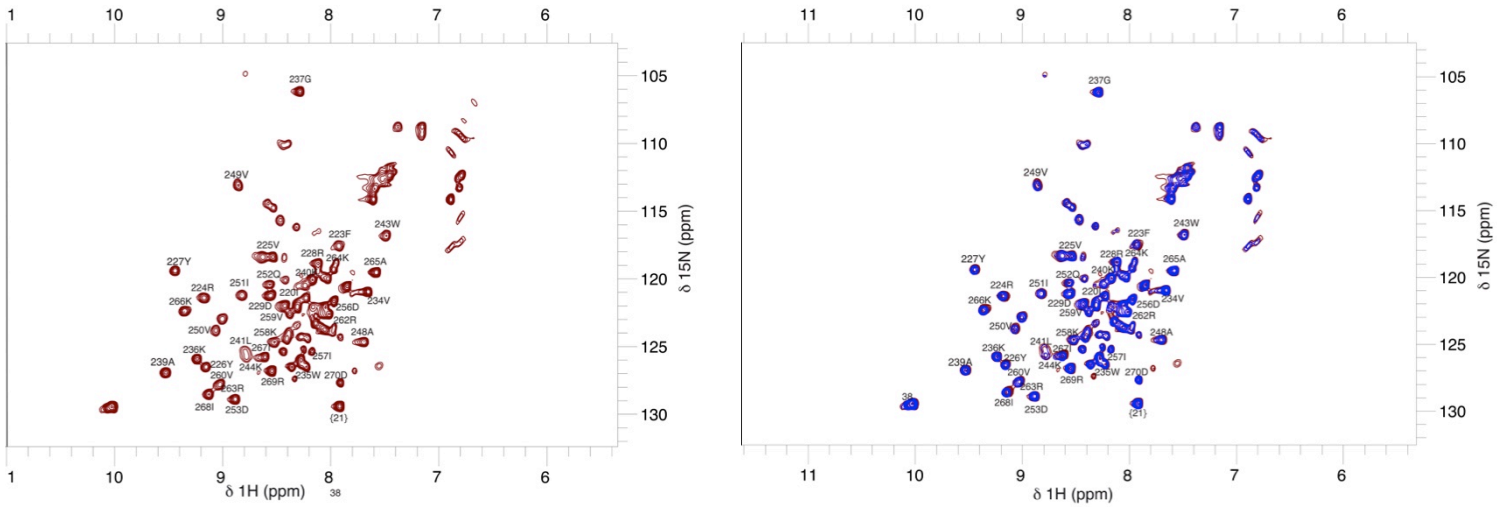


Figure 73: Residues affected by 1mM peptide at pH 7. Left: ^{15}N HSQC spectrum showing assignment of HIS-IN-CTD 220-270 results plus 1mM peptide at pH 7, Right: Superposition of spectra plus 1mM peptide (maroon) and minus peptide (blue)

When peptide concentration is increased to 2mM, further chemical shifts are observed on I268, I267/K244, K266, F223 and V250. These residues are also perturbed at pH 8, suggesting that they are directly affected by the addition of peptide.

Quantification of Chemical Shift Changes

Chemical shift changes were quantified using the formula below, with a scaling factor of 0.2 to balance the contributions of 1H and N15 to the perturbations observed (Ziarek, Peterson et al. 2011). They were ranked in different categories- high, medium, and low, with high being a chemical shift difference of 0.04ppm or higher, medium being between 0.03ppm and 0.04ppm, and low being between 0.02 and 0.03ppm.

$$\sqrt{(\Delta\delta H)^2 + 0.2(\Delta\delta N)^2}$$

- $(\Delta\delta H)^2$ - chemical shift difference in H dimension square
- $(\Delta\delta N)^2$ - chemical shift difference in N dimension squared
- 0.2 - scaling factor

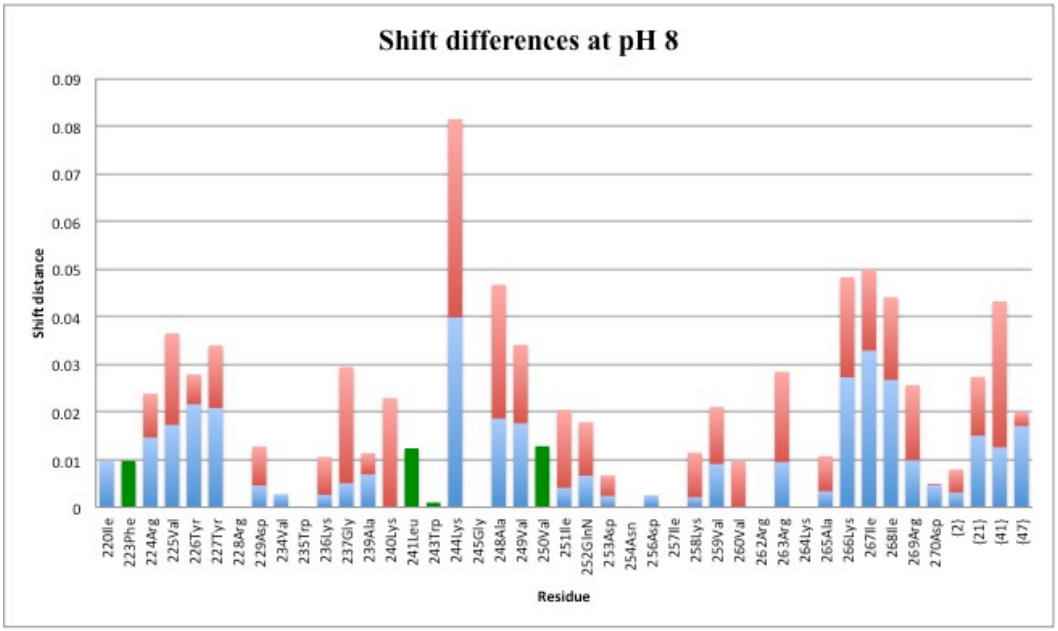


Figure 74: Graphical representation of chemical shift differences at pH 8 upon the addition of 1mM peptide (blue) and 2mM peptide (red). Highlighted in green are residues that are not visible at 2mM peptide. The chemical shifts of residues without bars are not affected by binding

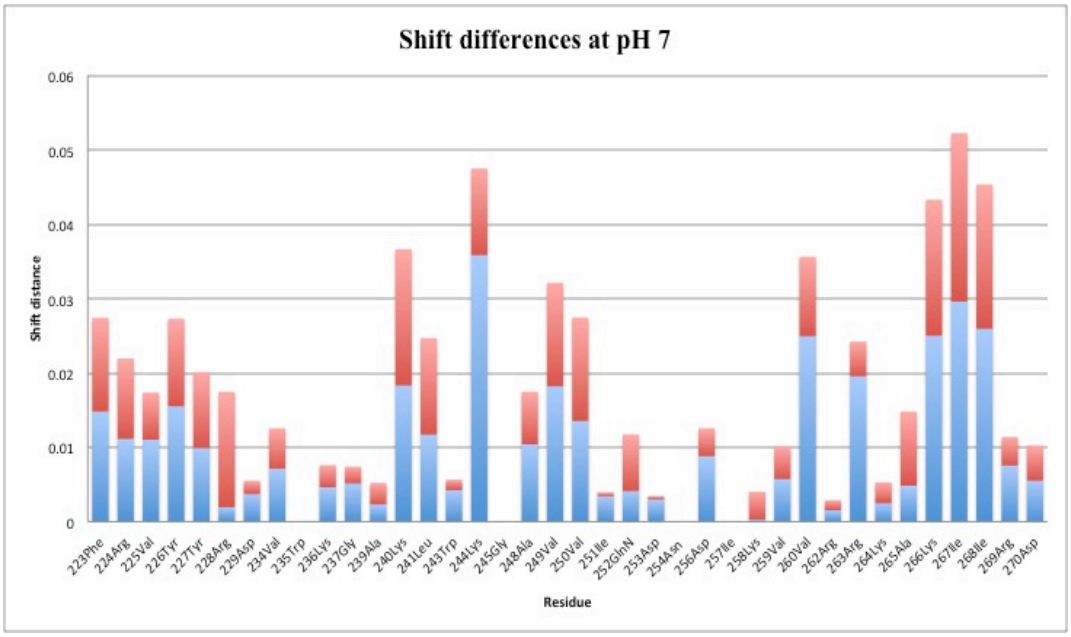


Figure 75: Graphical representation of chemical shift differences at pH 7 upon the addition of 1mM peptide (blue) and 2mM peptide (red). The chemical shifts of residues without bars are not affected by binding

Effect	**pH 7**	**pH 8**
Disappearing peaks		F223, W243, L241, V250
High shift difference	I268, I267, K266, K244	I268, I267, K266, K244, A248
Medium shift difference	K240, V249, V260	V225, Y227, V249
Low shift difference	F223, R224, Y226, L241, R263, Y227, V250	R224, Y226, G237, K240, V259, R263, R269

Table 5: Summary of all perturbed peaks in the presence of 2mM H4K20me1 peptide at pH 7 and pH 8

Overall, the residues being affected at pH 7 and pH 8 are consistent, indicating that these residues are indeed perturbed by peptide binding, and are involved in the interaction with the peptide. Of particular interest are K244, K266, I267, I268, which show the highest chemical shift difference in both conditions. The extent to which other residues are perturbed is dependent on the pH. For example, peaks corresponding to F223, L241, V250 disappear at pH 8, but are only slightly perturbed at pH 7 upon peptide addition.

Results from Circular Dichroism experiments

Due to the changes observed in ^{15}N HSQC between pH 7 and 8 upon peptide addition, we hypothesized that there might be differences in secondary structure content at each pH. CD experiments were performed in order to determine if there were changes in overall secondary structure content at each pH in the presence of 2mM peptide. For each pH, the reference data set (buffer) was subtracted before plotting the graphs. All spectra are characteristic of b-sheets, with negative bands at 218nm and positive bands at 195nm. The spectrum corresponding to protein only at pH 8 (green,

Figure 76) is shifted from the other spectra. However, the spectrum corresponding to pH + 2mM peptide (teal,

Figure 76) is superimposable with the spectra at pH 7. This data is consistent with NMR results, where the 15N HSQC spectrum at pH 8 plus peptide resembles the spectra at pH 7. This data confirms that the addition of peptide at pH 8 changes the conformation or dynamics of the protein, making it more similar to protein at pH 7.

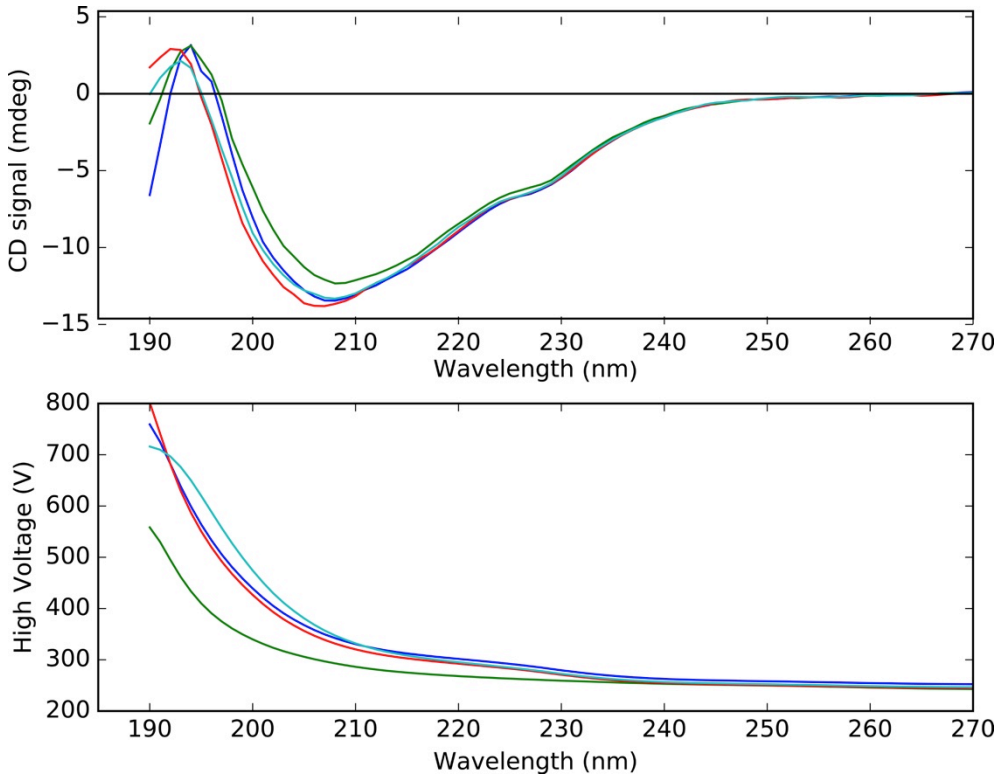


Figure 76: Superposition of CD spectra. HIS-IN-CTD 220-270 at pH 7 (blue), pH 7 plus 2mM peptide (red), pH 8 (green) and pH 8 plus 2mM peptide (teal

NMR structure

Data obtained from 2D and 3D experiments were used to generate distance restraints, which were used to calculate the solution structures of IN-CTD 220-270 using X-PLOR (Schwieters, Kuszewski et al. 2003). A superimposition of the 5 best structures shows consistent conformations within the β strands with an overall RMSD of 1.08Å. However, there was a lot of variation within the loops, and at the N- and C-terminal ends. This result was not surprising due to the quality of the ^{15}N HSQC. Moreover, these structures were generated from a limited set of constraints obtained from Cα and Cβ only, as all attempts to record 3D experiments for side chain were unsuccessful. New experiments are planned in order to generate more constraints. As expected, residues in the loops correspond to residues that were difficult on the ^{15}N HSQC.

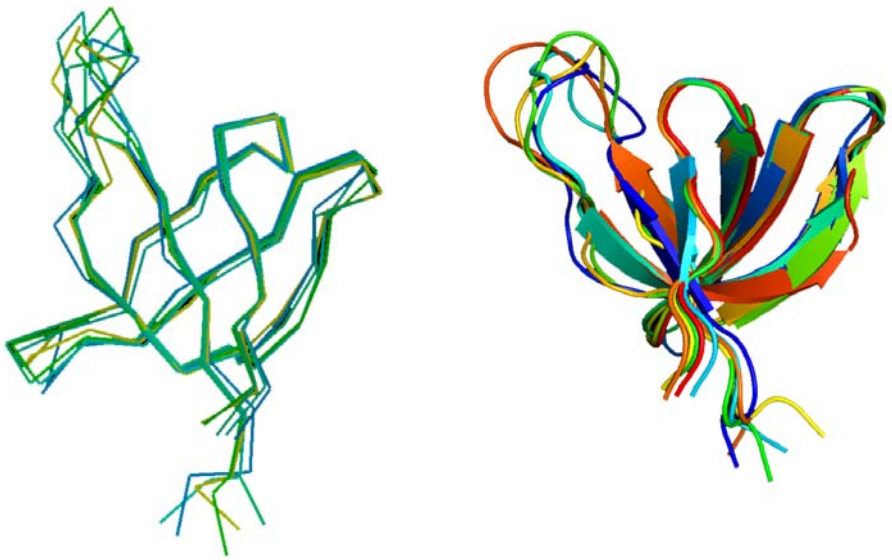
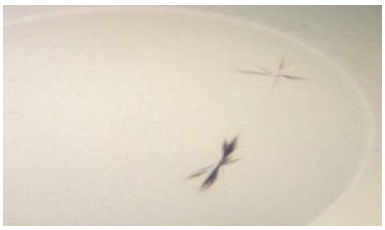


Figure 77: NMR structures of HIS-IN-CTD 220-270. Superposition of structures show high flexibility within the loops and at the N- and C-terminal domains

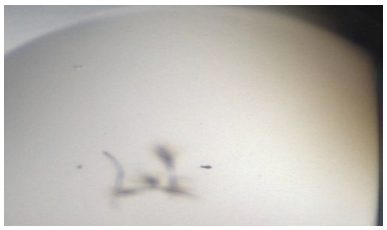
Results from Crystallization experiments

GST-IN-CTD crystallization

Attempts to obtain large 3D crystals with the GST-IN-CTD construct were unsuccessful. Initial hits were thin needles were obtained in 0.1M BICINE pH 9,10% MPD and 30% PEG 400, 0.1M CHES pH 9.5 from the JCSG and WIZARDS screens respectively.



0.1M BICINE pH9, 10% MPD



0.1M CHES pH 9.5, 30% MPD

Figure 78: Initial crystallization hits for GST-IN-CTD from the robot.

To optimize the hits, several factors were modified. Purification buffer was changed to Tris pH 9, 200 mM NaCl was added to the 30% PEG 400, 0.1M CHES pH 9.5 reservoirs. Seed crystals were added to drops. Additionally, additive screens were also tested to find reagents that would improve quality. However, no significant improvement in crystal quality was observed, as all hits obtained from the additive screen looked like the original thin needles. Additionally, adding DNA or peptide did not improve crystal quality.

Hits from additive screen

Thin needles with DNA

Figure 79: GST-IN-CTD crystallization attempts**.** All attempts to improve crystal quality were unsuccessful

HIS-IN-CTD 220-270 Crystallization

Protein only

Initial screens were set up at 5.3mg/ml of protein only. Crystals were obtained in several conditions, ranging from pH 4.6 to 8.5. However, the best quality crystals were obtained at pH 6.5. The best crystals were optimized by manual hanging drops at 5mg/ml by mixing 1µl protein + 1µl reservoir and equilibrating against 500µl of reservoir at 20°C. Diffraction data

were collected using a Pilatus 2 M detector on beamline X06DA (PXIII) at the Swiss Light Source, Paul Scherrer Institut, Villigen, Switzerland. Structure determination was carried out using the CCP4 suite of programs (Potterton, Briggs et al. 2003). The models were built using the program Coot (Emsley, Lohkamp et al. 2010) and structures were determined by molecular replacement using MOLREP (Vagin and Teplyakov 2010). Structures were generated using PYMOL (DeLano 2002) and Chimera (Pettersen, Goddard et al. 2004). Dimer interfaces were generated using PISA (Krissinel and Henrick 2007).

Precipitant	**Buffer/pH**
0.01$MgCl_2$, 1M Lithium Sulfate	0.01M MES, pH5.6
0.8M Sodium Malonate	0.1M MES, pH6.5
1M Sodium Malonate	0.1M MES, pH6.5
1.2M Sodium Malonate	0.1M MES, pH6.5
1.5M Sodium Malonate	**0.1M MES, pH6.5**
0.8M Potassium Sodium Tartrate, 0.2M Lithium Sulfate	0.1M TRIS, pH 7
1.5M Potassium Sodium Tartrate	**0.1M HEPES, pH 7.5**
0.5M Ammonium Sulfate, 25% PEG 3350	0.1M HEPES, pH 7.5
1M Ammonium Sulfate, 0.2M Lithium Sulfate	0.1M TRIS, pH 8.5

Table 6: Summary of crystallization conditions and data collected for HIS CTD 220-270. Highlighted in bold are conditions in which data collected have been processed.

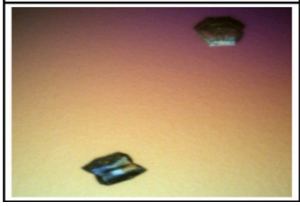
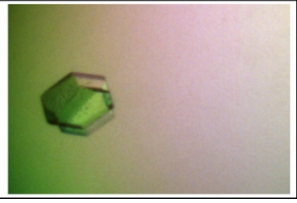
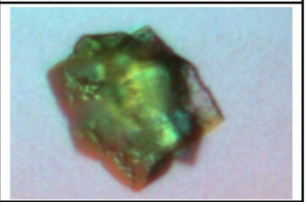
0.01M MES pH5.6, 0.01M $MgCl_2$, 1M Lithium Sulfate	**0.1M MES pH6.5 0.8M Sodium Malonate**	**0.1M Tris pH7, 1.5M Potassium Sodium Tartrate, 0.2M LiS04**	**0.1M Tris pH8.5, 1.26M$(NH_4)_2SO_4$, 0.2M Lithium Sulfate**
			

Figure 80: Crystals obtained after optimization at different pH

Data Processing and Refinement Statistics at each pH

	HIS-IN-CTD pH 6.5	HIS-IN-CTD pH 7	HIS-IN-CTD pH 7.5
Wavelength	1	1	1
Resolution range	42.9 - 1.501 (1.554 - 1.501)	18.03 - 1.5 (1.554 - 1.5)	19.29 - 1.5 (1.554 - 1.5)
Space group	P 63	P 63	P 63
Unit cell	49.5335 49.5335 41.58 90 90 120	49.4102 49.4102 39.764 90 90 120	49.81 49.81 43.1415 90 90 120
Total reflections	185318 (17637)	175848 (16825)	340867 (37120)
Unique reflections	9379 (934)	8925 (875)	8916 (997)
Multiplicity	19.8 (18.9)	19.7 (19.2)	38.2 (38.0)
Completeness (%)	97.16 (76.66)	99.91 (100.00)	91.94 (99.90)
Mean I/sigma(I)	25.73 (0.47)	35.59 (5.45)	26.74 (2.32)
Wilson B-factor	29.34	17.73	24.70
R-merge	0.04749 (8.887)	0.05086 (0.6408)	0.07057 (1.936)
R-meas	0.04879 (9.129)	0.05226 (0.6581)	0.07158 (1.962)
R-pim	0.01103 (2.071)	0.01188 (0.1494)	0.0117 (0.3175)
CC1/2	1 (0.454)	1 (0.887)	1 (0.865)
CC*	1 (0.79)	1 (0.97)	1 (0.963)
Reflections used in refinement	9119 (716)	8924 (875)	9060 (996)
Reflections used for R-free	461 (41)	436 (35)	444 (50)
R-work	0.2412 (0.6341)	0.1997 (0.2696)	0.2976 (0.4585)
R-free	0.2754 (0.4890)	0.2284 (0.3145)	0.3706 (0.5396)
CC(work)	0.603 (0.345)	0.693 (0.117)	0.944 (0.750)
CC(free)	0.551 (0.349)	0.860 (0.536)	0.890 (0.560)
Number of non-hydrogen atoms	514	548	615
macromolecules	474	474	474
ligands	1	1	1
solvent	39	73	140
Protein residues	56	56	56
RMS(bonds)	0.016	0.006	0.015
RMS(angles)	1.85	0.92	1.93
Ramachandran favored (%)	94.44	98.15	94.44
Ramachandran allowed (%)	3.70	1.85	1.85
Ramachandran outliers (%)	1.85	0.00	3.70
Rotamer outliers (%)	4.00	0.00	4.00
Clashscore	3.15	7.35	11.55
Average B-factor	47.00	23.83	35.39
macromolecules	46.68	22.66	34.49
ligands	68.30	18.42	51.36
solvent	50.40	31.55	38.34

Table 7: Refinement statistics at each pH. Statistics for the highest-resolution shell are shown in parentheses. Structures are in progress.

Crystal Structures of HIS-IN-CTD 220-270 at pH 7

The structure of HIS-IN-CTD 220-270 pH 7 was solved at a resolution of 1.5Å. As expected, it consists of 5 β strands that form 2 anti parallel β sheets connected by loops. Interestingly, electrostatic potential representation surface shows a patch of solvent exposed hydrophobic residues on β sheet, which may explain why the IN-CTD is prone to aggregation. Additionally, there is a patch of positive and negative residues exposed.

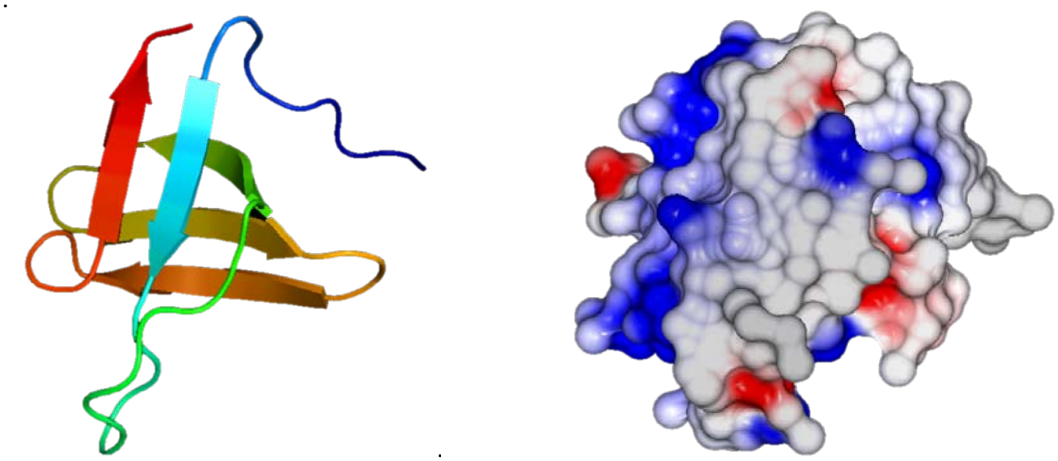


Figure 81: HIS-IN-CTD 220-270 monomer. Left: Cartoon representation of HIS-IN-CTD 220-270. The long dark blue loop on the N-terminus represents 5 of the 6X HIS Tag. Right: Electrostatic potential representation of HIS-IN-CTD 220-270, showing exposed hydrophobic residues

One crystal packing contact found to be important was driven by the hexahistidine tag, bound to metal ion that is likely to be a nickel ion. The only probable source of the nickel ion is the from the Ni^{2+} beads from purification. The histidines surround the nickel, as in a hexamer, and each histidine side chain is about 2.1Å from the Ni^{2+} ion

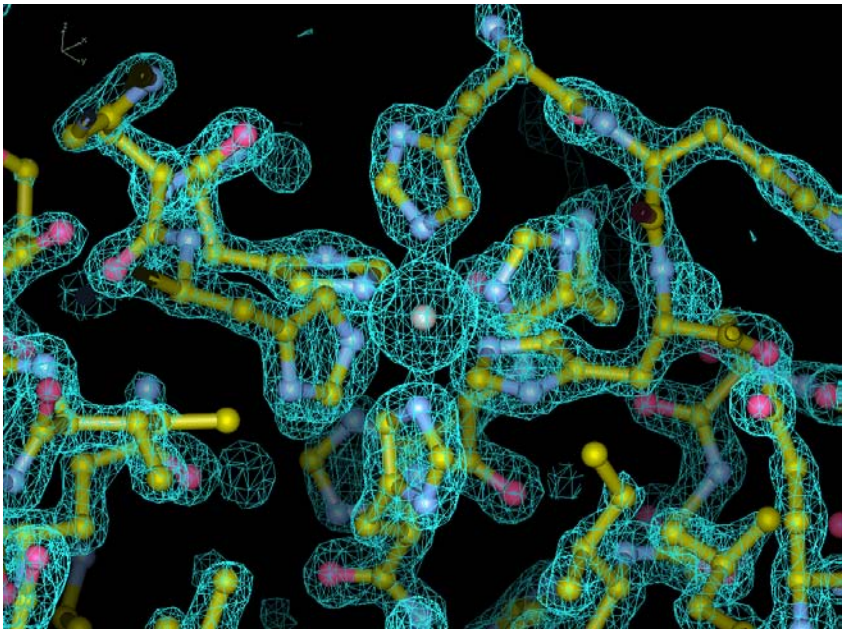


Figure 82: Hexahistidine tag bound to Ni^{2+} ion form crystal packing contacts. Map shows 2Fobs-Fcalc density contoured at 2.3 σ

We found three possible surface contacts for dimer formation, involving different residues in the protein. The first dimeric interface involves residues R231 to V234 on one monomer and residues Y226, W235 to P238, I268 to D270 on the second monomer. This interface involves charge based interactions (top half surface consists of charged residues) and hydrophobic interaction (second half surface consists of hydrophobic residues). Interestingly, peaks corresponding to R231 and D232 are not visible on the ^{15}N HSQC. This is possibly because they are buried in the dimer in the ^{15}N HSQC conditions. The interface area for this dimeric surface is $216A^2$, with solvation energy of -3.7kcal/mol, and 1 hydrogen bond.

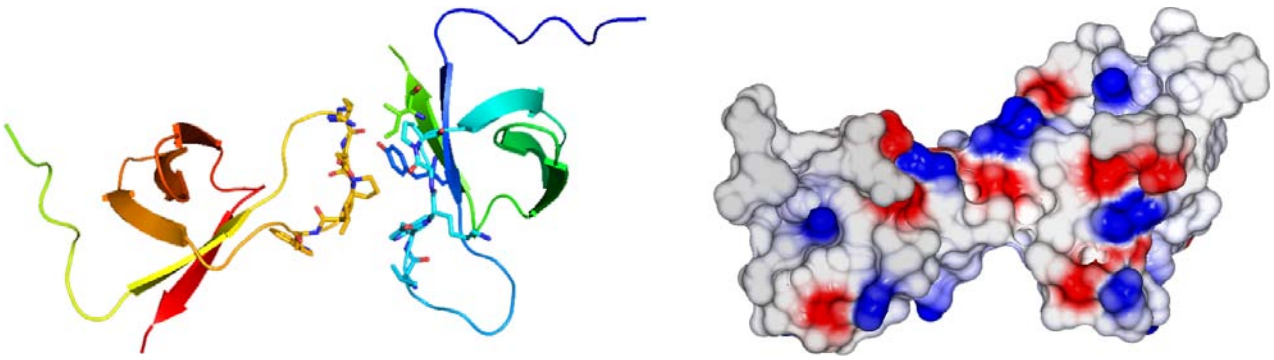


Figure 83: IN-CTD Interface 1. Left: Cartoon representation of the first dimeric interface, highlighting the side chains of residues R231 to V234 on the left and Y226, W235 to P238, I268 to D270 on the right. Right: Electrostatic representation of dimeric interface.

The second dimeric interface involves residues N254 and K240 on one monomer and R263 and K264 on the second monomer. This provides a plausible explanation for the peak corresponding to N254 not appearing on ^{15}N HSQC. This interface is driven by charge-based interactions. The interface area for this dimeric surface is 139A^2, with solvation energy of 2.47kcal/mol, and 3 hydrogen bonds.

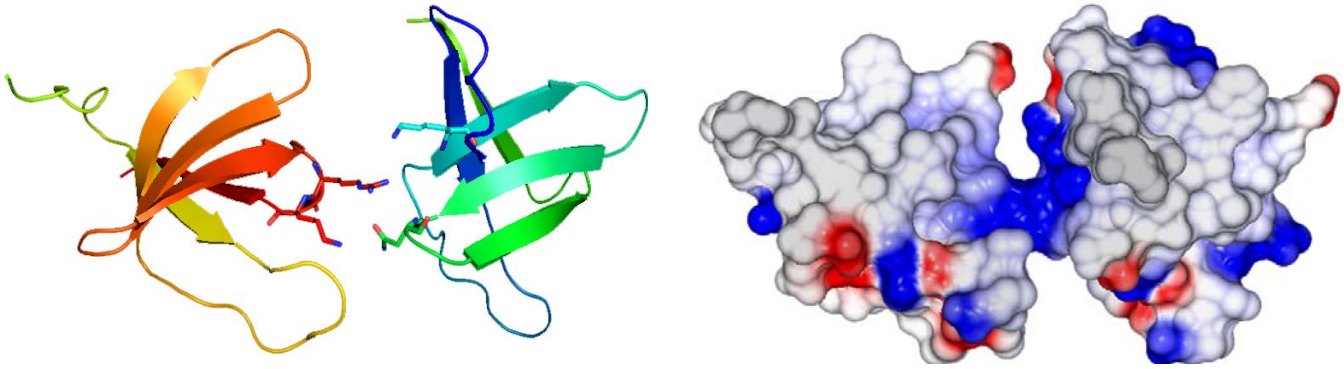


Figure 84: IN-CTD Interface 2. Left: Cartoon representation of the second dimeric interface, highlighting the side chains of residues 263R and 264K on the left, N254 and K240 on the right. Right: Electrostatic representation of dimeric interface.

The third dimeric interface involves residues S255 and D256 on one monomer and R269 on the second monomer. This is consistent with the peak corresponding to S255 not appearing on 15N HSQC. This interface involves charge-based interactions. The interface area for this dimeric surface is 62A^2, with solvation energy of 0.23kcal/mol, and 1 hydrogen bond.

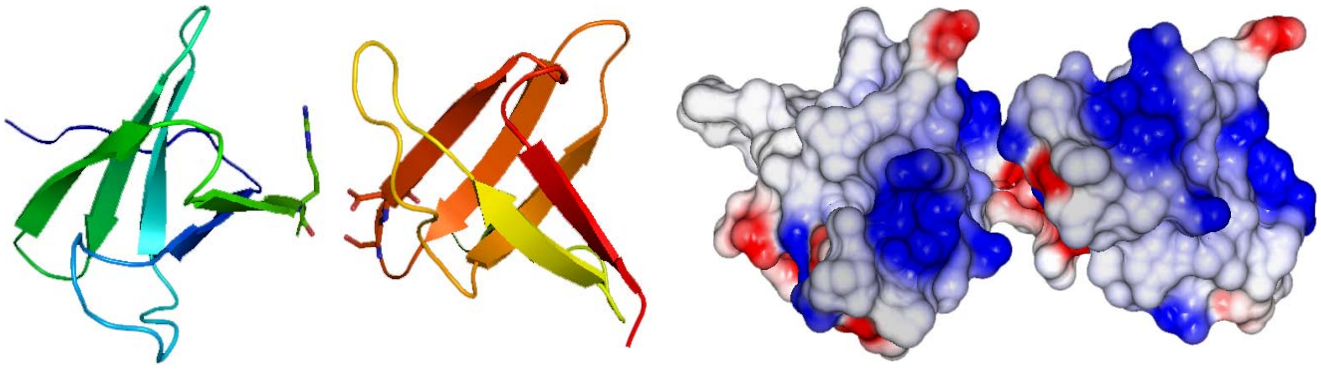


Figure 85: IN-CTD Interface 3. Left: Cartoon representation of the third dimeric interface, highlighting the side chains of residues R269 on the left, S255 and D256 on the right. Right: Electrostatic representation of dimeric interface

A superposition of structures at pH 6.5, 7 and 7.5 shows no differences in the core β-strands. However, differences were observed in the loops connecting the β-strands, especially between loops connecting residues R228 and V234, and P261 and D270. These results indicate that the residues in these regions are, and sensitive to changes in the protein environment.

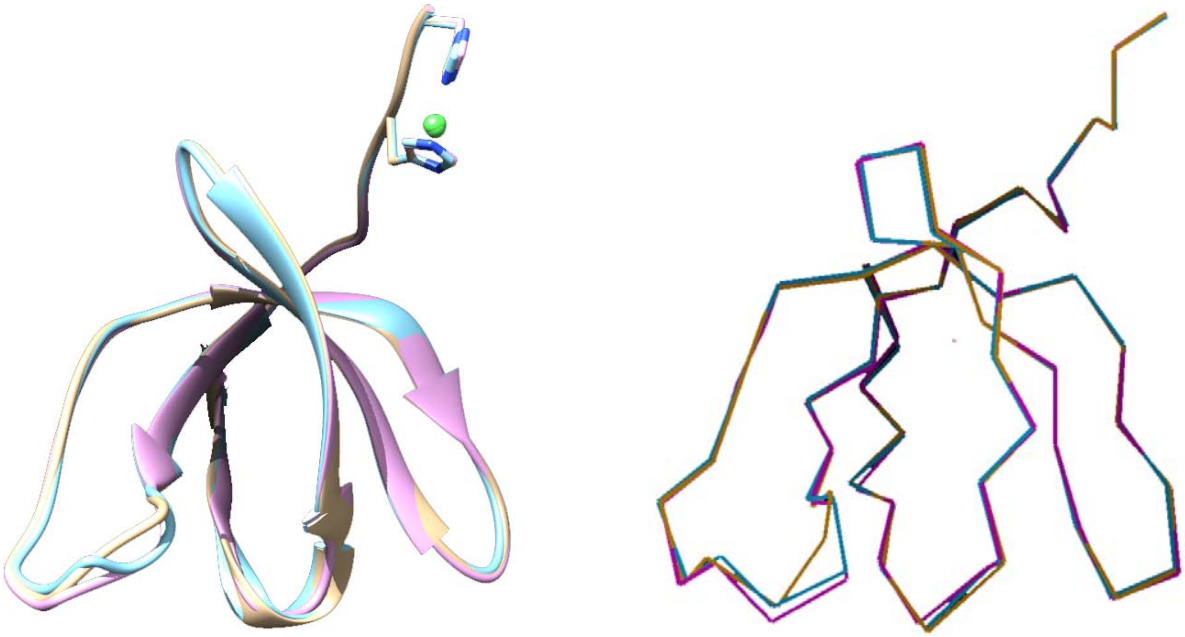


Figure 86: Comparison of structures at different pH. Left: Cartoon representation of superposition of HIS-IN-CTD 220-270 at pH 6.5 (gold), pH 7 (blue) and pH 7.5 (pink) Right: Ribbon representation of of HIS-IN-CTD 220-270 at pH 6.5 (gold), pH 7 (blue) and pH 7.5 (pink)

RMSD deviation was calculated using the formula:

$$\sqrt{\frac{\sum_{i=1}^{n}(Xi - Yi)^2}{n}}$$

where $(Xi - Yi)$ – distance between two atoms
n- number of atoms

Overall RMSD for Cα and all atoms was calculated and shown in the table below. The highest deviation on all atoms was found to be between pH 6.5 and pH 7.5, with a RMSD of 3.83Å, suggesting overall changes in conformation depending on pH.

RMSD Cα	pH 6.5	pH 7	pH 7.5
pH 6.5	0		
pH 7	0.41- Cα 1.44 – All atoms	0	
pH 7.5	0.43 – Cα 3.83 – All atoms	0.3 – Cα 1.43 – All atoms	0

Table 8: Overall RMSD differences between the structures at each pH

Between residues R228 and V234, Cα deviations were observed at pH 7 and pH 7.5. The deviation is consistent with increasing pH, indicating that residues in this region are sensitive to pH changes.

RMSD Cα	pH 6.5	pH 7	pH 7.5
pH 6.5	0		
pH 7	3.0	0	
pH 7.5	1.13	2.9	0

Table 9: RMSD differences between Cα of the structures at residues at R228 and V234

Deviations were also observed to a lesser extent between P261 and D270.

RMSD Cα	pH 6.5	pH 7	pH 7.5
pH 6.5	0		
pH 7	1.24	0	
pH 7.5	1.7	1.4	0

Table 10: RMSD differences between Cα of the structures at residues at P261 and D270

Co-crystallization with peptide

Initial screens with H4K20Me1 peptide were set up at 1:1.2 ratio with initial protein concentration of 5.3mg/ml. Crystals were obtained in several conditions, ranging from pH 4.6 to 7. The best crystals were optimized by manual hanging drops at 5mg/ml by mixing 1µl protein + 1µl reservoir and equilibrating against 500µl of reservoir at 20°C.

In order to ensure that the peptide co-crystallizes with the protein, manual screens were set up with a protein: peptide ratio of 1:5, or 1:10, starting from 2.5mg/ml or 5mg/ml of protein. No crystals were obtained in drops starting from 2.5mg/ml. With drops starting at higher protein: peptide ratio, the crystals took longer to appear, sometimes up to two weeks.

Precipitant	**Buffer/pH**	**Protein:Peptide Ratio**
0.01$MgCl_2$, 1M Lithium Sulfate	0.01M MES, pH5.6	1:5, 1:10
1.2M Sodium Malonate	0.1M MES, pH6.5	1:5, 1:10
1.5M Sodium Malonate	0.1M MES, pH6.5	1:10
0.8M Potassium Sodium Tartrate, 0.2M Lithium Sulfate	0.1M TRIS, pH 7	1:5, 1:10
1.5M Potassium Sodium Tartrate	**0.1M HEPES, pH 7.5**	1:10

Table 11: Summary of crystallization conditions and data collected for HIS CTD 220-270 plus peptide at a ratio of 1:5 or 1:10. Highlighted in bold are conditions in which data collected are being processed.

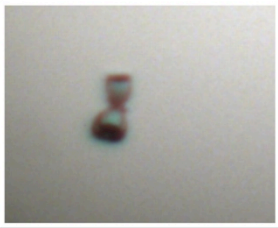
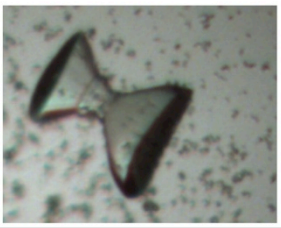
0.01M MES pH5.6, 0.01M $MgCl_2$, 1.2M Lithium Sulfate	0.1M MES pH6.5, 1.5M Sodium Malonate	0.1M Tris pH7, 1.5M PST, 0.2M LiSO4	0.1M HEPES pH7.5, 1.2M Potassium Sodium Tartrate
			

Figure 87: Crystals obtained at each pH with peptide at 1:10 ratio

Co-crystallization with peptide and DNA

Initial screens with H4K20Me1 peptide were set up at 1:1.2:1.2 ratio with starting protein concentration of 5.3mg/ml. Interestingly, crystals with DNA were obtained in conditions that are completely different from co-crystals without DNA. The best crystal form was obtained in 0.1M MES pH 6.5, 25% PEG monomethyl ether 550. However, these crystals could not be reproduced. Further screening with other low molecular weight PEGS like 300 were successful, with the best crystal at 0.1M MES pH 6.5, 30% PEG 300. However, the limit of diffraction of this crystal was 7Å.

In order to improve crystal diffraction quality, manual screens were set up at different temperatures (20°C, 27°C, 34°C), at varying protein concentrations (2.5mg/ml, 5mg/ml), and at varying protein/peptide/DNA ratios (1:1.2:1.2, 1:5:1.2). No crystals were obtained at higher ratios of 1:5:1.2. Additionally, no crystals were obtained at 34°C.

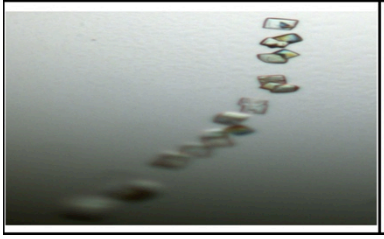
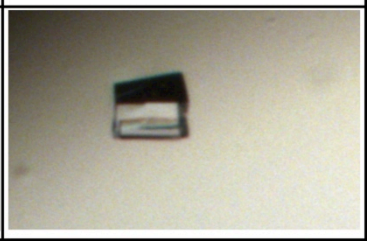
0.1 MES pH 6.5, 30% PEG 300 20°C 2.5mg/ml - 1:1.2:1.2	0.1M MES pH6.5, 30% PEG 300 27°C 5mg/ml – 1:1.2:1.2
	

Figure 88: Crystals obtained at 24°C and 27°C with peptide and DNA

The crystal at 27°C was tested at the SLS synchrotron. However, there was no improvement in diffraction from 7Å.

Precipitant	Buffer/pH	Protein:Peptide: DNA Ratio
30% PEG 300	0.1M MES, pH6.5	1:1.2:1.2
20% PEG 300	0.1M MES, pH6.5	1:1.2:1.2

Table 12: Summary of crystallization conditions and data collected for HIS CTD 220-270 plus peptide and DNA

The best diffracting DNA crystal was obtained from manual screens in 0.1M MES pH 6.5, 20% PEG 400 at 24°C, at a ratio of 1:1.2:1.2 with 5mg/ml protein. This crystallization trial was set up 6 months earlier, and diffracted to a resolution of 4Å from a synchrotron source. The unit cell parameters at 40.613 40.613 322.410 90.000 90.000 90.000 are larger than the parameters for protein only. The space group was determined to be either $P4_12_12$ or $P4_32_12$. This ambiguity can only be resolved by solving the crystal structure.

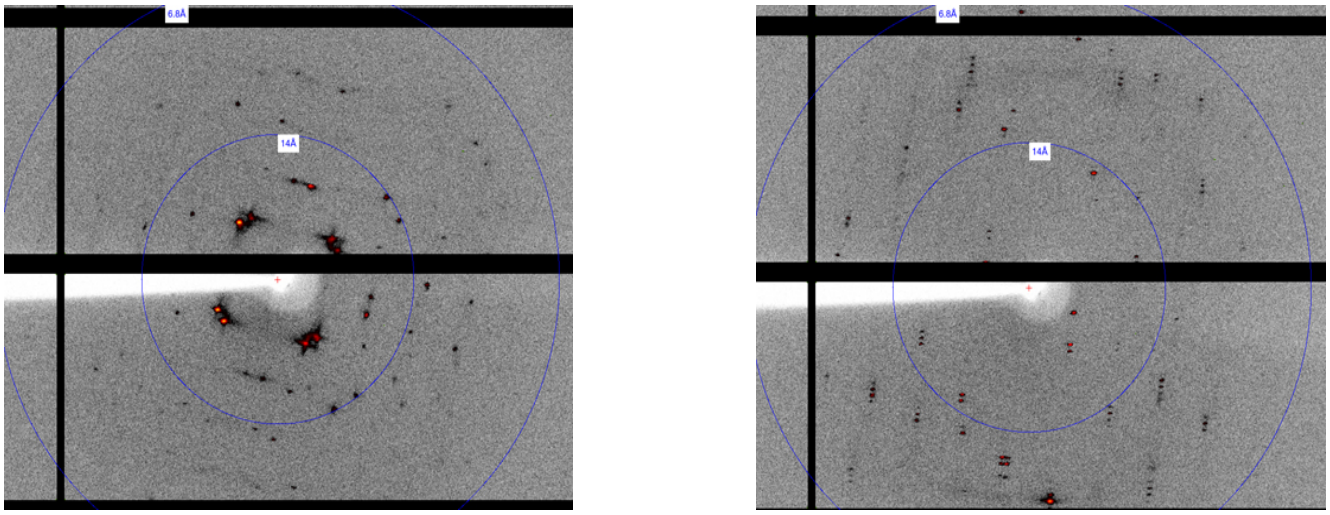


Figure 89: X-ray diffraction pattern for crystals with DNA

Model for H4K20Me1 peptide binding

Given information about perturbed residues in the presence of peptide at pH 8, we highlighted these residues on the x-ray structure.

The figure below shows residues that disappear in the presence of 2mM peptide – F223, W243, L241 and V250. It is surprising that most of these residues, with the exception of F223 are in the middle of a β strand. However, given their close proximity to each other and their hydrophobic nature, it is possible that they become more buried in some hydrophobic interactions in the presence of peptide.

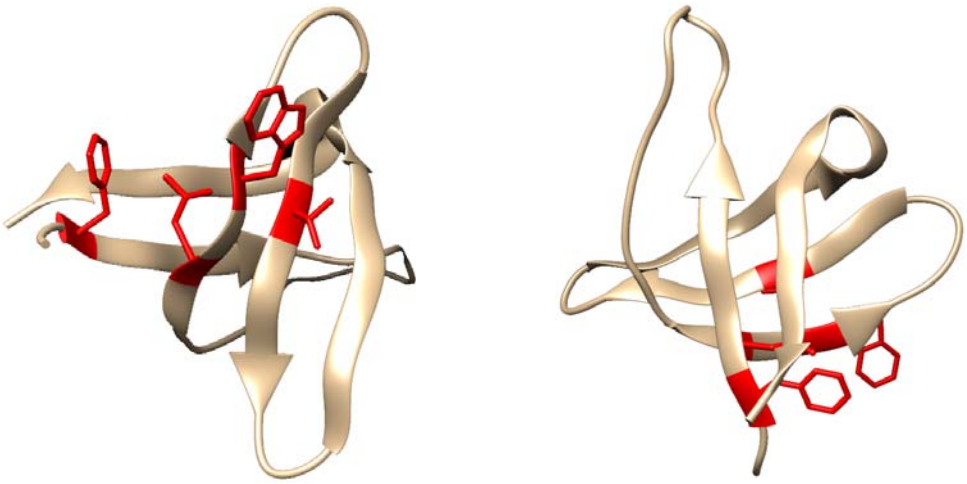


Figure 90: Residues that disappear in the presence of peptide. Side chains are highlighted in red**.**

More interestingly, mapping residues that shift in the presence of peptide presents a possible interaction surface. The location of these residues suggests the peptide binds to this "hydrophobic patch" on β strand 1 and 5, which correspond to the N and C terminus of the IN-CTD. Residues in this patch include I268, I267, K266 and R224, V225, Y226 and Y227. Given the close proximity of the other perturbed residues, it is plausible that peptide binding to nearby residues also induces chemical shift changes on these residues, especially A248 and K244.

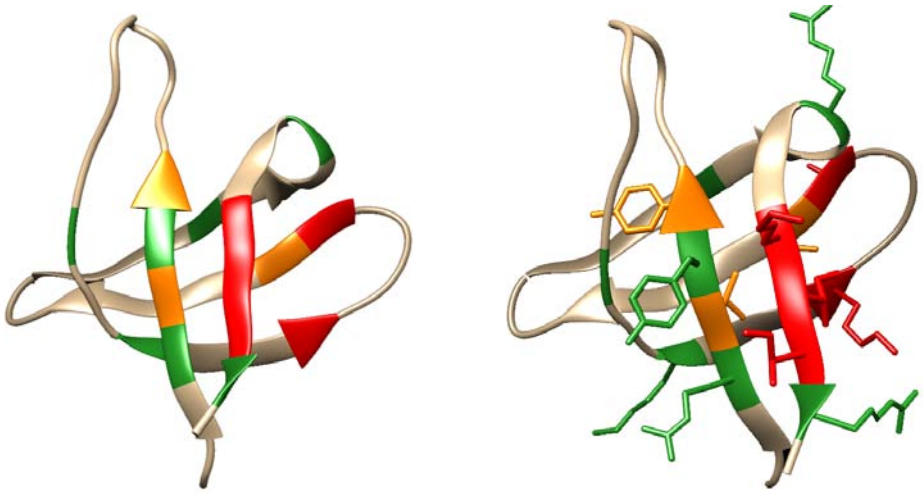


Figure 91: Residues shifting in the presence of peptide. Figure on the right shows side chains. Residues are color coded by the degree of perturbation, with red showing residues with high chemical shift, orange showing residues with medium chemical shift, and green showing the residues with the lowest chemical shift.

With the NMR structure and perturbation data, HADDOCK (Dominguez, Boelens et al. 2003, van Zundert, Rodrigues et al. 2016) was used to model the binding of the peptide to the protein. There were several docking possibilities predicted by HADDOCK, probably due to the lack of distance restraints from NMR experimental data. More experimental data would be needed to improve the docking prediction.

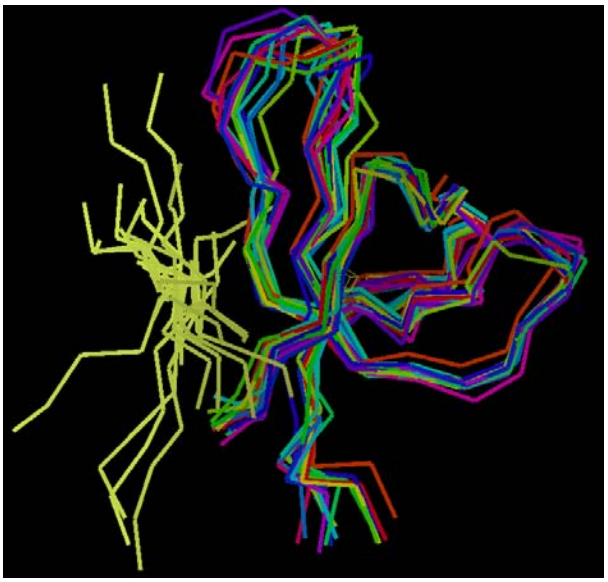


Figure 92: First round of docking run. No conclusions can be made due to bad quality of data

Using DOCKSCORE(Malhotra, Mathew et al. 2015), a webserver that ranks docked poses bases on factors such as surface area, conservation of residues, and the presence of hydrophobic resides, a weighted score was assigned for each predicted model. The docking model with the highest score is shown. The methylated lysine seems to be in close proximity to the predicted binding pocket. However, no conclusions can be made based on this structure.

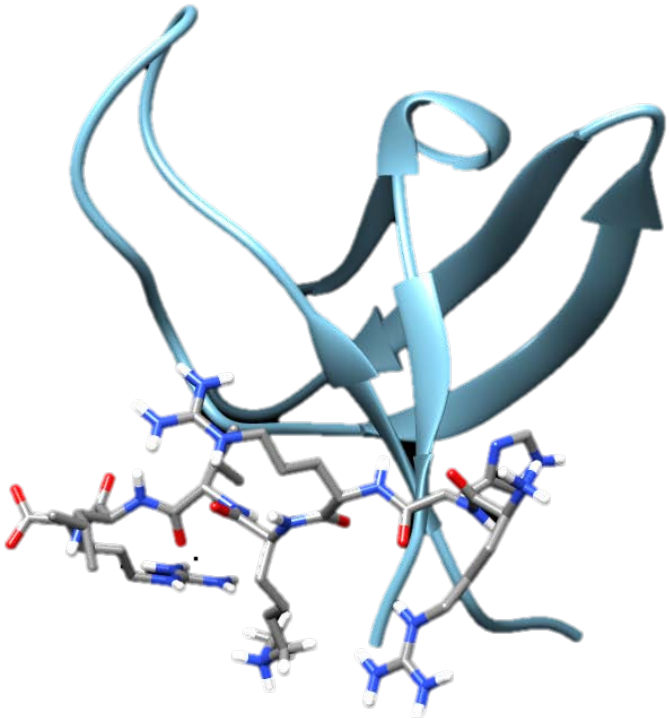


Figure 93: Docking model with best-weighted score from DOCKRUN.

DISCUSSION

Protein Production and Purification

Published protocols of IN-CTD purification were performed in the presence of denaturing agents such as urea or guanidinium chloride (Lutzke, Vink et al. 1994, Lodi, Ernst et al. 1995). In order to avoid denaturing agents, we opted to purify in high salt conditions, with detergents such as CHAPS, or low salt with additives such as arginine(Leibly, Nguyen et al. 2012).

With the GST-tagged IN-CTD, we could produce stable and soluble protein for biochemical studies. However, this protein could not be used for structural studies. The HIS-tagged protein produced in the Gateway construct was poorly folded, probably due to the 32 extra residues at the N-terminus. While the Gateway technology offered a big advantage in the ease of cloning, it produced protein of poor quality in our conditions. By switching to the pET15 expression vector, we could improve purification yield, protein stability and solubility.

Biochemical Studies

It was possible to confirm the interaction of the H4K20Me1 peptide with the IN-CTD using several techniques. trNOESY and WaterLOGSY confirm the presence of a complex evident by a change in NOEs sign in the presence of GST-IN-CTD and peptide. Similar to results shown by our collaborators (Vincent Parissi, Bordeaux), the binding of the isolated IN-CTD to histone tails seems to be dependent on methylation state, as IN-CTD binds to the peptide representing the monomethylated histone tail, with a K_d of 0.8µM. The IN-CTD binds to the non-, d- and tri- methylated peptide with a K_d of H4K20me1, H4K20me2 and H4K20me3 to be, 4.7µM 3.75µM, 5.2µM respectively. In the presence of a competing non-methylated peptide, the IN-CTD binds to the mono-methylated peptide with a similar Kd of 1.8µM, suggesting a preference for the mono-methylated state. Overall, our results confirm that the IN-CTD is capable of mediating interactions with histone tails and plays an important role in nucleosomal integration.

While the IN-CTD of PFV has been shown to interact with H2A/H2B histone dimer surface (Maskell, Renault et al. 2015), there is currently no data about the affinity of these interactions. However, our results are consistent with the MSL3 chromodomain was shown to bind H4K20 in the micro-molar range (Kim, Blus et al. 2010)

Perspectives

It would be interesting to perform MST experiments with the full length IN to determine the binding constant with all the IN domains present.

Structures of protein only

We solved a high resolution-ray structure at 3 different pH conditions. At pH 7, we detected 3 possible dimer interfaces. Interestingly, many of the residues detected to be involved in dimer formation (S230, R231, D232, P233, N254 and S255) are not visible in ^{15}N HSQC experiments. This data indicates that these residues are probably buried in NMR assignment conditions, and these contacts are not artefacts due to crystal packing. These structures display high flexibility, especially in residues in loops. Superposition of structures at varying pH shows no movement in the β strands, suggesting the loops are important for the flexibility of the IN-CTD and may play a role in its interactions. Our NMR structure correlates with the crystal structure, since β strands are superimposable, and high flexibility are observed in the loops.

Interactions with peptide (^{15}N HSQC data)

We observed binding with the peptide corresponding to monomethylated H4K20me1 under all the conditions we tested including 1M NaCl, 150mM NaCl, using the full length (220-288) and truncated construct (220-270) and at pH 7 and pH 8. Most experiments were carried out using the 220-270 constructs due to better HSQC quality. This is consistent with pushed x-ray structures where the density of 271-288 was not clearly defined (Chen, Krucinski et al. 2000), confirming high flexibility in this region.

In all conditions, three effects were observed in the presence of peptide: chemical shift moving, new peaks appearing with increasing intensity, and some peaks disappearing. In order to identify residues that are perturbed in the presence of peptide, we carried out 3D experiment and assignments. 38 out of 51 residues could be successfully assigned. 3 out of 12 were expected, because they are prolines. The unassigned residues were predicted to be in flexible regions like loops.

Overall, similar residues were observed in the interaction between HIS-IN-CTD 220-270 and pH 7 and pH 8. Residues such as I267, I268, K266 and K244 are perturbed to same extent in both pH conditions, indicating that they play a crucial role in peptide interactions. However, other residues are affected to different extents at each pH. At pH 8 in the presence of 2mM peptide, there is a severe change in ^{15}N HSQC between 7.9ppm and 8.5ppm. Additionally, several new peaks appear that were visible at pH 7. Assignment data from Eijkelenboom et al (personal communication) suggests that these residues correspond to E246, G247, S255 and N254.

Based on our results, we have come up with a hypothesis about the differences in conformation observed in the presence of peptide at each pH. We hypothesize that the differences observed in ^{15}N HSQC at pH 7 and pH 8 with protein only are due to different oligomeric states. For example, at pH 7, we might have a mix of dimers or other higher order oligomers, which accounts for the number of peaks we observe at pH 7. At pH 8, this equilibrium is shifted, and we have a more homogeneous mix, evident by the better quality of ^{15}N HSQC. In the presence of 2mM peptide at pH 8, the IN-CTD becomes more oligomeric, changing the quality of spectrum between 7.9ppm and 8.5ppm. This is further confirmed by the similarities between the spectrum at pH 8 plus peptide and the spectrum at pH 7. Movements in the loops probably drive this change in multimerization states. At pH 7, this dramatic shift is not observed because the protein is already multimerized and the conformation is less perturbed with peptide binding. This hypothesis is further supported by the intasome structure(Passos, Li et al. 2017). Residues in the CTD that are implicated in higher order multimerization such as K240, K244 and R269 are affected by peptide binding in our experiments. Moreover, the functional oligomerization state of integrase is at least a tetramer. Therefore, it is plausible that the IN-CTD oligomerizes upon binding to histone tails in order to perform the strand transfer reaction.

Interactions with peptide (Structural data)

We mapped the residues with chemical shift changes at pH 8 onto our crystal structure. Here, we observe that the major residues involved in peptide binding are on the N and C terminus of IN-CTD. Interestingly, several of these residues are hydrophobic in nature, consistent with members of the Royal Domain Superfamily. While the IN-CTD may not possess a hydrophobic core in its monomeric form, we hypothesize that these resides form a

“hydrophobic patch” that serves as the binding surface for the monomethylated H4K20Me1. Our preliminary docking experiments suggests the methylated lysine is in close proximity and may interact with this patch.

Based on our data we propose that in the presence of the histone H4 tail, IN interacts with the H4K20 via hydrophobic patch on the CTD. This binding induces movements in the loops, possibly causing the CTD to multimerize, or form new contacts, and allowing integrase to form stable oligomers in order to integration to occur.

In conclusion, we show that the IN-CTD is indeed able to interact with the monomethylated H4K20, and we identify the potential binding sites for these interactions. Crystal or NMR structures in the presence of peptide are in progress and would be needed to prove the correctness of our model.

CHAPTER 3 – CO-EVOLUTION STUDIES

MATERIALS AND METHODS

DNA cloning

The IGBMC Molecular Biology Service carried out the cloning of all constructs generated in this project using the restriction-ligation cloning strategies.

Restriction Cloning

One round of PCR was performed with the Phusion polymerase and p8.91 plasmids corresponding to each WT and mutant protein as the template. At the end, NcoATGglyHis6_220-270_StopBamHI constructs for were generated for each construct. The protocol used above (PCR mix, agarose gel electrophoresis, PCR clean up) was repeated. All primer sequences are described in the table below.

Desired Product	**Group/ Subtype**	**Primer Sequence**
HIS-220-270	A2 WT	F-GAGATATACCATGGGCCATCATCATCATCATCAC ATTCAAAATTTTCGGGTTTATTACAGG R- GCAGCCGGATCCTTA ATCCCTAATGATCTTTGCTTTTCTTCTTGG
HIS-220-270	N222K/K240Q	F- GAGATATACCATGGGCCATCATCATCATCATCAC ATTCAAAAATTTCGGGTTTATTACAGG R- GCAGCCGGATCCTTA ATCCCTAATGATCTTTGCTTTTCTTCTTGG
HIS-220-270	K240Q/N254K	F- GAGATATACCATGGGCCATCATCATCATCATCAC ATTCAAAATTTTCGGGTTTATTACAGG R- GCAGCCGGATCCTTA ATCCCTAATGATCTTTGCTTTTCTTCTTGG
HIS-220-270	O WT	F - GAGATATACCATGGGCCATCATCATCATCATCAC ATTCAAAAATTTCGGGTCTATTACAG R- GCAGCCGGATCCTTA ATCTCTGATTATTTTTGCCTTCCTTCTTGG

Table 13: Table describing primers used in restriction cloning

Protein Production

All protein production was performed as described in the previous chapter (38.). All IN-CTDs used in this project were purified using the same strategy. Cells were resuspended in lysis buffer -25mM HEPES pH 8, 1M NaCl, 10mM imidazole at a ratio of 10ml/gram of cells. Cells were lysed using a sonicator with a 13mm probe for 1 min/g of cells with pulse every 2

seconds at 40% amplitude at 4°C. Following ultracentrifugation at 185000xg for 1 hour, the supernatant was loaded on the 1ml HisTrap FF Crude column (GE Healthcare) with flow rate of 1ml/min using the AKTA purifier. Protein was eluted using a gradient up to 500mM Imidazole. Protein concentration/quality was analyzed using the Nanodrop. Subsequently, the protein sample was concentrated using the Amicon Ultra 15ml with a MWCO of 3 kDa for the next purification step. A second step of purification was carried out using the S7516/60 column (GE Healthcare) in 25mM HEPES pH 8, 1M NaCl. Samples were dialyzed into 25mM HEPES pH 8, 150mM NaCl for crystallization.

Crystallization

Manual drops were setup in Hampton Research 24 well VDX plates with reservoir conditions in which the HIS-IN-CTD 220-270 crystals were obtained. In parallel, the TTP Labtech's mosquito crystal, using the Sparse Matrix strategy was used to determine other crystallization conditions. Screens tested included ANIONS, CATIONS and WIZARDS.

RESULTS

Purification Results

A2 CTD purification

5.3g of cells from 1L production were resuspended in 53ml of lysis buffer. 10mg of protein was recovered following affinity chromatography. The sample was concentrated to 2.4mg/ml in 4.3ml for analysis by gel filtration. Following size exclusion chromatography, the amount of protein was reduced to 5mg. This was due to the presence of a large amount of aggregates. Following GF, the purified protein was stable and could be concentrated up to 5mg/ml in 150mM NaCl.

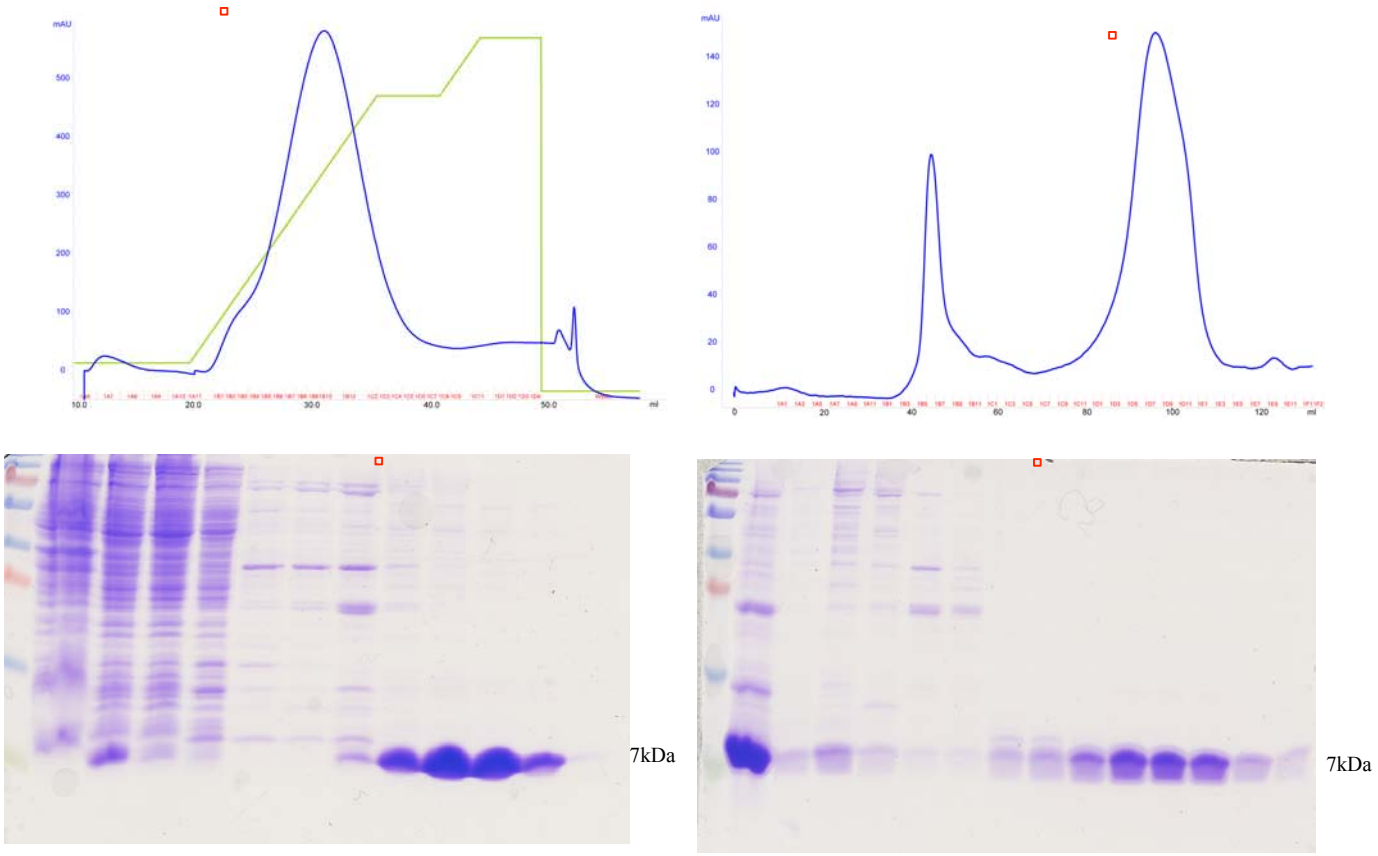


Figure 94: 1L Purification of A2 HIS-IN-CTD Left: Gel and Chromatogram following affinity purification with HISTrap FF crude column. Right: Gel and Chromatogram following purification with S75 16/60 column. Red boxes highlight pooled fractions

N222K/K240Q CTD purification

4.7g of cells from 1L production were resuspended in 47ml of lysis buffer. 8.85mg of protein was recovered following affinity chromatography. The sample was concentrated to 1.99mg/ml in 4.3ml for analysis by gel filtration. Following size exclusion chromatography, the amount of protein was reduced to 4.8mg. As with the A2 WT, there was a presence of a large amount of aggregates. This mutant was less stable than the WT and had a higher tendency to precipitate in solution. However, it could also be concentrated up to 5mg/ml in 150mM NaCl.

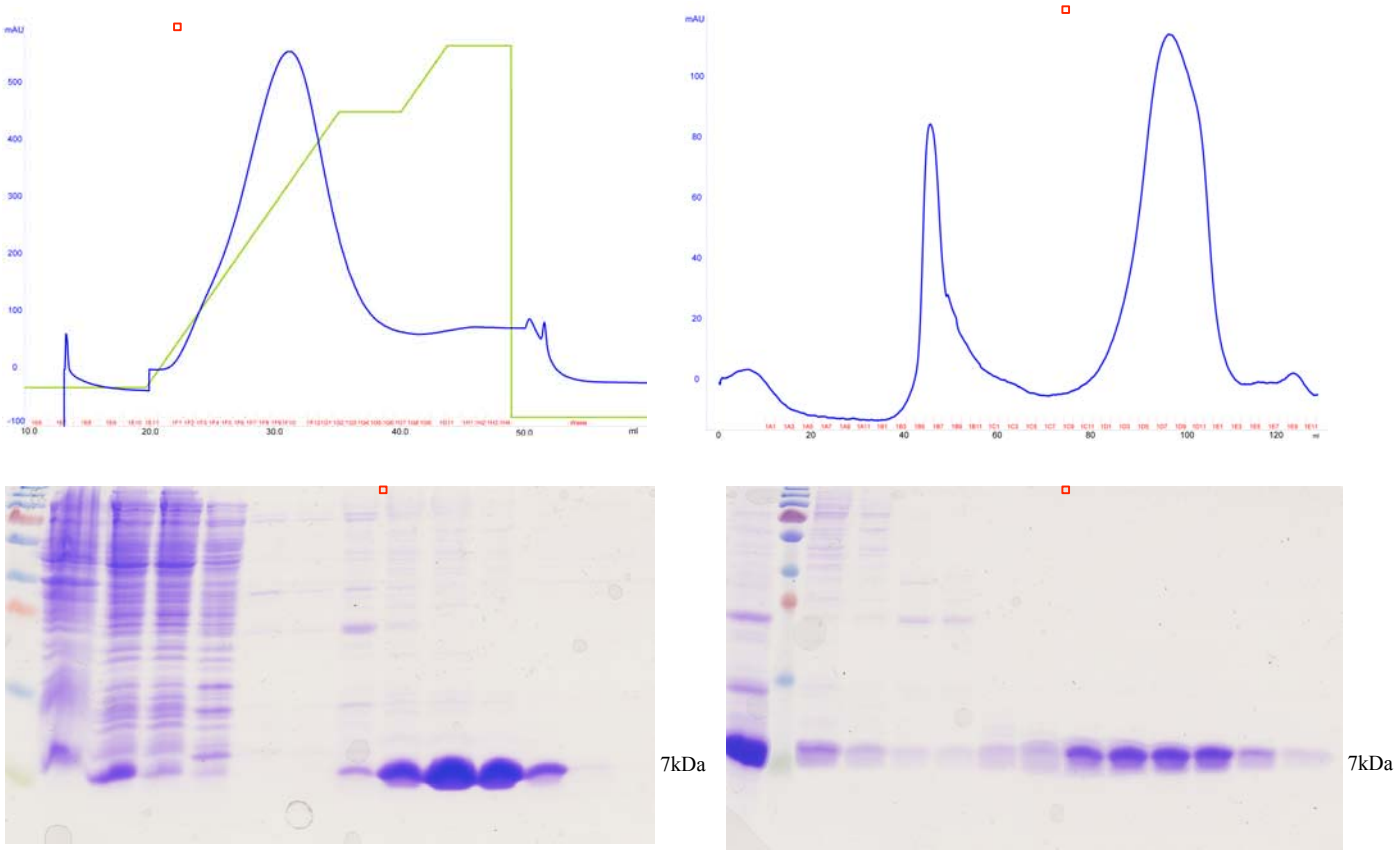


Figure 95: 1L Purification of N222K/K240Q HIS-IN-CTD. Left: Gel and Chromatogram following affinity purification with HISTrap FF crude column. Right: Gel and Chromatogram following purification with S75 16/60 column. Red boxes highlight pooled fractions

K240Q/N254K CTD purification

4.8g of cells from 1L production were resuspended in 48ml of lysis buffer. 9.7mg of protein was recovered following affinity chromatography. The sample was concentrated to 2.43mg/ml in 4ml for analysis by gel filtration. Following size exclusion chromatography, the yield was reduced to 4mg. Compared to the other proteins; this mutant contained the highest amount of aggregates during size exclusion chromatography. Additionally, it was the least stable, and was most prone to precipitation. It was possible to concentrate to 5mg/ml, but it precipitated more than the other proteins during concentration.

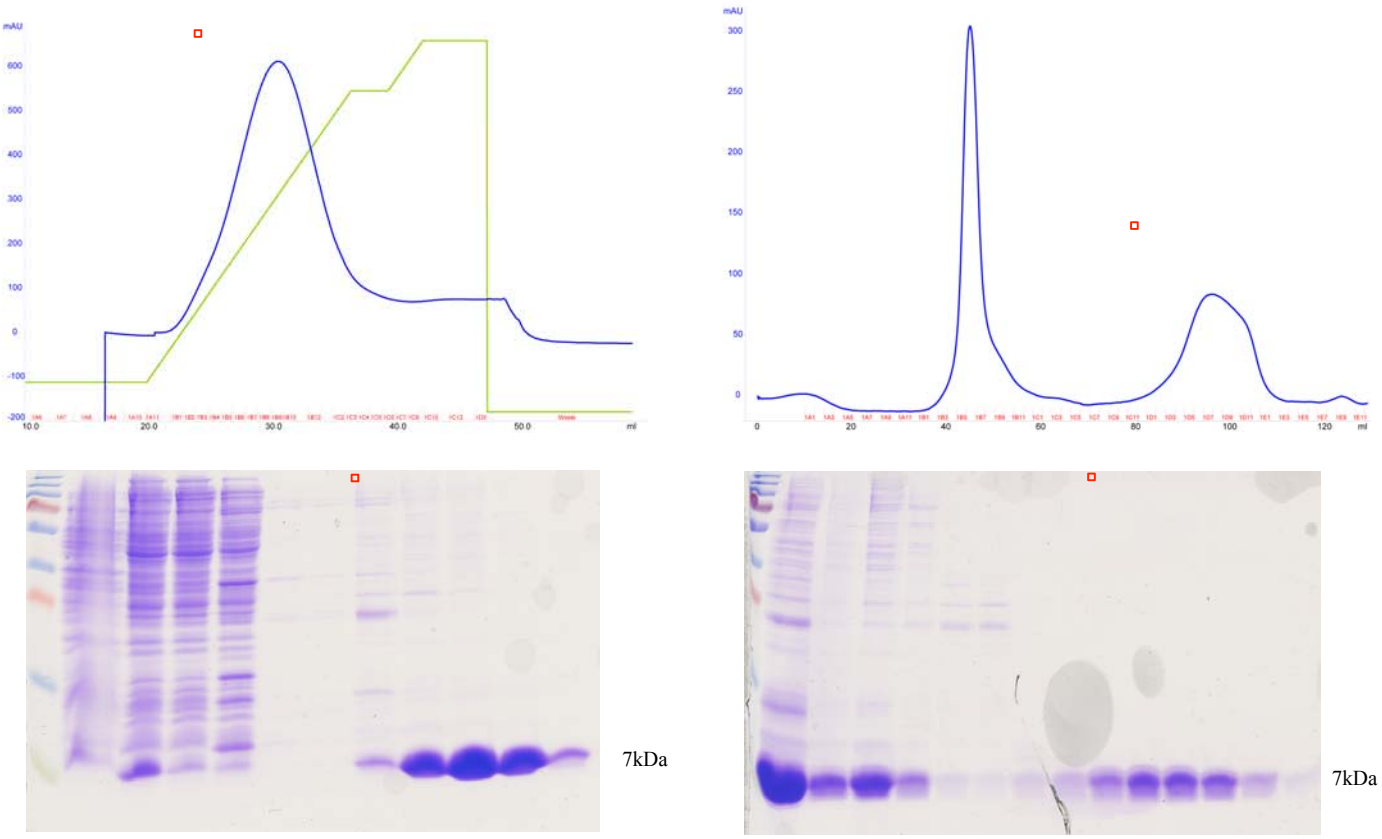


Figure 96: 1L Purification of K240Q/N254K HIS-IN-CTD. Left: Gel and Chromatogram following affinity purification with HISTrap FF crude column. Right: Gel and Chromatogram following purification with S75 16/60 column. Red boxes highlight pooled fractions

O CTD purification

6.2g of cells from 1L production were resuspended in 62ml of lysis buffer. 15mg of protein was recovered following affinity chromatography. The sample was concentrated to 3.5mg/ml in 4.5ml for analysis by gel filtration. Following size exclusion chromatography, the yield was reduced to 10.5mg. The O CTD was the most soluble protein, with the least amount of aggregates during gel filtration. Additionally, it was the most stable in solution, and could be concentrated up to 7mg/ml in 150mM NaCl without precipitating.

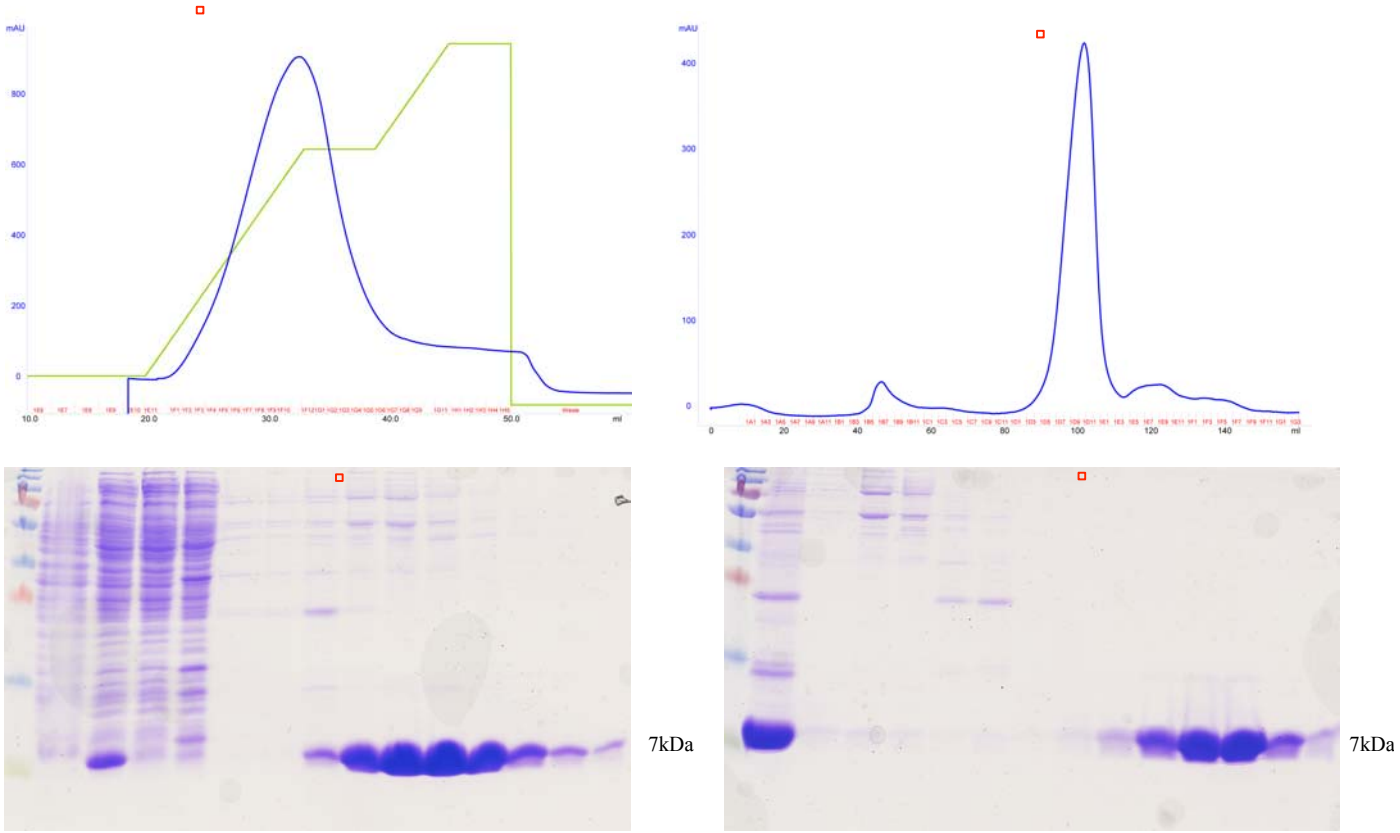


Figure 97: **1L Purification of O HIS-IN-CTD.**
Left: Gel and Chromatogram following affinity purification with HISTrap FF crude column. Right: Gel and Chromatogram following purification with S75 16/60 column. Red boxes highlight pooled fractions.

Expression Vector	Purification Tag	Cloning Strategy	Sequence limit	Group/Subtype	Molecular Weight	Extinction Coefficient	Theoritical isoelectric Point
pET 15b	6X HIS-	Restriction	220-270	A2 Wild-type	7021.1	13980	10.2
pET 15b	6X HIS	Restriction	220-270	A2 N22K/K240Q	7035.12	13980	10.2
pET 15b	6X HIS	Restriction	220-270	A2 K240Q/N254K	7035.12	13980	10.2
pET 15b	6X HIS	Restriction	220-270	O wild type	7019.12	13980	10.28

Table 14: Summary of protein used in this study

Crystallization

A2

Manual hanging drops of the A2 CTD were set up by mixing 2µl protein at 5mg/ml+ 2µl reservoir and equilibrating against 500µl of reservoir at 20°C. Crystals were obtained in several conditions. Additionally, several crystals were obtained in screens set up using the mosquito robot with protein concentration at 4.3mg/ml that were set up in parallel.

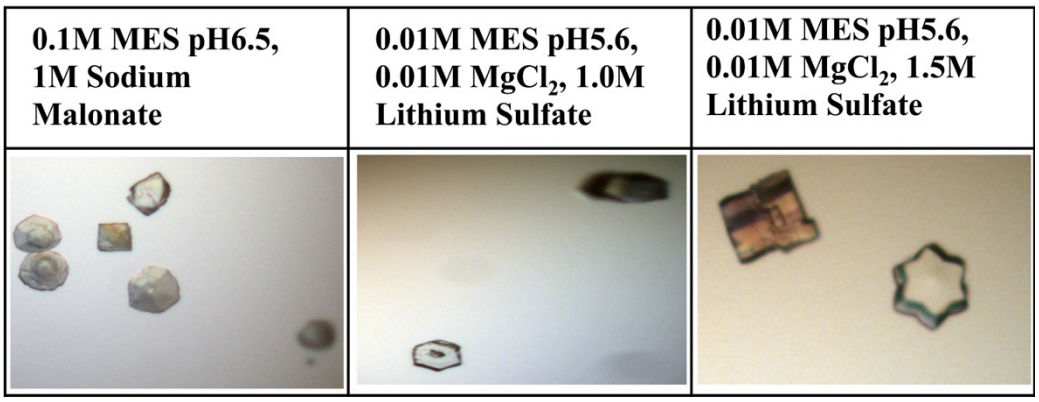


Figure 98**: Crystals of A2 CTD obtained at 20°C**

Two datasets were collected on crystals obtained in 0.1M MES pH 6.5, 1M Sodium Malonate at a resolution of 2.1Å. Diffraction data were collected using a Pilatus 2 M detector on beamline X06DA (PXIII) at the Swiss Light Source, Paul Scherrer Institut, Villigen, Switzerland

Data Processing and Refinement Statistics for each A2 dataset

	Dataset 1	Dataset 2
Wavelength		
Resolution range	42.89 - 1.702 (1.762 - 1.702)	29.83 - 1.8 (1.865 - 1.8)
Space group	P 6_3	P 6_3
Unit cell	49.523 49.523 42.011 90 90 120	49.4562 49.4562 41.582 90 90 120
Total reflections	131436 (13563)	109074 (10576)
Unique reflections	6521 (652)	5451 (534)
Multiplicity	20.2 (20.8)	20.0 (19.8)
Completeness (%)	99.80 (100.00)	99.82 (99.63)
Mean I/sigma(I)	29.62 (3.05)	18.57 (2.05)
Wilson B-factor	31.05	33.51
R-merge	0.05045 (1.229)	0.08596 (2.15)
R-meas	0.05181 (1.26)	0.08826 (2.205)
R-pim	0.01165 (0.2744)	0.01982 (0.4877)
CC1/2	1 (0.551)	1 (0.525)
CC*	1 (0.843)	1 (0.83)
Reflections used in refinement	6508 (652)	5442 (532)
Reflections used for R-free	293 (25)	269 (22)
R-work	0.2592 (0.5692)	0.2427 (0.5254)
R-free	0.3185 (0.6115)	0.3016 (0.4564)
CC(work)	0.957 (0.513)	0.704 (0.512)
CC(free)	0.818 (0.325)	0.690 (0.540)
Number of non-hydrogen atoms	509	501
macromolecules	475	475
ligands	1	1
solvent	33	25
Protein residues	56	56
RMS(bonds)	0.014	0.015
RMS(angles)	1.80	1.73
Ramachandran favored (%)	92.59	94.44
Ramachandran allowed (%)	5.56	5.56
Ramachandran outliers (%)	1.85	0.00
Rotamer outliers (%)	4.00	8.00
Clashscore	6.28	3.14
Average B-factor	55.03	52.87
macromolecules	55.03	52.90
ligands	102.09	95.33
solvent	53.53	50.64

Table 15: Refinement statistics for each A2 dataset. Statistics for the highest-resolution shell are shown in parentheses

Structure determination was carried out using the CCP4 suite of programs (Potterton, Briggs et al. 2003). The models were built using the program Coot (Emsley, Lohkamp et al. 2010) and structures were determined by molecular replacement using MOLREP (Vagin and

Teplyakov 2010) . Structures were generated using PYMOL (DeLano 2002) and Chimera (Pettersen, Goddard et al. 2004). Dimer interfaces were generated using PISA (Krissinel and Henrick 2007).

Crystal Structure of A2 – Dataset 1

The structure of A2 IN-CTD at pH 6.5 was solved at a resolution of 1.7Å. Consistent with our other structure (Figure 81) it consists of 5 β strands that form 2 anti-parallel β sheets connected by loops, indicating that the overall architecture is conserved between subtype A2 and B.

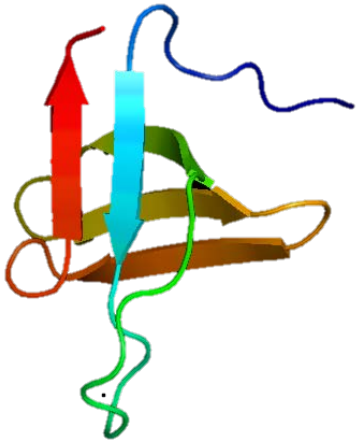


Figure 99: Cartoon representation of A2 IN-CTD monomer

Like the HIS-IN-CTD 220-270 described in the previous chapter (Figure 83), we found three possible surface contacts for dimer formation, involving similar residues in the protein. The first dimeric interface involves residues R231 to I234 on one monomer and residues Y226, W235 to P238, I268 to D270 on the second monomer. An electrostatic representation shows that this dimer is driven by charged based and hydrophobic interactions. The mutation from V234 to I234 did not change the interface area as the dimeric surface area is and 216A^2, consistent with the previous structure. The solvation energy of this dimer is -3.3kcal/mol, indicative of hydrophobic interfaces, with a p value of 0.15, implying that this interaction is specific, and not an artifact of crystal packing. Additionally, there are 2 hydrogen bonds present.

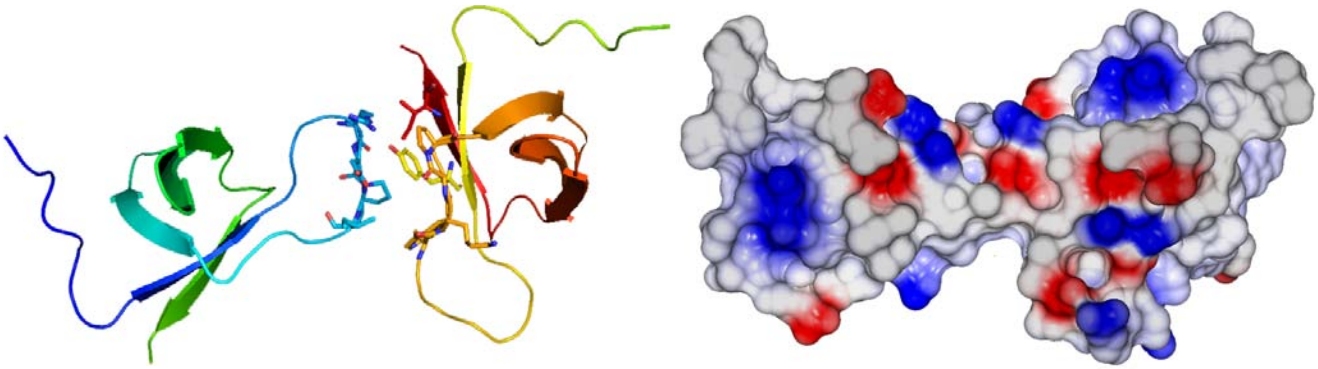


Figure 100: A2 IN-CTD Interface 1 .Left: Cartoon representation of the first dimeric interface, highlighting the side chains of residues R231 to I234 on the left and Y226, W235 to P238, I268 to D270 on the right. Right: Electrostatic representation of dimeric interface.

The second interface involves interactions between residues P238, N254 and K240 on one monomer and R263 and K264 on the second monomer. This interface is primarily mediated by charge-based interactions. It is noteworthy that two of the residues (K240Q and N254K) involved in this dimer interaction are in the $N_{222}K_{240}N_{254}K_{273}$ motif, which was suggested by our collaborators to be important for integration assays. The interface area for this dimeric surface is 127A^2, with solvation energy of 2.47kcal/mol, confirming that it is not a hydrophobic interface. There are 2 hydrogen bonds present. The solvation energy has a p value of 0.79, could be indicative of an artifact of crystal packing.

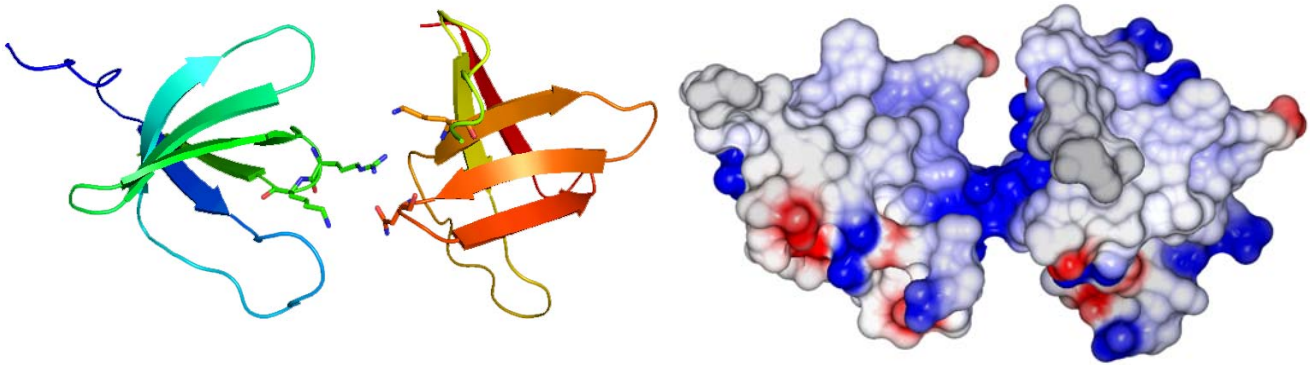


Figure 101: A2 IN-CTD Interface 2 . Left: Cartoon representation of the second dimeric interface, highlighting the side chains of residues 263R and 264K on the left, N254 and K240 on the right. Right: Electrostatic representation of dimeric interface

The third dimeric interface involves interactions between R269 on one monomer and S255 and D256 on the second monomer. The interface area for this dimeric surface is 46A^2, with 1

hydrogen bond. With solvation energy of 0.36 kcal/mol, charged residues mediate this interface. The p value of the solvation energy is 0.59, which could imply that this interface is an artifact.

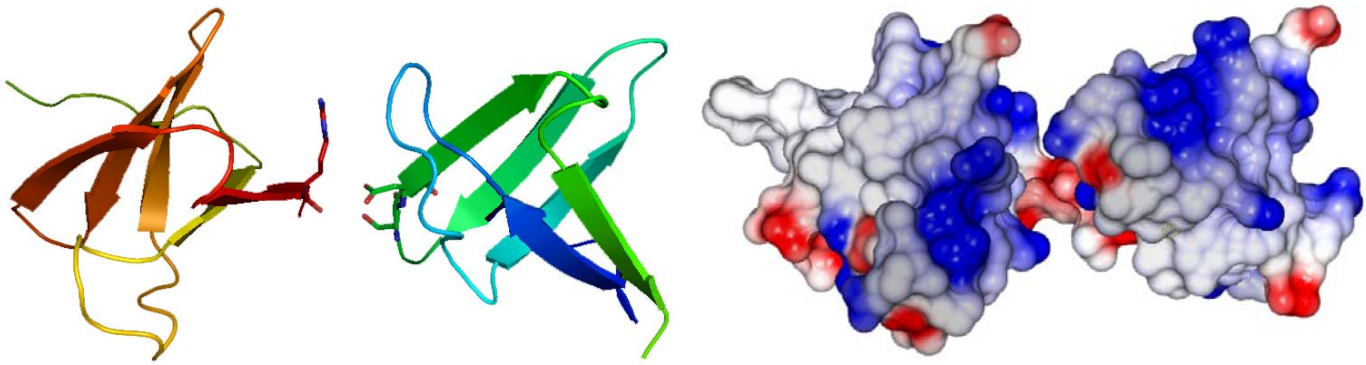


Figure 102: A2 IN-CTD Interface 3. Left: Cartoon representation of the second dimeric interface, highlighting the side chains of residues R269 on the left, S255 and D256 on the right. Right: Electrostatic representation of dimeric interface

N222K/K240Q

Manual hanging drops of the N222K/K240Q CTD were set up by mixing 1µl protein at 5mg/ml+ 1µl reservoir and equilibrating against 500µl of reservoir at 20°C. No crystals were obtained in any conditions.

Several commercial screens were tested, including ANIONS, CATIONS, WIZARDS, CLASSICS, JSCG+, PEGS and INDEX using the mosquito robot with protein at 5.6 mg/ml Additionally, no crystals were obtained in screens set up using the mosquito robot.

K240Q/N254K

Manual hanging drops of the K240Q/N254K CTD were set up by mixing 2µl protein at 5mg/ml+ 2µl reservoir and equilibrating against 500µl of reservoir at 20°C. Crystals were obtained in two conditions. Crystals obtained in 0.1M MES pH 6.5, 1M Sodium Malonate diffracted in a resolution range from 2Å to 4Å. However, most of the data sets were of poor quality due to twinning.

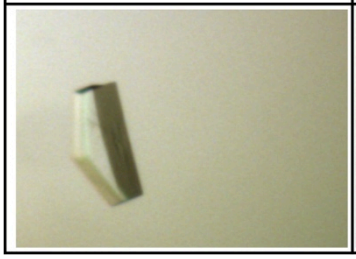
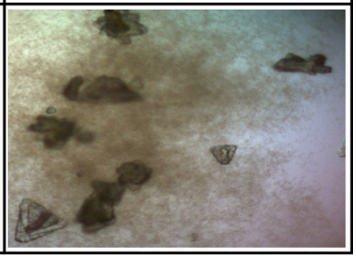
0.1M MES pH6.5, 1M Sodium Malonate	0.01M MES pH5.6, 0.01M $MgCl_2$, 1.5M Lithium Sulfate
	

Figure 103: Crystals of K240Q/N254K CTD obtained at 20°C

New screens were set up reproduce more crystals by manual hanging drops with protein at 4.5mg/ml. Reservoir conditions included 0.1M MES pH6.5, 0.8M Sodium Malonate, 0.1M MES pH6.5, 1M Sodium Malonate and 0.1M HEPES pH 7.5, 1.5M Ammonium Sulfate. Data was collected at the SLS synchrotron and crystals diffracted to 2Å. Additionally, several crystals were obtained in screens set up using the mosquito robot, with protein concentration of 4mg/ml.

Data Processing and Refinement Statistics for K240Q/N254K dataset at pH 7.5

	K240Q/N254K
Wavelength	1.0000
Resolution range	39.84 - 2.0 (2.072 - 2.0)
Space group	C 1 2 1
Unit cell	84.8868 48.796 66.518 90 106.615 90
Total reflections	117731 (11249)
Unique reflections	17801 (1768)
Multiplicity	6.6 (6.4)
Completeness (%)	99.94 (99.83)
Mean I/sigma(I)	7.03 (0.97)
Wilson B-factor	42.16
R-merge	0.1363 (1.597)
R-meas	0.1479 (1.739)
R-pim	0.05686 (0.6797)
CC1/2	0.994 (0.377)
CC*	0.999 (0.74)
Reflections used in refinement	17812 (1768)
Reflections used for R-free	885 (89)
R-work	0.2350 (0.3568)
R-free	0.2916 (0.3988)
CC(work)	0.923 (0.442)
CC(free)	0.907 (0.379)
Number of non-hydrogen atoms	1566
macromolecules	1428
ligands	1
solvent	137
Protein residues	168
RMS(bonds)	0.015
RMS(angles)	1.69
Ramachandran favored (%)	94.44
Ramachandran allowed (%)	4.94
Ramachandran outliers (%)	0.62
Rotamer outliers (%)	6.00
Clashscore	5.92
Average B-factor	52.33
macromolecules	52.12
ligands	31.14
solvent	54.73

Table 16: Refinement statistics for data collected at pH 7.5_. Statistics for the highest-resolution shell are shown in parentheses

Crystal Structure of K240Q/N254K

Unlike the subtype A2/B WT IN-CTD structure that had 3 similar dimeric interfaces, the structure of this mutant displays only 2 interfaces. The first interface is similar to the A2 WT interface; involving residues R231 to I234 on one monomer and residues Y226, W235 to P238, I268 to D270 on the second monomer. The surface area for this interface is 242.8A^2 with 1 hydrogen bond. Consistent with other structures, the solvation energy is -3.9kcal/mol/mol, as the interface is driven by hydrophobic interactions. As expected, the p value of 0.16 indicates that this interface is not a crystal artifact.

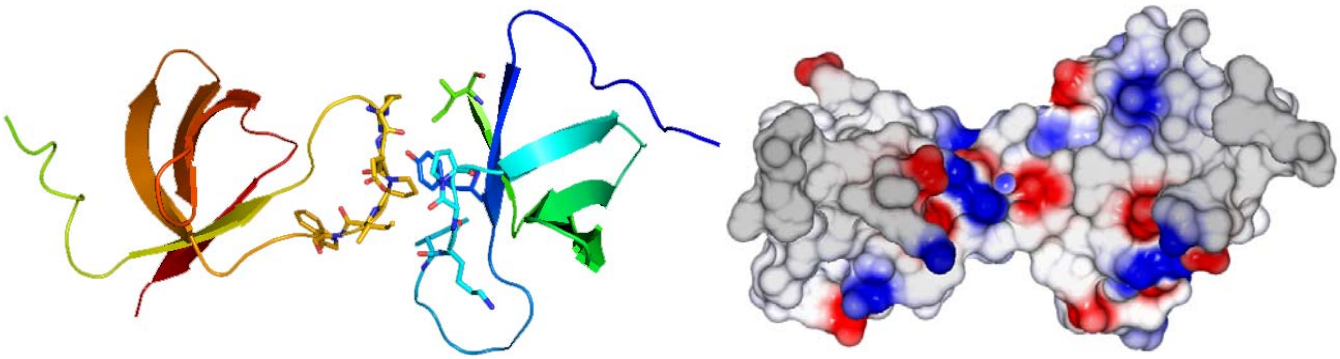


Figure 104: K240Q/N254K Interface 1. Left: Cartoon representation of the first dimeric interface, highlighting the side chains of residues R231 to I234 on the left and Y226, W235 to P238, I268 to D270 on the right. Right: Electrostatic representation of dimeric interface.

The second dimeric interface observed in this IN-CTD was completely novel and not previously observed in the structures of the subtype A2/B IN-CTD. Interestingly, this interface involves the same residue contacts (E246, G247, R263 to I267), on both monomers. The surface area for this interface is 238.7A^2 with solvation energy of 1.64kcal/mol as charged residues drive the interface interactions. Additionally, there are 2 hydrogen bonds and 2 salt bridges at the interface. The p-value for the solvation energy is 0.7, which could indicate that this interaction is a crystal artifact.

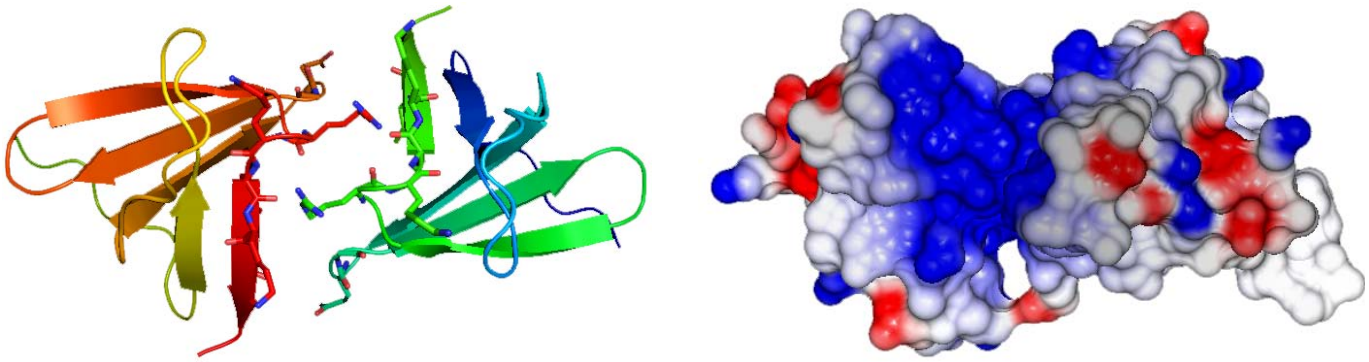


Figure 105: K240Q/N254K Interface 2 . Left: Cartoon representation of the first dimeric interface, highlighting the side chains of residues E246, G247, R263 to I267 on both monomers. Right: Electrostatic representation of dimeric interface.

O

Manual hanging drops of the K240Q/N254K CTD were set up by mixing 2µl protein at 5mg/ml+ 2µl reservoir and equilibrating against 500µl of reservoir at 20°C. No crystals were obtained in any of the crystallization conditions that were successful for A2 and K240Q/N254K. However, new crystallization conditions were discovered from the screens using the mosquito robot. These conditions were optimized by manual hanging drops, mixing 2µl protein at 5mg/ml+ 2µl reservoir and equilibrating against 500µl of reservoir at 20°C.

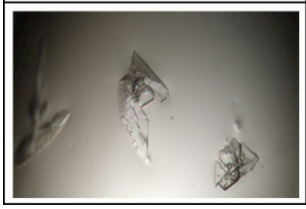
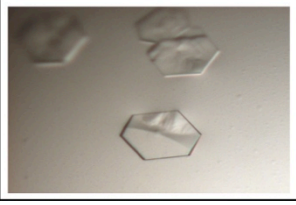
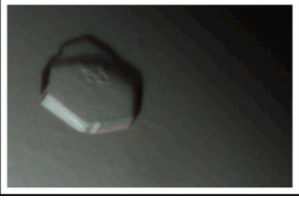
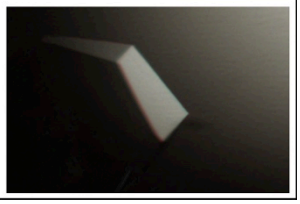
0.1M sodium acetate pH4.5 25% PEG 8000 0.2M LiSO4	0.1M sodium acetate pH4.5 28% PEG 8000 0.2M LiSO4	0.1M tri-Sodium Citrate pH5.5 2.2M AmSO4	0.1M tri-Sodium Citrate pH5.5 2M AmSO4
			

Figure 106: Crystals of O CTD obtained at 20°C

O CTD crystallizes preferentially in PEG 8000, and high Ammonium Sulfate concentrations. Crystals obtained in Ammonium Sulfate don't diffract. Crystals obtained in 28% PEG8000 diffract weakly to a maximum of 4Å. At 25% PEG 8000, they diffract slightly better to 3Å.

Discussions

The main aim of this project was to understand the structural differences that could account for differences in IN activity in subtype A2/group O chimeric constructs produced by our collaborators (Matteo Negroni, IBMC). The efforts to purify and crystallize the CTD of these proteins highlight the intrinsic flexibility of the IN-CTD.

A2

The sequence of the A2 CTD differs from pNL4.3 (HIS-IN-CTD 220-270) by one residue, where V234 is substituted for I234 in the A2 CTD. The pNL4.3 subtype is representative of subtype B. It was possible to obtain crystals in the same conditions as the pNL4.3 IN-CTD. These crystals diffracted to high resolution, and it was possible to solve the structure using molecular replacement. The crystal structure of IN-CTD of subtype A2 is similar to subtype B, as they display similar dimeric interface. 2 out of 3 interfaces had solvation energy with p values higher than 0.5, which is usually indicative of artfefacts. However, the presence of these interfaces in subtype B suggests that they may be significant. Moreover, interface 2 (Figure 101) involves residues in the motif of interest, suggesting that this interface is biologically relevant.

N222K/K240Q

Crystals could not be obtained for the N222K/K240Q mutant. This is surprising because it was as soluble as WT during purification. From the A2 WT structure, it was observed that K240Q is involved in dimer formation. One of the factors responsible for crystal packing is the interaction of the HIS-tag with the Ni2+. N222 is located on the same loop as the HIS-tag on the N-terminus. It is possible that the N222K mutation inhibits this interaction of the loop with the Ni2+ ion, thereby reducing the chances of crystal formation.

K240Q/N254K

Similar to the A2 WT protein, K240Q/N254K crystallizes in similar conditions as HIS-IN-CTD 220-270 (subtype B), but in a different space group (C2). During purification, K240Q/N254K was less soluble and stable in solution, and more prone to aggregation and precipitation. It is noteworthy that despite reduced protein solubility and stability, crystals diffract to a resolution of 2Å. Interestingly, the crystal structure of this IN-CTD display 2

potential dimeric interfaces, as opposed to 3 observed in the A2 WT. Additionally, the interface 2 involving K240 and N254 (Figure 101) observed in the A2 WT is not observed in this mutant. Instead, a new interface involving a new set of residues (Figure 105) was observed. It is possible that this new dimer compensates for the loss of K240 and N25K. However, with the high p value of this interface, it is difficult to make any conclusions. With the reduced stability of this mutant observed during purification, we hypothesize that the K240Q/N25K mutation explains increased aggregation and precipitation observed during purification, as mutating the residues could lead to the formation of weaker dimers and less stable protein.

O WT

The O WT IN-CTD protein differs from the A2 by 4 residues: N222 is substituted for K222; K240 is substituted for Q240, N254 and S255 are substituted for K254 and G255 respectively. 4 of out 5 residues are involved in dimer formation. These substitutions make the O WT IN-CTD more soluble and stable during purification. We hypothesize that these series of mutations have been systematically designed by the virus in order to improve IN oligomerization and stability.

Due to the high flexibility of the IN-CTD, these 5 differences greatly influence the dynamics, and probably structure of the protein, as the crystallization conditions are completely different from the other proteins. While A2 preferentially crystallize in high concentrations of salt like sodium malonate or lithium sulfate, crystallization drops of O CTD are completely clear under these conditions. The O CTD crystallizes in reservoirs with high Ammonium Sulfate or high PEGS. Surprisingly, O CTD crystals that grow as large hexagonal plates either diffract poorly or not at all.

Structure-Function Analysis

Our structure-function analysis suggests that the predicted dimers are functional, as residues in this interface seem to play an important role in IN functionality.

When compared to A2 WT, N222K/K240Q was shown to be as active as WT during integration assays by our collaborators. We have shown that K240 is important for dimer formation. A possible explanation for this activity level is that N222K is a compensatory mutation, allowing integrase to maintain its function, even with the destabilizing mutation at

K240. N222 has been shown to be in a loop. We hypothesize that substituting a polar residue (N) with a basic residue (K) form new charge based interactions within the protein, possible stabilizing the loop, and allowing integrase to retain its function

K240Q/N254K was shown to retain only 15% of its activity. This is plausible since K240 and N254K are important for dimer formation. It has been established that the formation of higher order oligomers are necessary for integrase function. With these 2 mutations and no compensatory mutation at N222K, we hypothesize that integrase loses its ability to form stable oligomers, thereby losing its functionality.

Preliminary experiments by our group and our collaborators suggest that the full length O integrase is more active than the full length A2 integrase in 3' processing reactions. These results are in agreement with the higher stability and solubility observed during purification. This data suggests that the virus, probably in response to environmental pressure, systemically designed these IN mutations to maintain infectivity. In light of this, it is interesting that group O viruses have displayed less fitness that group M viruses (Ariën, Abraha et al. 2005).

Conclusions/Perspectives

Put together, our results suggest that the $N_{222}K_{240}N_{254}K_{273}$ in A2 IN-CTD motif plays a role in the stability of the IN-CTD, and the mutation to $K_{222}Q_{240}K_{254}Q_{273}$ in the O-CTD helps retain infectivity. Unfortunately, the K273 residue is out of the limits of our construct, so we cannot comment on this residue. However, we show that K240 and N254 are important for oligomer formation.

Further data is needed to support our hypothesis. We need to produce better diffracting crystals of the O-IN-CTD. Additionally, we need to obtain crystals of the N222K/K240Q to explain the activity levels observed. A possible approach to obtaining these crystals is with seeding with A2 WT crystals.

The progress made in the process of this study raises new insights into the importance of the IN-CTD for HIV integration. The methods developed using NMR and X-ray crystallography would be used to screen HIV inhibitors. Particularly, our results could be used as the basis for high throughput screening for conformational inhibitors in vitro and in silico.

Appendix

IN-CTD Protein Sequences (pET Vectors)

HIS-CTD 220-270

M G H H H H H H I Q N F R V Y Y R D S R D P V W K G P A K L L W K G E G A V V I Q D N S D I K V V P R R K A K I I R D

HIS-CTD 220-288

M G H H H H H H I Q N F R V Y Y R D S R D P V W K G P A K L L W K G E G A V V I Q D N S D I K V V P R R K A K I I R D Y G K Q M A G D D C V A S R Q D E D

HIS-P3C- CTD 220-288

M G H H H H H H L E V L F Q G P I Q N F R V Y Y R D S R D P V W K G P A K L L W K G E G A V V I Q D N S D I K V V P R R K A K I I R D Y G K Q M A G D D C V A S R Q D E D

HIS-P3C-CTD 220-270

M G H H H H H H L E V L F Q G P I Q N F R V Y Y R D S R D P V W K G P A K L L W K G E G A V V I Q D N S D I K V V P R R K A K I I R D

IN-CTD Protein Sequences (Gateway Vectors)

HIS-P3C-L-CTD 221-288

M G S S H H H H H H G T G S Y I T S L Y K K A G F L E V L F Q G P M Q N F R V Y Y R D S R D P V W K G P A K L L W K G E G A V V I Q D N S D I K V V P R R K A K I I R D Y G K Q M A G D D C V A S R Q D E D

GST-P3C-L-CTD 221-288

MSPILGYWKI KGLVQPTRLL LEYLEEKYEE HLYERDEGDK WRNKKFELGL EFPNLPYYID GDVKLTQSMA IIRYIADKHN MLGGCPKERA EISMLEGAVL DIRYGVSRIA YSKDFETLKV DFLSKLPEML KMFEDRLCHK TYLNGDHVTH PDFMLYDALD VVLYMDPMCL DAFPKLVCFK KRIEAIPQID KYLKSSKYIA WPLQGWQATF GGGDHPPKSD LVPRPWSNQT SLYKKAGFLE VLFQGPMQNF RVYYRDSRDP VWKGPAKLLW KGEGAVVIQD NSDIKVVPRR KAKIIRDYGK QMAGDDCVAS RQDED

STREP-P3C-L-CTD 221-288

MASWSHPQFE KGAVTSLYKK AGFLEVLFQG PMQNFRVYYR DSRDPVWKGP AKLLWKGEGA VVIQDNSDIK VVPRRKAKII RDYGKQMAGD DCVASRQDED

LB Composition

Tryptone	10g
Yeast Extract	5g
NaCl	10g

Add distilled water up to 1L, and autoclave at 121°C for 2 hours

LB-Agar Composition

Tryptone	10g
Yeast Extract	5g
NaCl	10g
Agar	15g

Minimal Medium Composition

Prepare 10X M9 medium

Na_2HPO_4 X $2H_2O$	76.75g	430mM
KH_2PO4	30g	220mM
NaCl	5g	86mM
$^{15}NH_4Cl$	5g	94mM
Adjust pH to 7.4 with 10M NaOH		

Autoclave and in a cool dark place

Prepare 100X Trace element solution
For 1L solution

$ZnSO_4$ x $7H_2O$	167mg	582µM
$FeCl_3$ x $6H_2O$	833mg	3mM
$CuCl_2$ x $2H_2O$	13mg	76.5 µM
$CoCl_2$ x $6H_2O$	10mg	42 µM
H_3BO_3	10mg	164 µM
$MnCl_2$ x $6H_2O$	1.6g	4.3mM

Filter and store at -20°C

20% ^{13}C Glucose
1M $MgSO_4$ (autoclaved)
1M $CaCl_2$ (autoclaved)
10mg/ml Biotin (filtered and stored at -20°C)
10mg/m Thiamin (filtered and stored at -20°C)

To prepare 1L of minimal medium on the day of culture, mix:

100ml	10X M9 medium
10ml	100X trace elements solution
10ml	20% 13C Glucose
1ml	1M $MgSO_4$
0.3ml	1M $CaCl_2$
0.1ml	10mg/ml Biotin
0.1ml	10mg/ml Thiamin

Adjust volume with autoclaved MilliQ water

Sodium Dodecyl Sulfate Polyacrylamide (SDS-PAGE) Electrophoresis

Resolving gel composition (2 gels)

	10%	12.5%	15%
Distilled Water	6ml	5.2ml	4.4ml
1.5M Tris HCl pH 8.8, 0.4% SDS	3.15ml	3.15ml	3.15ml
40% Acrylamide :bisacrylamide 29:1	3.1ml	3.9ml	4.7ml
10% Ammonium persulfate	125µl	125µl	125µl
TEMED	12.5µl	12.5µl	12.5µl

Stacking gel composition (2 gels)

Distilled Water	3.8ml
0.5M Tris HCl pH 6.8, 0.4% SDS	1.6ml
40% Acrylamide :bisacrylamide 29:1	0.8ml
10% Ammonium persulfate	60µl
TEMED	6µl

Résumé en Français

Introduction

Les rétrovirus possèdent deux enzymes clefs permettant la réplication de leur génome : la transcriptase inverse virale (RT) et l'intégrase (IN). Peu après l'infection d'une cellule cible par le virus, au cours de la phase précoce du cycle de réplication viral, l'ARN est rétrotranscrit par la RT pour générer un ADN viral double-brin (DNAv). Ce DNAv interagit avec des protéines virales et cellulaires pour former le complexe de pré-intégration (PIC). Ce PIC transite ensuite dans le cytoplasme le long des microtubules, jusqu'à la membrane nucléaire au niveau de laquelle, à travers le pore nucléaire, il rejoindra le noyau de la cellule. Le PIC du virus de l'immunodéficience humaine de type 1 (VIH-1) est constitué de l'ADN viral, de protéines de la cellule hôte et de protéines virales. IN est un élément central et permanent du PIC et est impliquée dans de nombreuses étapes du cycle de réplication virale, comme la transcription inverse, l'import nucléaire, le ciblage à la chromatine ainsi que l'intégration de l'ADNv.

IN est une ADN recombinase qui catalyse deux réactions endonucléolytiques permettant l'intégration de l'ADNv dans l'ADN de la cellule hôte et par conséquent la réplication virale et la production de nouvelles particules infectieuses. La première réaction est une maturation des extrémités 3' de l'ADNv (*3' processing*) exposant un groupement 3'OH libre à chaque extrémité. La deuxième réaction est un transfert de brins au cours de laquelle IN clive les deux brins de l'ADN cible pour y abouter les deux extrémités de l'ADNv précédemment modifié. Par la suite, les facteurs de réparation de l'ADN de la cellule cible combleront les lacunes restantes complétant ainsi le processus d'intégration. IN est formée de trois domaines structuraux et fonctionnels : le domaine N-terminal (NTD) qui forme un motif en doigt de zinc, le domaine catalytique central (CCD) qui contient le site actif et le domaine C-terminal (CTD) qui possède la capacité de lier l'ADN de manière non spécifique. Une revue récente met en lumière l'importance de ce CTD pour la liaison d'autres protéines virales telles que la RT. Plus récemment, une étude en cryo-microscopie électronique a permis de résoudre la structure de l'intasome de HIV-1.

L'intégration de l'ADNv dans le génome de la cellule hôte requiert un ciblage vers des zones spécifiques de la chromatine. Des facteurs cellulaires tels que BET ou LEDGF/p75 participent à ces évènements par leur faculté à reconnaitre et lier des histones portant des modifications précises (Kvaratskhelia et al., 2014). Le mécanisme précis d'association au nucléosome et d'intégration à la chromatine reste en revanche très mal connu. Le CTD de IN possède un domaine de type SH3 formé d'une soixantaine d'acides aminés formant un tonneau β. Ce motif SH3 rapproche le CTD de la superfamille de protéines à motif *Royal Domain*, connues pour reconnaître des queues d'histones

portant des modifications post-traductionnelles via une cage aromatique située au sein de ce tonneau β, qui constitue le site de liaison. Le CTD partage de nombreuses similarités structurales avec les domaines Tudor, mais est dépourvu de cage aromatique. Toutefois, le travail de collaborateurs (Vincent Parissi, Bordeaux) a pu montrer que l'extrémité N-terminale mono-méthylée de l'histone H4 (H4K20Me1) interagit de manière forte avec IN. Par une étude en *far-dot blot*, cette interaction a pu être localisée de manière spécifique au niveau du CTD.

Une caractéristique majeure du virus de l'immunodéficience humaine est sa forte variabilité, illustrant sa grande flexibilité génétique. Cette flexibilité, qui permet la variation antigénique, est centrale pour l'adaptation du virus face à la réponse immunitaire de l'hôte, ainsi que pour le développement de résistance aux traitements antiviraux. L'inconvénient de cette forte variabilité est la possibilité de perte de fonction au niveau des protéines ou des acides nucléiques nécessaires à l'infectivité virale. La façon dont est maintenue l'équilibre entre ces deux aspects est essentielle pour notre compréhension de l'évolution virale. A ce titre, l'étude des réseaux de coévolution permettant de conserver les fonctions des protéines virales en dépit des variations de séquences est de première importance. Ces études permettent en outre d'appréhender de nouveaux aspects de la biologie du VIH-1 et ainsi de proposer de nouvelles cibles thérapeutiques potentielles.

Des travaux menés par nos collaborateurs (Matteo Negroni, IBMC, Strasbourg) à ce sujet sur l'IN ont permis de démontrer l'importance du domaine CTD. En effet, lors de l'étude de mutations ou de substitutions de séquences entre différents groupes ou sous-types de VIH-1, des substitutions affectant l'activité de l'IN ont été localisées au niveau du CTD. Certaines mutations réduisent son activité, d'autres l'abolissent totalement. Nous nous sommes donc intéressés aux différences structurales entre les CTD de ces différents groupes et sous-types.

Objectifs de ces recherches

L'objectif global de ce projet de recherche est la compréhension des interactions médiées par le CTD en ayant recours à des techniques de résonnance magnétique nucléaire (RMN) et de cristallographie aux rayons X. La première partie est centrée sur la caractérisation des interactions entre le domaine CTD de l'IN et un peptide de H4K20Me1. La deuxième partie est consacrée à la compréhension des changements co-évolutifs au niveau du CTD de différents groupes et sous-types de VIH-1.

Méthodes

Clonage : Au cours de ce projet, deux stratégies de clonage ont été utilisées. La première basée sur la technologie Gateway® a été utilisée pour construire des vecteurs d'entrée servant ensuite à générer 3 vecteurs d'expression. Ceux-ci contiennent différentes étiquettes de purification ou solubilisation (6His, Strep, GST). La deuxième méthode est basée sur l'utilisation d'enzymes de restriction et d'amplification par PCR. Quatre constructions ont ainsi été conçues pour améliorer la solubilité du CTD et cette méthode a aussi permis de générer quatre vecteurs d'expression pour l'étude co-évolutive de ce domaine.

Vector d'expression	Etiquette	Clonage	Bornes	Groupe/sous-type
pHGWA	6X HIS-P3C	Gateway	220-288	
pDEST 15	GST-P3C	Gateway	220-288	
pET 15	6X HIS-P3C	Restriction	220-288	
pET 15	6X HIS-P3C	Restriction	220-270	
pET 15	6X HIS	Restriction	220-288	
pET 15	6X HIS	Restriction	220-270	
pET 15	6X HIS-P3C	Restriction	220-270	A2 Wild-type
pET 15	6X HIS	Restriction	220-270	A2 N22K/K240Q
pET 15	6X HIS	Restriction	220-270	A2 K240Q/N254K
pET 15	6X HIS	Restriction	220-270	O wild type

Table 1: Vecteurs générés et purifiés au cours de ce projet

Purification des protéines: De manière générale, les protéines recombinantes sont purifiées en deux étapes : Après lyse des cellules dans le tampon approprié (variant selon la protéine exprimée et son utilisation), la protéine est soumise à une première étape de chromatographie par affinité au nickel ou GST en fonction du tag choisi. Une deuxième étape de purification par chromatographie d'exclusion stérique est ensuite réalisée sur colonne Superdex 75 ou Superdex 200 en fonction du poids moléculaire de la protéine d'intérêt.

Synthèse de peptide : Des peptides correspondant à l'extrémité N-terminale de l'histone H4 ont été synthétisés sur la plateforme de synthèse peptidique de l'IGBMC. Ces peptides ont pour certains été modifiés par ajout de groupements méthyle sur le résidu correspondant à la lysine 20. Certains ont aussi été marqués à l'aide d'un fluorophore (fluorescéine).

Peptide	**Description**	**Sequence**
H4K20me0	Peptide non-fluorescent non-méthylé	RHRKVLR
H4K20me1	Peptide non-fluorescent mono-méthylé	RHRK(me1)VLR
Fluo-H4K20me0	Peptide fluorescent non-méthylé	Fluo-KGG-RHRKVLR
Fluo-H4K20me1	Peptide fluorescent mono-méthylé	Fluo-KGG-RHRK(me1)VLR
Fluo-H4K20me2	Peptide fluorescent di-méthylé	Fluo-KGG-RHRK(me2)VLR
Fluo-H4K20me3	Peptide fluorescent tri-méthylé	Fluo-KGG-RHRK(me3)VLR

Table 2: Peptides synthétisés au cours de ce projet

Etudes d'interaction : Afin de confirmer les interactions entre le CTD de IN et H4K20Me1, différentes expériences ont été réalisées. Une étude de thermophorèse (*microscale thermophoresis*, MST) a permis de déterminer la constante de dissociation entre les deux partenaires. Des expériences de compétition ont permis de montrer la spécificité de cette interaction entre CTD et le peptide mono-méthylé H4K20Me1. Cette interaction a pu être confirmée par des études en RMN (waterLOGSY, trNOESY etc…).

Etudes par RMN : Différentes techniques de RMN ont été mises en œuvre pour comprendre en détails les interactions en solution entre CTD et H4K20Me1. La forme courte du CTD (6His-220-270), marquée aux isotopes ^{15}N et ^{13}C, a été produite et purifiée. Des expériences 2D et 3D ont été conduites afin de réaliser les attributions et d'identifier chaque résidu dans la protéine. Par la suite des expériences de titration avec le peptide H4K20Me1 ont mis en évidence des décalages dans les spectres RMN. Les résidus correspondants à ces déplacements chimiques sont probablement directement impliqués dans la liaison au peptide ou subissent des changements de conformation à la suite de cette liaison.

Cristallographie aux rayons X : En plus des expériences de RMN, des études cristallographiques ont été entreprises pour étudier les interactions entre le CTD (6His-220-270) et le peptide H4K20Me1. Des criblages des conditions de cristallisation ont été réalisés pour la protéine seule et pour la protéine en présence du peptide. Dans le cadre des études de coévolution, différentes formes du CTD ont aussi pu être cristallisées (chimères A2/O).

Résultats

Etudes d'interaction : Les expériences de thermophorèse nous ont permis de montrer la liaison préférentielle du CTD au peptide fluorescent monométhylé H4K20Me1, avec un kD de 0.8 μM. Des constantes de dissociations plus faibles ont pu être déterminées pour les peptides H4K20Me2 et H4K20Me3 à hauteur de 4.3 μM and 5.2 μM respectivement. De plus, en présence d'un peptide compétiteur non fluorescent (H4K20Me0), l'affinité pour H4K20Me1 est conservée (1.17 μM), confirmant la spécificité de cette interaction.

Etudes par RMN : Dans les deux types d'expériences réalisées, un changement d'effet Overhauser nucléaire (NOE) du positif vers le négatif a été observé pour la protéine GST-CTD en présence du peptide H4K20Me1.

Le spectre ^{15}N HSQC nous a appris que les résidus A239 et G237 ne sont pas affectés par la liaison du peptide alors que d'autres résidus comme K266, Y226, K236, I268 ou V260 présentent une modification de déplacement chimique en présence du peptide. Par ailleurs, les pics correspondant aux résidus I251, L241 et K258 sont très fortement réduits. On note enfin la présence d'un nouveau pic non identifié à 7.7ppm/126.9ppm. Ces changements indiquent que ces résidus sont impliqués de manière directe dans la liaison au peptide ou tout du moins subissent des changements de conformation majeurs en présence de ce dernier. De plus, les différences observées pour la liaison au peptide entre les pH7 et pH8 suggèrent des changements de conformation de l'intégrase dépendants de son environnement.

Cristallographie aux rayons X : des structures du CTD de l'IN seule ont été résolues à 3 différents pH. La structure à pH 7 a révélé 3 interfaces formant des dimères du CTD révélant la flexibilité de ce domaine. De plus, les changements pH dépendants dans les boucles suggèrent que la structure du CTD est sensible aux changements de sont environnement.

Modèle basé sur les données structurale provenant de la cristallographie des rayons X et de la RMN:

A partir de la structure de la protéine seule et les informations obtenues par RMN, nous proposons que l'interaction IN-histone est médiée par une surface hydrophobe formé par les extrémités N-ter et C-ter du CTD. Nous émettons l'hypothèse que cette interaction est accompagnée par des mouvements dans les boucles et par la formation d'oligomères qui sont nécessaires pour l'intégration.

Nous avons résolu les structures cristallographiques du CTD de l'IN du sous-groupe A2 et d'une IN chimère. La structure de l'IN-CTD de A2 montre des interfaces dimériques similaires au CTD du sous-groupe B confirmant l'importance de ces interfaces pour les fonctions de l'intégrase. De manière intéressante dans le mutant K240Q/N254K, une interface différente est observée, qui représente probablement un mécanisme de compensation. Les résultats de cette étude permettent de mieux comprendre les processus d'adaptation et d'évolution virale pour l'IN-CTD

En conclusion, mes résultats mettent en évidence l'importance de l'IN-CTD en posant les bases structurales du rôle du CTD pour les interactions avec les histones et dans les processus d'évolution du virus.

Bibliography

Abecasis, A. B., A.-M. Vandamme and P. Lemey (2009). "Quantifying Differences in the Tempo of Human Immunodeficiency Virus Type 1 Subtype Evolution." Journal of Virology **83**(24): 12917-12924.

Aghokeng, A. F., E. Mpoudi-Ngole, H. Dimodi, A. Atem-Tambe, M. Tongo, C. Butel, E. Delaporte and M. Peeters (2009). "Inaccurate Diagnosis of HIV-1 Group M and O Is a Key Challenge for Ongoing Universal Access to Antiretroviral Treatment and HIV Prevention in Cameroon." PLOS ONE **4**(11): e7702.

AIDS.gov. (2015). "Stages of HIV Infection."

AIDSinfo. (2016). "The Stages of HIV Infection."

Archer, J., J. W. Pinney, J. Fan, E. Simon-Loriere, E. J. Arts, M. Negroni and D. L. Robertson (2008). "Identifying the Important HIV-1 Recombination Breakpoints." PLoS Computational Biology **4**(9): e1000178.

Ariën, K. K., A. Abraha, M. E. Quiñones-Mateu, L. Kestens, G. Vanham and E. J. Arts (2005). "The Replicative Fitness of Primary Human Immunodeficiency Virus Type 1 (HIV-1) Group M, HIV-1 Group O, and HIV-2 Isolates." Journal of Virology **79**(14): 8979-8990.

Arthur, L. O., J. W. Bess, R. C. Sowder, R. E. Benveniste, D. L. Mann, J. C. Chermann and L. E. Henderson (1992). "Cellular proteins bound to immunodeficiency viruses: implications for pathogenesis and vaccines." Science **258**(5090): 1935.

Asherie, N. (2004). "Protein crystallization and phase diagrams." Methods **34**(3): 266-272.

Baeten, J. M., B. Chohan, L. Lavreys, V. Chohan, R. S. McClelland, L. Certain, K. Mandaliya, W. Jaoko and O. Julie (2007). "HIV-1 Subtype D Infection Is Associated with Faster Disease Progression than Subtype A in Spite of Similar Plasma HIV-1 Loads." The Journal of Infectious Diseases **195**(8): 1177-1180.

Baid, R., A. K. Upadhyay, T. Shinohara and U. B. Kompella (2013). "Biosynthesis, Characterization, and Efficacy in Retinal Degenerative Diseases of Lens Epithelium-derived Growth Factor Fragment (LEDGF1–326), a Novel Therapeutic Protein." Journal of Biological Chemistry **288**(24): 17372-17383.

Barat, C., V. Lullien, O. Schatz, G. Keith, M. T. Nugeyre, F. Grüninger-Leitch, F. Barré-Sinoussi, S. F. LeGrice and J. L. Darlix (1989). "HIV-1 reverse transcriptase specifically interacts with the anticodon domain of its cognate primer tRNA." The EMBO Journal **8**(11): 3279-3285.

Barre-Sinoussi, F., J. C. Chermann, F. Rey, M. T. Nugeyre, S. Chamaret, J. Gruest, C. Dauguet, C. Axler-Blin, F. Vezinet-Brun, C. Rouzioux, W. Rozenbaum and L. Montagnier

(1983). "Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS)." Science **220**(4599): 868.

Benleulmi, M. S., J. Matysiak, D. R. Henriquez, C. Vaillant, P. Lesbats, C. Calmels, M. Naughtin, O. Leon, A. M. Skalka, M. Ruff, M. Lavigne, M. L. Andreola and V. Parissi (2015). "Intasome architecture and chromatin density modulate retroviral integration into nucleosome." Retrovirology **12**: 13.

Bieniasz, P. D., T. A. Grdina, H. P. Bogerd and B. R. Cullen (1998). "Recruitment of a protein complex containing Tat and cyclin T1 to TAR governs the species specificity of HIV-1 Tat." The EMBO Journal **17**(23): 7056-7065.

Bothner, A. A. and R. Gassend (1973). "BINDING OF SMALL MOLECULES TO PROTEINS *." Annals of the New York Academy of Sciences **222**(1): 668-676.

Boutwell, C. L., J. M. Carlson, T.-H. Lin, A. Seese, K. A. Power, J. Peng, Y. Tang, Z. L. Brumme, D. Heckerman, A. Schneidewind and T. M. Allen (2013). "Frequent and Variable Cytotoxic-T-Lymphocyte Escape-Associated Fitness Costs in the Human Immunodeficiency Virus Type 1 Subtype B Gag Proteins." Journal of Virology **87**(7): 3952-3965.

Boutwell, C. L., C. F. Rowley and M. Essex (2009). "Reduced Viral Replication Capacity of Human Immunodeficiency Virus Type 1 Subtype C Caused by Cytotoxic-T-Lymphocyte Escape Mutations in HLA-B57 Epitopes of Capsid Protein." Journal of Virology **83**(6): 2460-2468.

Briggs, J. A. G., J. D. Riches, B. Glass, V. Bartonova, G. Zanetti and H.-G. Kräusslich (2009). "Structure and assembly of immature HIV." Proceedings of the National Academy of Sciences **106**(27): 11090-11095.

Bukrinsky, M. I., N. Sharova, M. P. Dempsey, T. L. Stanwick, A. G. Bukrinskaya, S. Haggerty and M. Stevenson (1992). "Active nuclear import of human immunodeficiency virus type 1 preintegration complexes." Proceedings of the National Academy of Sciences of the United States of America **89**(14): 6580-6584.

Bushman, F. D., A. Engelman, I. Palmer, P. Wingfield and R. Craigie (1993). "Domains of the integrase protein of human immunodeficiency virus type 1 responsible for polynucleotidyl transfer and zinc binding." Proceedings of the National Academy of Sciences of the United States of America **90**(8): 3428-3432.

Bushman, F. D., T. Fujiwara and R. Craigie (1990). "Retroviral DNA integration directed by HIV integration protein in vitro." Science **249**(4976): 1555.

Busschots, K., J. Vercammen, S. Emiliani, R. Benarous, Y. Engelborghs, F. Christ and Z. Debyser (2005). "The interaction of LEDGF/p75 with integrase is lentivirus-specific and promotes DNA binding." J Biol Chem **280**.

Cai, M., R. Zheng, M. Caffrey, R. Craigie, G. M. Clore and A. M. Gronenborn (1997). "Solution structure of the N-terminal zinc binding domain of HIV-1 integrase." Nat Struct Mol Biol **4**(7): 567-577.

Cancio, R., S. Spadari and G. Maga (2004). "Vif is an auxiliary factor of the HIV-1 reverse transcriptase and facilitates abasic site bypass." Biochemical Journal **383**(3): 475.

Carlson, J. M., A. Q. Le, A. Shahid and Z. L. Brumme (2015). "HIV-1 adaptation to HLA: a window into virus-host immune interactions." Trends in Microbiology **23**(4): 212-224.

Carlson, L.-A., J. A. G. Briggs, B. Glass, J. D. Riches, M. N. Simon, M. C. Johnson, B. Müller, K. Grünewald and H.-G. Kräusslich (2008). "Three-Dimensional Analysis of Budding Sites and Released Virus Suggests a Revised Model for HIV-1 Morphogenesis." Cell Host & Microbe **4**(6): 592-599.

Cavanagh, J., W. J. Fairbrother, A. G. Palmer Iii, M. Rance and N. J. Skelton (2007). CHAPTER 7 - HETERONUCLEAR NMR EXPERIMENTS. Protein NMR Spectroscopy (Second Edition). Burlington, Academic Press**:** 533-678.

CDC. (2016). "Testing ".

Chayen, N. E. and E. Saridakis (2008). "Protein crystallization: from purified protein to diffraction-quality crystal." Nat Meth **5**(2): 147-153.

Checkley, M. A., B. G. Luttge and E. O. Freed (2011). "HIV-1 Envelope Glycoprotein Biosynthesis, Trafficking, and Incorporation." Journal of molecular biology **410**(4): 582-608.

Chen, A., I. T. Weber, R. W. Harrison and J. Leis (2006). "Identification of Amino Acids in HIV-1 and Avian Sarcoma Virus Integrase Subsites Required for Specific Recognition of the Long Terminal Repeat Ends." The Journal of biological chemistry **281**(7): 4173-4182.

Chen, C., T. J. Nott, J. Jin and T. Pawson (2011). "Deciphering arginine methylation: Tudor tells the tale." Nat Rev Mol Cell Biol **12**(10): 629-642.

Chen, J. C. H., J. Krucinski, L. J. W. Miercke, J. S. Finer-Moore, A. H. Tang, A. D. Leavitt and R. M. Stroud (2000). "Crystal structure of the HIV-1 integrase catalytic core and C-terminal domains: A model for viral DNA binding." Proceedings of the National Academy of Sciences of the United States of America **97**(15): 8233-8238.

Cherepanov, P. (2007). "LEDGF/p75 interacts with divergent lentiviral integrases and modulates their enzymatic activity in vitro." Nucleic Acids Research **35**(1): 113-124.

Cherepanov, P., A. L. B. Ambrosio, S. Rahman, T. Ellenberger and A. Engelman (2005). "Structural basis for the recognition between HIV-1 integrase and transcriptional coactivator p75." Proceedings of the National Academy of Sciences of the United States of America **102**(48): 17308-17313.

Cherepanov, P., E. Devroe, P. A. Silver and A. Engelman (2004). "Identification of an Evolutionarily Conserved Domain in Human Lens Epithelium-derived Growth Factor/Transcriptional Co-activator p75 (LEDGF/p75) That Binds HIV-1 Integrase." Journal of Biological Chemistry **279**(47): 48883-48892.

Cherepanov, P., G. Maertens, P. Proost, B. Devreese, J. Van Beeumen, Y. Engelborghs, E. De Clercq and Z. Debyser (2003). "HIV-1 integrase forms stable tetramers and associates with LEDGF/p75 protein in human cells." J Biol Chem **278**.

Cherepanov, P., Z.-Y. J. Sun, S. Rahman, G. Maertens, G. Wagner and A. Engelman (2005). "Solution structure of the HIV-1 integrase-binding domain in LEDGF/p75." Nat Struct Mol Biol **12**(6): 526-532.

Cheung, M.-S., M. L. Maguire, T. J. Stevens and R. W. Broadhurst (2010). "DANGLE: A Bayesian inferential method for predicting protein backbone dihedral angles and secondary structure." Journal of Magnetic Resonance **202**(2): 223-233.

Chow, S. A., K. A. Vincent, V. Ellison and P. O. Brown (1992). "Reversal of integration and DNA splicing mediated by integrase of human immunodeficiency virus." Science **255**(5045): 723.

Cihlar, T. and M. Fordyce (2016). "Current status and prospects of HIV treatment." Curr Opin Virol **18**: 50-56.

Ciuffi, A., M. Llano, E. Poeschla, C. Hoffmann, J. Leipzig, P. Shinn, J. R. Ecker and F. Bushman (2005). "A role for LEDGF/p75 in targeting HIV DNA integration." Nat Med **11**.

Clapham, P. R. and Á. McKnight (2002). "Cell surface receptors, virus entry and tropism of primate lentiviruses." Journal of General Virology **83**(8): 1809-1829.

Coffin, J. and R. Swanstrom (2013). "HIV Pathogenesis: Dynamics and Genetics of Viral Populations and Infected Cells." Cold Spring Harbor Perspectives in Medicine **3**(1): a012526.

Cohen, M. S., G. M. Shaw, A. J. McMichael and B. F. Haynes (2011). "Acute HIV-1 Infection." The New England journal of medicine **364**(20): 1943-1954.

Cong, M.-e., W. Heneine and J. G. García-Lerma (2007). "The Fitness Cost of Mutations Associated with Human Immunodeficiency Virus Type 1 Drug Resistance Is Modulated by Mutational Interactions." Journal of Virology **81**(6): 3037-3041.

Craigie, R. and F. D. Bushman (2012). "HIV DNA Integration." Cold Spring Harbor Perspectives in Medicine **2**(7).

Cuevas, J. M., R. Geller, R. Garijo, J. López-Aldeguer and R. Sanjuán (2015). "Extremely High Mutation Rate of HIV-1 In Vivo." PLOS Biology **13**(9): e1002251.

Dahabieh, M., E. Battivelli and E. Verdin (2015). "Understanding HIV Latency: The Road to an HIV Cure." Annual review of medicine **66**: 407-421.

Dalvit, C., G. Fogliatto, A. Stewart, M. Veronesi and B. Stockman (2001). "WaterLOGSY as a method for primary NMR screening: Practical aspects and range of applicability." Journal of Biomolecular NMR **21**(4): 349-359.

De Rijck, J., C. de Kogel, J. Demeulemeester, S. Vets, S. E. Ashkar, N. Malani, F. D. Bushman, B. Landuyt, S. J. Husson, K. Busschots, R. Gijsbers and Z. Debyser (2013). "The BET family of proteins targets Moloney Murine Leukemia Virus integration near transcription start sites." Cell reports **5**(4): 886-894.

DeLano, W. L. (2002). The PyMOL Molecular Graphics System. Palo Alto, CA, Delano Scientific LLC.

Di Primio, C., V. Quercioli, A. Allouch, R. Gijsbers, F. Christ, Z. Debyser, D. Arosio and A. Cereseto (2013). "Single-Cell Imaging of HIV-1 Provirus (SCIP)." Proceedings of the National Academy of Sciences **110**(14): 5636-5641.

Dominguez, C., R. Boelens and A. M. J. J. Bonvin (2003). "HADDOCK: A Protein−Protein Docking Approach Based on Biochemical or Biophysical Information." Journal of the American Chemical Society **125**(7): 1731-1737.

Dyda, F., A. B. Hickman, T. M. Jenkins, A. Engelman, R. Craigie and D. R. Davies (1994). "Crystal structure of the catalytic domain of HIV-1 integrase: similarity to other polynucleotidyl transferases." Science **266**(5193): 1981.

Eidahl, J. O., B. L. Crowe, J. A. North, C. J. McKee, N. Shkriabai, L. Feng, M. Plumb, R. L. Graham, R. J. Gorelick, S. Hess, M. G. Poirier, M. P. Foster and M. Kvaratskhelia (2013). "Structural basis for high-affinity binding of LEDGF PWWP to mononucleosomes." Nucleic Acids Research **41**(6): 3924-3936.

Eijkelenboom, A. P. A. M., R. A. Puras Lutzke, R. Boelens, R. H. A. Plasterk, R. Kaptein and K. Hard (1995). "The DNA-binding domain of HIV-1 integrase has an SH3-like fold." Nat Struct Mol Biol **2**(9): 807-810.

Emsley, P., B. Lohkamp, W. G. Scott and K. Cowtan (2010). "Features and development of Coot." Acta Crystallogr D Biol Crystallogr **66**.

Engelman, A., G. Englund, J. M. Orenstein, M. A. Martin and R. Craigie (1995). "Multiple effects of mutations in human immunodeficiency virus type 1 integrase on viral replication." Journal of Virology **69**(5): 2729-2736.

Engelman, A., J. J. Kessl and M. Kvaratskhelia (2013). "Allosteric inhibition of HIV-1 integrase activity." Current opinion in chemical biology **17**(3): 339-345.

Engelman, A., K. Mizuuchi and R. Craigie (1991). "HIV-1 DNA integration: Mechanism of viral DNA cleavage and DNA strand transfer." Cell **67**(6): 1211-1221.

Eron, J. J. J. (2000). "HIV-1 Protease Inhibitors." Clinical Infectious Diseases **30**(Supplement_2): S160-S170.

F.D.A. (2016). "Antiretroviral drugs used in the treatment of HIV infection."

Farmer, B. T., R. A. Venters, L. D. Spicer, M. G. Wittekind and L. Müller (1992). "A refocused and optimized HNCA: Increased sensitivity and resolution in large macromolecules." Journal of Biomolecular NMR **2**(2): 195-202.

Fassati, A. and S. P. Goff (2001). "Characterization of Intracellular Reverse Transcription Complexes of Human Immunodeficiency Virus Type 1." Journal of Virology **75**(8): 3626-3635.

Fischl, M. A., D. D. Richman, M. H. Grieco, M. S. Gottlieb, P. A. Volberding, O. L. Laskin, J. M. Leedom, J. E. Groopman, D. Mildvan, R. T. Schooley, G. G. Jackson, D. T. Durack and D. King (1987). "The Efficacy of Azidothymidine (AZT) in the Treatment of Patients with AIDS and AIDS-Related Complex." New England Journal of Medicine **317**(4): 185-191.

Forshey, B. M., U. von Schwedler, W. I. Sundquist and C. Aiken (2002). "Formation of a Human Immunodeficiency Virus Type 1 Core of Optimal Stability Is Crucial for Viral Replication." Journal of Virology **76**(11): 5667-5677.

Frankel, A. D. and J. A. T. Young (1998). "HIV-1: Fifteen Proteins and an RNA." Annual Review of Biochemistry **67**: 1-25.

Freed, E. O. (2015). "HIV-1 assembly, release and maturation." Nat Rev Micro **13**(8): 484-496.

Fujinaga, K., T. P. Cujec, J. Peng, J. Garriga, D. H. Price, X. Graña and B. M. Peterlin (1998). "The Ability of Positive Transcription Elongation Factor b To Transactivate Human Immunodeficiency Virus Transcription Depends on a Functional Kinase Domain, Cyclin T1, and Tat." Journal of Virology **72**(9): 7154-7159.

Ganser-Pornillos, B., M. Yeager and W. I. Sundquist (2008). "The Structural Biology of HIV Assembly." Current opinion in structural biology **18**(2): 203.

Gao, F., E. Bailes, D. L. Robertson, Y. Chen, C. M. Rodenburg, S. F. Michael, L. B. Cummins, L. O. Arthur, M. Peeters, G. M. Shaw, P. M. Sharp and B. H. Hahn (1999). "Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes." Nature **397**(6718): 436-441.

Greenberg, M. L. and N. Cammack (2004). "Resistance to enfuvirtide, the first HIV fusion inhibitor." J Antimicrob Chemother **54**(2): 333-340.

Greenfield, N. J. (2006). "Using circular dichroism spectra to estimate protein secondary structure." Nat Protoc **1**(6): 2876-2890.

Greenwald, J., V. Le, S. L. Butler, F. D. Bushman and S. Choe (1999). "The Mobility of an HIV-1 Integrase Active Site Loop Is Correlated with Catalytic Activity." Biochemistry **38**(28): 8892-8898.

Grzesiek, S. and A. Bax (1992). "Correlating backbone amide and side chain resonances in larger proteins by multiple relayed triple resonance NMR." Journal of the American Chemical Society **114**(16): 6291-6293.

Grzesiek, S. and A. Bax (1992). "An efficient experiment for sequential backbone assignment of medium-sized isotopically enriched proteins." Journal of Magnetic Resonance (1969) **99**(1): 201-207.

Gupta, S. S., T. Maetzig, G. N. Maertens, A. Sharif, M. Rothe, M. Weidner-Glunde, M. Galla, A. Schambach, P. Cherepanov and T. F. Schulz (2013). "Bromo- and Extraterminal Domain Chromatin Regulators Serve as Cofactors for Murine Leukemia Virus Integration." Journal of Virology **87**(23): 12721-12736.

Hahn, B. H., G. M. Shaw, K. M. De, Cock and P. M. Sharp (2000). "AIDS as a Zoonosis: Scientific and Public Health Implications." Science **287**(5453): 607.

Heinzinger, N. K., M. I. Bukinsky, S. A. Haggerty, A. M. Ragland, V. Kewalramani, M. A. Lee, H. E. Gendelman, L. Ratner, M. Stevenson and M. Emerman (1994). "The Vpr protein of human immunodeficiency virus type 1 influences nuclear localization of viral nucleic acids in nondividing host cells." Proceedings of the National Academy of Sciences **91**(15): 7311-7315.

Hemelaar, J. (2012). "The origin and diversity of the HIV-1 pandemic." Trends in Molecular Medicine **18**(3): 182-192.

Henrich, T. J. and D. R. Kuritzkes (2013). "HIV-1 entry inhibitors: recent development and clinical use." Curr Opin Virol **3**(1): 51-57.

Hicks, C. and R. M. Gulick (2009). "Raltegravir: The First HIV Type 1 Integrase Inhibitor." Clinical Infectious Diseases **48**(7): 931-939.

Higman, V. (2012). "Spectrum Descriptions."

Hu, Z. and D. R. Kuritzkes (2014). "Altered Viral Fitness and Drug Susceptibility in HIV-1 Carrying Mutations That Confer Resistance to Nonnucleoside Reverse Transcriptase and Integrase Strand Transfer Inhibitors." Journal of Virology **88**(16): 9268-9276.

Hulme, A. E., Z. Kelley, D. Foley and T. J. Hope (2015). "Complementary Assays Reveal a Low Level of CA Associated with Viral Complexes in the Nuclei of HIV-1-Infected Cells." Journal of Virology **89**(10): 5350-5361.

Invitrogen. (2003). "Gateway Technology."

Isel, C. and J. Karn (1999). "Direct evidence that HIV-1 tat stimulates RNA polymerase II carboxyl-terminal domain hyperphosphorylation during transcriptional elongation1." Journal of Molecular Biology **290**(5): 929-941.

Jain, C. and J. G. Belasco (2001). "Structural Model for the Cooperative Assembly of HIV-1 Rev Multimers on the RRE as Deduced from Analysis of Assembly-Defective Mutants." Molecular Cell **7**(3): 603-614.

Jenkins, T. M., D. Esposito, A. Engelman and R. Craigie (1997). "Critical contacts between HIV-1 integrase and viral DNA identified by structure-based analysis and photo-crosslinking." The EMBO Journal **16**(22): 6849-6859.

Jenkins, T. M., A. B. Hickman, F. Dyda, R. Ghirlando, D. R. Davies and R. Craigie (1995). "Catalytic domain of human immunodeficiency virus type 1 integrase: identification of a soluble mutant by systematic replacement of hydrophobic residues." Proc Natl Acad Sci USA **92**.

Jerabek-Willemsen, M., T. André, R. Wanner, H. M. Roth, S. Duhr, P. Baaske and D. Breitsprecher (2014). "MicroScale Thermophoresis: Interaction analysis and beyond." Journal of Molecular Structure **1077**: 101-113.

Jerabek-Willemsen, M., C. J. Wienken, D. Braun, P. Baaske and S. Duhr (2011). "Molecular Interaction Studies Using Microscale Thermophoresis." Assay and Drug Development Technologies **9**(4): 342-353.

Kay, L. E., M. Ikura, R. Tschudin and A. Bax (1990). "Three-dimensional triple-resonance NMR spectroscopy of isotopically enriched proteins." Journal of Magnetic Resonance (1969) **89**(3): 496-514.

Kelly, S. M., T. J. Jess and N. C. Price (2005). "How to study proteins by circular dichroism." Biochim Biophys Acta **1751**(2): 119-139.

Khurshid, S., E. Saridakis, L. Govada and N. E. Chayen (2014). "Porous nucleating agents for protein crystallization." Nat. Protocols **9**(7): 1621-1633.

Kim, D., B. J. Blus, V. Chandra, P. Huang, F. Rastinejad and S. Khorasanizadeh (2010). "Corecognition of DNA and a methylated histone tail by the MSL3 chromodomain." Nat Struct Mol Biol **17**(8): 1027-1029.

Kim, J., J. Daniel, A. Espejo, A. Lake, M. Krishna, L. Xia, Y. Zhang and M. T. Bedford (2006). "Tudor, MBT and chromo domains gauge the degree of lysine methylation." EMBO Reports **7**(4): 397-403.

Köhler, C., R. Recht, M. Quinternet, F. de Lamotte, M.-A. Delsuc and B. Kieffer (2015). Accurate Protein–Peptide Titration Experiments by Nuclear Magnetic Resonance Using Low-

Volume Samples. Affinity Chromatography: Methods and Protocols. S. Reichelt. New York, NY, Springer New York**:** 279-296.

Krishnan, L., X. Li, H. L. Naraharisetty, S. Hare, P. Cherepanov and A. Engelman (2010). "Structure-based modeling of the functional HIV-1 intasome and its inhibition." Proceedings of the National Academy of Sciences **107**(36): 15910-15915.

Krissinel, E. and K. Henrick (2007). "Inference of Macromolecular Assemblies from Crystalline State." Journal of Molecular Biology **372**(3): 774-797.

Kwong, P. D., R. Wyatt, J. Robinson, R. W. Sweet, J. Sodroski and W. A. Hendrickson (1998). "Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody." Nature **393**(6686): 648-659.

Le Noury, D. A., S. Mosebi, M. A. Papathanasopoulos and R. Hewer (2015). "Functional roles of HIV-1 Vpu and CD74: Details and implications of the Vpu–CD74 interaction." Cellular Immunology **298**(1–2): 25-32.

Le Rouzic, E., D. Bonnard, S. Chasset, J.-M. Bruneau, F. Chevreuil, F. Le Strat, J. Nguyen, R. Beauvoir, C. Amadori, J. Brias, S. Vomscheid, S. Eiler, N. Lévy, O. Delelis, E. Deprez, A. Saïb, A. Zamborlini, S. Emiliani, M. Ruff, B. Ledoussal, F. Moreau and R. Benarous (2013). "Dual inhibition of HIV-1 replication by integrase-LEDGF allosteric inhibitors is predominant at the post-integration stage." Retrovirology **10**(1): 144.

Leibly, D. J., T. N. Nguyen, L. T. Kao, S. N. Hewitt, L. K. Barrett and W. C. Van Voorhis (2012). "Stabilizing Additives Added during Cell Lysis Aid in the Solubilization of Recombinant Proteins." PLOS ONE **7**(12): e52482.

Lesbats, P., A. N. Engelman and P. Cherepanov (2016). "Retroviral DNA Integration." Chem Rev **116**(20): 12730-12757.

Lessells, R. J., D. K. Katzenstein and T. de Oliveira (2012). "Are subtype differences important in HIV drug resistance?" Current opinion in virology **2**(5): 636-643.

Lewis, P., M. Hensel and M. Emerman (1992). "Human immunodeficiency virus infection of cells arrested in the cell cycle." The EMBO Journal **11**(8): 3053-3058.

Liu, J., A. Bartesaghi, M. J. Borgnia, G. Sapiro and S. Subramaniam (2008). "Molecular architecture of native HIV-1 gp120 trimers." Nature **455**(7209): 109-113.

Llano, M., S. Delgado, M. Vanegas and E. M. Poeschla (2004). "Lens Epithelium-derived Growth Factor/p75 Prevents Proteasomal Degradation of HIV-1 Integrase." Journal of Biological Chemistry **279**(53): 55570-55577.

Llano, M., D. T. Saenz, A. Meehan, P. Wongthida, M. Peretz, W. H. Walker, W. Teo and E. M. Poeschla (2006). "An Essential Role for LEDGF/p75 in HIV Integration." Science **314**(5798): 461.

Llano, M., M. Vanegas, O. Fregoso, D. Saenz, S. Chung, M. Peretz and E. M. Poeschla (2004). "LEDGF/p75 Determines Cellular Trafficking of Diverse Lentiviral but Not Murine Oncoretroviral Integrase Proteins and Is a Component of Functional Lentiviral Preintegration Complexes." Journal of Virology **78**(17): 9524-9537.

Llano, M., M. Vanegas, N. Hutchins, D. Thompson, S. Delgado and E. M. Poeschla (2006). "Identification and Characterization of the Chromatin-binding Domains of the HIV-1 Integrase Interactor LEDGF/p75." Journal of Molecular Biology **360**(4): 760-773.

Lodi, P. J., J. A. Ernst, J. Kuszewski, A. B. Hickman, A. Engelman, R. Craigie, G. M. Clore and A. M. Gronenborn (1995). "Solution Structure of the DNA Binding Domain of HIV-1 Integrase." Biochemistry **34**(31): 9826-9833.

Lutzke, R. A., C. Vink and R. H. Plasterk (1994). "Characterization of the minimal DNA-binding domain of the HIV integrase protein." Nucleic Acids Research **22**(20): 4125-4131.

Maertens, G., P. Cherepanov, Z. Debyser, Y. Engelborghs and A. Engelman (2004). "Identification and Characterization of a Functional Nuclear Localization Signal in the HIV-1 Integrase Interactor LEDGF/p75." Journal of Biological Chemistry **279**(32): 33421-33429.

Maertens, G., P. Cherepanov, W. Pluymers, K. Busschots, E. De Clercq, Z. Debyser and Y. Engelborghs (2003). "LEDGF/p75 Is Essential for Nuclear and Chromosomal Targeting of HIV-1 Integrase in Human Cells." Journal of Biological Chemistry **278**(35): 33528-33539.

Maillot, B., N. Lévy, S. Eiler, C. Crucifix, F. Granger, L. Richert, P. Didier, J. Godet, K. Pradeau-Aubreton, S. Emiliani, A. Nazabal, P. Lesbats, V. Parissi, Y. Mely, D. Moras, P. Schultz and M. Ruff (2013). "Structural and Functional Role of INI1 and LEDGF in the HIV-1 Preintegration Complex." PLOS ONE **8**(4): e60734.

Malhotra, S., O. K. Mathew and R. Sowdhamini (2015). "DOCKSCORE: a webserver for ranking protein-protein docked poses." BMC Bioinformatics **16**(1): 127.

Marini, B., A. Kertesz-Farkas, H. Ali, B. Lucic, K. Lisek, L. Manganaro, S. Pongor, R. Luzzati, A. Recchia, F. Mavilio, M. Giacca and M. Lusic (2015). "Nuclear architecture dictates HIV-1 integration site selection." Nature **521**(7551): 227-231.

Marion, D., P. C. Driscoll, L. E. Kay, P. T. Wingfield, A. Bax, A. M. Gronenborn and G. M. Clore (1989). "Overcoming the overlap problem in the assignment of proton NMR spectra of larger proteins by use of three-dimensional heteronuclear proton-nitrogen-15 Hartmann-Hahn-multiple quantum coherence and nuclear Overhauser-multiple quantum coherence spectroscopy: application to interleukin 1.beta." Biochemistry **28**(15): 6150-6156.

Marion, D., L. E. Kay, S. W. Sparks, D. A. Torchia and A. Bax (1989). "Three-dimensional heteronuclear NMR of nitrogen-15 labeled proteins." Journal of the American Chemical Society **111**(4): 1515-1517.

Martin, S. R. and P. M. Bayley (2002). Absorption and Circular Dichroism Spectroscopy. Calcium-Binding Protein Protocols: Volume 2: Methods and Techniques. H. J. Vogel. Totowa, NJ, Springer New York**:** 43-55.

Maskell, D. P., L. Renault, E. Serrao, P. Lesbats, R. Matadeen, S. Hare, D. Lindemann, A. N. Engelman, A. Costa and P. Cherepanov (2015). "Structural basis for retroviral integration into nucleosomes." Nature **523**(7560): 366-369.

Matreyek, K. A. and A. Engelman (2013). "Viral and Cellular Requirements for the Nuclear Entry of Retroviral Preintegration Nucleoprotein Complexes." Viruses **5**(10): 2483-2511.

Matreyek, K. A., S. S. Yücel, X. Li and A. Engelman (2013). "Nucleoporin NUP153 Phenylalanine-Glycine Motifs Engage a Common Binding Pocket within the HIV-1 Capsid Protein to Mediate Lentiviral Infectivity." PLOS Pathogens **9**(10): e1003693.

McIntosh, L. P. and F. W. Dahlquist (2009). "Biosynthetic Incorporation of 15N and 13C for Assignment and Interpretation of Nuclear Magnetic Resonance Spectra of Proteins." Quarterly Reviews of Biophysics **23**(1): 1-38.

McNatt, M. W., T. Zang and P. D. Bieniasz (2013). "Vpu Binds Directly to Tetherin and Displaces It from Nascent Virions." PLOS Pathogens **9**(4): e1003299.

Mesplede, T., P. K. Quashie, N. Osman, Y. Han, D. N. Singhroy, Y. Lie, C. J. Petropoulos, W. Huang and M. A. Wainberg (2013). "Viral fitness cost prevents HIV-1 from evading dolutegravir drug pressure." Retrovirology **10**.

Michel, F., C. Crucifix, F. Granger, S. Eiler, J. F. Mouscadet, S. Korolev, J. Agapkina, R. Ziganshin, M. Gottikh and A. Nazabal (2009). "Structural basis for HIV-1 DNA integration in the human genome, role of the LEDGF/P75 cofactor." EMBO J **28**.

Miller, M. D., C. M. Farnet and F. D. Bushman (1997). "Human immunodeficiency virus type 1 preintegration complexes: studies of organization and composition." Journal of Virology **71**(7): 5382-5390.

Mitchell, R. S., B. F. Beitzel, A. R. W. Schroder, P. Shinn, H. Chen, C. C. Berry, J. R. Ecker and F. D. Bushman (2004). "Retroviral DNA Integration: ASLV, HIV, and MLV Show Distinct Target Site Preferences." PLoS Biology **2**(8): e234.

Morikawa, Y., D. J. Hockley, M. V. Nermut and I. M. Jones (2000). "Roles of Matrix, p2, and N-Terminal Myristoylation in Human Immunodeficiency Virus Type 1 Gag Assembly." Journal of Virology **74**(1): 16-23.

Moscoso, C. G., Y. Sun, S. Poon, L. Xing, E. Kan, L. Martin, D. Green, F. Lin, A. G. Vahlne, S. Barnett, I. Srivastava and R. H. Cheng (2011). "Quaternary structures of HIV Env immunogen exhibit conformational vicissitudes and interface diminution elicited by ligand binding." Proceedings of the National Academy of Sciences of the United States of America **108**(15): 6091-6096.

Munro, J. B., J. Gorman, X. Ma, Z. Zhou, J. Arthos, D. R. Burton, W. C. Koff, J. R. Courter, A. B. Smith, P. D. Kwong, S. C. Blanchard and W. Mothes (2014). "Conformational dynamics of single HIV-1 envelope trimers on the surface of native virions." Science **346**(6210): 759.

Naughtin, M., Z. Haftek-Terreau, J. Xavier, S. Meyer, M. Silvain, Y. Jaszczyszyn, N. Levy, V. Miele, M. S. Benleulmi, M. Ruff, V. Parissi, C. Vaillant and M. Lavigne (2015). "DNA Physical Properties and Nucleosome Positions Are Major Determinants of HIV-1 Integrase Selectivity." PLoS One **10**(6): e0129427.

Nickle, D. C., M. Rolland, M. A. Jensen, S. L. K. Pond, W. Deng, M. Seligman, D. Heckerman, J. I. Mullins and N. Jojic (2007). "Coping with Viral Diversity in HIV Vaccine Design." PLOS Computational Biology **3**(4): e75.

Ocwieja, K. E., T. L. Brady, K. Ronen, A. Huegel, S. L. Roth, T. Schaller, L. C. James, G. J. Towers, J. A. T. Young, S. K. Chanda, R. König, N. Malani, C. C. Berry and F. D. Bushman (2011). "HIV Integration Targeting: A Pathway Involving Transportin-3 and the Nuclear Pore Protein RanBP2." PLOS Pathogens **7**(3): e1001313.

Parent, L. J. (2011). "New insights into the nuclear localization of retroviral gag proteins." Nucleus **2**(2): 92-97.

Pasi, M., D. Mornico, S. Volant, A. Juchet, J. Batisse, C. Bouchier, V. Parissi, M. Ruff, R. Lavery and M. Lavigne (2016). "DNA minicircles clarify the specific role of DNA structure on retroviral integration." Nucleic Acids Research **44**(16): 7830-7847.

Passos, D. O., M. Li, R. Yang, S. V. Rebensburg, R. Ghirlando, Y. Jeon, N. Shkriabai, M. Kvaratskhelia, R. Craigie and D. Lyumkis (2017). "Cryo-EM structures and atomic model of the HIV-1 strand transfer complex intasome." Science **355**(6320): 89.

Peng, K., W. Muranyi, B. Glass, V. Laketa, S. R. Yant, L. Tsai, T. Cihlar, B. Müller and H.-G. Kräusslich (2014). "Quantitative microscopy of functional HIV post-entry complexes reveals association of replication with the viral capsid." eLife **3**: e04114.

Pettersen, E. F., T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng and T. E. Ferrin (2004). "UCSF Chimera—A visualization system for exploratory research and analysis." Journal of Computational Chemistry **25**(13): 1605-1612.

Pettit, S. C., M. D. Moody, R. S. Wehbie, A. H. Kaplan, P. V. Nantermet, C. A. Klein and R. Swanstrom (1994). "The p2 domain of human immunodeficiency virus type 1 Gag regulates sequential proteolytic processing and is required to produce fully infectious virions." Journal of Virology **68**(12): 8017-8027.

Piantadosi, A., B. Chohan, V. Chohan, R. S. McClelland and J. Overbaugh (2007). "Chronic HIV-1 Infection Frequently Fails to Protect against Superinfection." PLOS Pathogens **3**(11): e177.

Plantier, J.-C., M. Leoz, J. E. Dickerson, F. De Oliveira, F. Cordonnier, V. Lemee, F. Damond, D. L. Robertson and F. Simon (2009). "A new human immunodeficiency virus derived from gorillas." Nat Med **15**(8): 871-872.

Potterton, E., P. Briggs, M. Turkenburg and E. Dodson (2003). "A graphical user interface to the CCP4 program suite." Acta Crystallogr D Biol Crystallogr **59**.

Poulsen, F. M. (2002). "<A brief introduction to NMR spectroscopy of proteins>."

Powell, R. L. R., M. M. Urbanski, S. Burda, T. Kinge and P. N. Nyambi (2009). "High Frequency of HIV-1 Dual Infections Among HIV-Positive Individuals in Cameroon, West Central Africa." JAIDS Journal of Acquired Immune Deficiency Syndromes **50**(1): 84-92.

Pradeepa, M. M., H. G. Sutherland, J. Ule, G. R. Grimes and W. A. Bickmore (2012). "Psip1/Ledgf p52 Binds Methylated Histone H3K36 and Splicing Factors and Contributes to the Regulation of Alternative Splicing." PLOS Genetics **8**(5): e1002717.

Price, A. J., A. J. Fletcher, T. Schaller, T. Elliott, K. Lee, V. N. KewalRamani, J. W. Chin, G. J. Towers and L. C. James (2012). "CPSF6 Defines a Conserved Capsid Interface that Modulates HIV-1 Replication." PLOS Pathogens **8**(8): e1002896.

Quashie, P. K., T. Mesplède, Y.-S. Han, M. Oliveira, D. N. Singhroy, T. Fujiwara, M. R. Underwood and M. A. Wainberg (2012). "Characterization of the R263K Mutation in HIV-1 Integrase That Confers Low-Level Resistance to the Second-Generation Integrase Strand Transfer Inhibitor Dolutegravir." Journal of Virology **86**(5): 2696-2705.

Raghavendra, N. K., N. Shkriabai, R. L. J. Graham, S. Hess, M. Kvaratskhelia and L. Wu (2010). "Identification of host proteins associated with HIV-1 preintegration complexes isolated from infected CD4+ cells." Retrovirology **7**(1): 66.

Rausch, W. J. and F. S. Grice (2015). "HIV Rev Assembly on the Rev Response Element (RRE): A Structural Perspective." Viruses **7**(6).

Ren, X. and J. H. Hurley (2011). "Proline-rich regions and motifs in trafficking: From ESCRT interaction to viral exploitation." Traffic (Copenhagen, Denmark) **12**(10): 1282-1290.

Rizzuto, C. D., R. Wyatt, N. Hernández-Ramos, Y. Sun, P. D. Kwong, W. A. Hendrickson and J. Sodroski (1998). "A Conserved HIV gp120 Glycoprotein Structure Involved in Chemokine Receptor Binding." Science **280**(5371): 1949.

Sax, P. E., J. L. Meyers, M. Mugavero and K. L. Davis (2012). "Adherence to antiretroviral treatment and correlation with risk of hospitalization among commercially insured HIV patients in the United States." PLoS One **7**(2): e31591.

Schröder, A. R. W., P. Shinn, H. Chen, C. Berry, J. R. Ecker and F. Bushman (2002). "HIV-1 Integration in the Human Genome Favors Active Genes and Local Hotspots." Cell **110**(4): 521-529.

Schwieters, C. D., J. J. Kuszewski, N. Tjandra and G. Marius Clore (2003). "The Xplor-NIH NMR molecular structure determination package." Journal of Magnetic Resonance **160**(1): 65-73.

Seidel, S. A. I., P. M. Dijkman, W. A. Lea, G. van den Bogaart, M. Jerabek-Willemsen, A. Lazic, J. S. Joseph, P. Srinivasan, P. Baaske, A. Simeonov, I. Katritch, F. A. Melo, J. E. Ladbury, G. Schreiber, A. Watts, D. Braun and S. Duhr (2013). "Microscale Thermophoresis Quantifies Biomolecular Interactions under Previously Challenging Conditions." Methods (San Diego, Calif.) **59**(3): 301-315.

Serrao, E., L. Krishnan, M.-C. Shun, X. Li, P. Cherepanov, A. Engelman and G. N. Maertens (2014). "Integrase residues that determine nucleotide preferences at sites of HIV-1 integration: implications for the mechanism of target DNA binding." Nucleic Acids Research **42**(8): 5164-5176.

Sharma, A., R. C. Larue, M. R. Plumb, N. Malani, F. Male, A. Slaughter, J. J. Kessl, N. Shkriabai, E. Coward, S. S. Aiyer, P. L. Green, L. Wu, M. J. Roth, F. D. Bushman and M. Kvaratskhelia (2013). "BET proteins promote efficient murine leukemia virus integration at transcription start sites." Proceedings of the National Academy of Sciences of the United States of America **110**(29): 12036-12041.

Sharp, P. M. and B. H. Hahn (2011). "Origins of HIV and the AIDS Pandemic." Cold Spring Harbor Perspectives in Medicine: **1**(1): a006841.

Shaw, G. M. and E. Hunter (2012). "HIV Transmission." Cold Spring Harbor Perspectives in Medicine **2**(11): a006965.

Shun, M.-C., Y. Botbol, X. Li, F. Di Nunzio, J. E. Daigle, N. Yan, J. Lieberman, M. Lavigne and A. Engelman (2008). "Identification and Characterization of PWWP Domain Residues Critical for LEDGF/p75 Chromatin Binding and Human Immunodeficiency Virus Type 1 Infectivity." Journal of Virology **82**(23): 11555-11567.

Shun, M. C., N. K. Raghavendra, N. Vandegraaff, J. E. Daigle, S. Hughes, P. Kellam, P. Cherepanov and A. Engelman (2007). "LEDGF/p75 functions downstream from preintegration complex formation to effect gene-specific HIV-1 integration." Genes Dev **21**.

Simon-Loriere, E., D. P. Martin, K. M. Weeks and M. Negroni (2010). "RNA Structures Facilitate Recombination-Mediated Gene Swapping in HIV-1." Journal of Virology **84**(24): 12675-12682.

Sloan, R. D. and M. A. Wainberg (2011). "The role of unintegrated DNA in HIV infection." Retrovirology **8**(1): 52.

Sluis-Cremer, N., M. A. Wainberg and R. F. Schinazi (2015). "Resistance to reverse transcriptase inhibitors used in the treatment and prevention of HIV-1 infection." Future microbiology **10**(11): 1773-1782.

Smyth, R. P., T. E. Schlub, A. J. Grimm, C. Waugh, P. Ellenberg, A. Chopra, S. Mallal, D. Cromer, J. Mak and M. P. Davenport (2014). "Identifying Recombination Hot Spots in the HIV-1 Genome." Journal of Virology **88**(5): 2891-2902.

Sowd, G. A., E. Serrao, H. Wang, W. Wang, H. J. Fadel, E. M. Poeschla and A. N. Engelman (2016). "A critical role for alternative polyadenylation factor CPSF6 in targeting HIV-1 integration to transcriptionally active chromatin." Proceedings of the National Academy of Sciences **113**(8): E1054-E1063.

Ssemwanga, D., R. N. Nsubuga, B. N. Mayanja, F. Lyagoba, B. Magambo, D. Yirrell, L. Van der Paal, H. Grosskurth and P. Kaleebu (2013). "Effect of HIV-1 Subtypes on Disease Progression in Rural Uganda: A Prospective Clinical Cohort Study." PLOS ONE **8**(8): e71768.

Stephenson, K. E., H. T. D'Couto and D. H. Barouch (2016). "New concepts in HIV-1 vaccine development." Current Opinion in Immunology **41**: 39-46.

Turlure, F., G. Maertens, S. Rahman, P. Cherepanov and A. Engelman (2006). "A tripartite DNA-binding element, comprised of the nuclear localization signal and two AT-hook motifs, mediates the association of LEDGF/p75 with chromatin in vivo." Nucleic Acids Research **34**(5): 1653-1665.

Vagin, A. and A. Teplyakov (2010). "Molecular replacement with MOLREP." Acta Crystallogr D Biol Crystallogr **66**.

Van Heuverswyn, F., Y. Li, C. Neel, E. Bailes, B. F. Keele, W. Liu, S. Loul, C. Butel, F. Liegeois, Y. Bienvenue, E. M. Ngolle, P. M. Sharp, G. M. Shaw, E. Delaporte, B. H. Hahn and M. Peeters (2006). "Human immunodeficiency viruses: SIV infection in wild gorillas." Nature **444**(7116): 164-164.

van Nuland, R., F. M. A. van Schaik, M. Simonis, S. van Heesch, E. Cuppen, R. Boelens, H. T. M. Timmers and H. van Ingen (2013). "Nucleosomal DNA binding drives the recognition of H3K36-methylated nucleosomes by the PSIP1-PWWP domain." Epigenetics & Chromatin **6**: 12-12.

van Zundert, G. C. P., J. P. G. L. M. Rodrigues, M. Trellet, C. Schmitz, P. L. Kastritis, E. Karaca, A. S. J. Melquiond, M. van Dijk, S. J. de Vries and A. M. J. J. Bonvin (2016). "The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes." Journal of Molecular Biology **428**(4): 720-725.

Vranken, W. F., W. Boucher, T. J. Stevens, R. H. Fogh, A. Pajon, M. Llinas, E. L. Ulrich, J. L. Markley, J. Ionides and E. D. Laue (2005). "The CCPN data model for NMR spectroscopy: Development of a software pipeline." Proteins: Structure, Function, and Bioinformatics **59**(4): 687-696.

W.H.O. (2016). "Global Summary of the AIDS epidemic 2015." from http://www.who.int/hiv/data/epi_core_2016.png?ua=1.

Wang, D., W. Lu and F. Li (2015). "Pharmacological intervention of HIV-1 maturation." Acta Pharmaceutica Sinica B **5**(6): 493-499.

Wang, J.-Y., H. Ling, W. Yang and R. Craigie (2001). "Structure of a two-domain fragment of HIV-1 integrase: implications for domain organization in the intact protein." The EMBO Journal **20**(24): 7333-7343.

Watts, J. M., K. K. Dang, R. J. Gorelick, C. W. Leonard, J. W. Bess, Jr., R. Swanstrom, C. L. Burch and K. M. Weeks (2009). "Architecture and secondary structure of an entire HIV-1 RNA genome." Nature **460**(7256): 711-716.

Weng, Z., R. J. Rickles, S. Feng, S. Richard, A. S. Shaw, S. L. Schreiber and J. S. Brugge (1995). "Structure-function analysis of SH3 domains: SH3 binding specificity altered by single amino acid substitutions." Molecular and Cellular Biology **15**(10): 5627-5634.

Wertheim, J. O. and M. Worobey (2009). "Dating the Age of the SIV Lineages That Gave Rise to HIV-1 and HIV-2." PLoS Computational Biology **5**(5): e1000377.

Williamson, M. P. (2013). "Using chemical shift perturbation to characterise ligand binding." Prog Nucl Magn Reson Spectrosc **73**: 1-16.

Wlodawer, A., W. Minor, Z. Dauter and M. Jaskolski (2008). "Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures." The FEBS journal **275**(1): 1-21.

Yan, N., E. O'Day, L. A. Wheeler, A. Engelman and J. Lieberman (2011). "HIV DNA is heavily uracilated, which protects it from autointegration." Proceedings of the National Academy of Sciences **108**(22): 9244-9249.

Yap, K. L. and M.-M. Zhou (2010). "Keeping it in the family: diverse histone recognition by conserved structural folds." Critical Reviews in Biochemistry and Molecular Biology **45**(6): 488-505.

Yoder, K. E. and F. D. Bushman (2000). "Repair of Gaps in Retroviral DNA Integration Intermediates." Journal of Virology **74**(23): 11191-11200.

Zheng, R., T. M. Jenkins and R. Craigie (1996). "Zinc folds the N-terminal domain of HIV-1 integrase, promotes multimerization, and enhances catalytic activity." Proceedings of the National Academy of Sciences of the United States of America **93**(24): 13659-13664.

Zhou, C. and T. M. Rana (2002). "A Bimolecular Mechanism of HIV-1 Tat Protein Interaction with RNA Polymerase II Transcription Elongation Complexes." Journal of Molecular Biology **320**(5): 925-942.

Ziarek, J. J., F. C. Peterson, B. L. Lytle and B. F. Volkman (2011). "Binding site identification and structure determination of protein-ligand complexes by NMR." Methods in enzymology **493**: 241-275

Résumé

L'Integrase du VIH est une ADN recombinase catalysant deux réactions qui permettent l'intégration de l'ADN viral dans l'ADN hôte. L'intégrase du VIH comprend 3 domaines : N-terminal impliqué dans la réaction de « 3' processing » et le transfert de brin, le domaine catalytique contenant le site actif et le domaine C-terminal liant l'ADN non-spécifiquement (CTD). Des recherches récentes mettent en évidence l'importance du CTD dans la liaison avec d'autres protéines virales comme la transcriptase inverse. Le but de la thèse était de comprendre les rôles et l'importance du domaine C-terminal de l'intégrase dans deux contextes : l'intégration dans la chromatine et la coévolution, avec l'objectif de comprendre le rôle de la multimerisation dans la fonction de l'intégrase. Globalement, les résultats de mon projet indiquent que l'IN-CTD joue un rôle important, en contribuant à la formation de multimères d'ordre supérieur importants pour la fonction de l'IN.

Mots clés : L'Integrase du VIH, le domaine C-terminal, chromatine, coévolution, multimerisation

Résumé en anglais

HIV Integrase is a DNA recombinase that catalyzes two endonucleolytic reactions that allow the viral DNA integration into host DNA for replication and subsequent viral protein production. HIV Integrase consists of 3 structural and functional domains: The N-terminal zinc domain involved in 3' processing and strand transfer, the catalytic core domain which contains the active site, and the C-terminal domain that binds DNA non-specifically. Recent research highlights the importance of the CTD in binding with other viral proteins such as Reverse Transcriptase. The aim of the thesis was to understand the roles and importance of the C-terminal domain of HIV-1 Integrase in two contexts: chromatin integration, and co-evolution, with the overall purpose of understanding the role of multimerization in IN function. Overall, results from my project indicate that the IN-CTD plays an important role, by contributing to the formation of higher order multimers that are important for IN functionality.

Keywords : HIV Integrase, C-terminal domain, chromatin, coevolution, multimerization