

University of Alberta

Homology modeling of Suv39h1 and discovery of its small molecule inhibitors by virtual screening and their *in vitro* validation

by

Ishwar Vithal Hosamani

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

**Master of Science
in
Experimental Oncology**

Department of Oncology

©Ishwar V. Hosamani
Fall 2013
Edmonton,
Alberta

Permission is hereby granted to the University of Alberta Libraries to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly or scientific research purposes only. Where the thesis is converted to, or otherwise made available in digital form, the University of Alberta will advise potential users of the thesis of these terms.

The author reserves all other publication and other rights in association with the copyright in the thesis and, except as herein before provided, neither the thesis nor any substantial portion thereof may be printed or otherwise reproduced in any material form whatsoever without the author's prior written permission.

Dedication

To my parents, brother and sister

Abstract

Epigenetic modifications are carried out by specific enzymes and are reversible making them a viable and attractive target to design inhibitors to reset the epigenetic regulatory machinery of the cell and restore it to its normal state. Suppressor of variegation 3-9 homolog 1 (Suv39h1) is a histone 3 lysine 9 (H3K9) trimethylase that plays an important role in heterochromatin formation, mitosis, and its misregulation has been implicated in several types of cancer. In this thesis, a homology model of human Suv39h1 was constructed, optimized and validated and used for virtual screening against several small molecule databases to find its potential small molecule inhibitors. The binding sites of three ligands i.e. S-Adenosyl Methionine, Chaetocin and N-terminal of H3K9 peptide, on the homology model of Suv39h1 were determined and used for virtual screening. The small molecules predicted to bind to Suv39h1 with high affinity were tested by an *in vitro* assay.

Acknowledgements

I remember the day when I had my first meeting with my supervisors Dr. Gordon Chan and Dr. Jack Tuszynski and nodding my head along to what they were explaining to me. Needless to say I was struggling to completely understand what they were saying. The journey since then has been one with constant learning and discovery, which has finally culminated in the form of this thesis.

I am very grateful to Drs. Chan and Tuszynski for giving me this opportunity to learn and work in their labs. Thank you for your guidance over the past three years and at the same time giving me the freedom to develop my own ideas.

I would also like to thank my supervisory committee members Drs. Michael Hendzel and Alan Underhill for the several discussions and invaluable advice that I have received from both of you, both academic and otherwise.

I would also like to thank the members of the Chan lab (Devinderjit Kaur, Dr. Larissa Vos) and the Tuszynski lab (especially Philip Winter and Dr. Richard Tseng) for being supportive, understanding and teaching me everything that I did not know. Special thanks to Dr. Dawn Macdonald for being my mentor and guide and helping me to overcome any difficulties I faced during my program.

My time here in Edmonton and the department would not have been as enjoyable and fruitful if not for the support and encouragement from some of the most awesome friendships that I have developed over the past few years. Thank you Dr. David Sharon, Dr. Nikhil Raghuram, Dr. Hilmar Strickfaden and Dr. Sheena Macleod for sharing probably a few hundred gallons of beer and the inevitable deeply intellectual, philosophical, scientific and the silly discussions that the evenings presented over the last few years. I would also like to thank my friends Anil Menon, Vijetha Bhat, Lalit Pant and Sunil DM for their support and encouragement without which this would not have been possible.

Last but not the least I would like to thank my parents and my siblings for believing in me and giving me the freedom to do what I always wanted to do.

Table of Contents

1. Chapter 1 – Introduction.....	1
1.1. Histones and nucleosomes.....	1
1.2. Epigenetics and histone modifications	3
1.3. Histone methylation.....	5
1.3.1. Arginine methylation	7
1.3.2. Lysine methylation.....	9
1.4. H3K9 methylation and its significance	17
1.5. <i>SUV39h1</i> (Suppressor of variegation 3-9 homolog 1).....	18
1.5.1. <i>Suv39h1</i> and its role in mitosis.....	20
1.5.2. <i>Suv39h1</i> and cancer.....	23
1.6. Histone methyltransferases and cancer.....	24
1.7. Epigenetic modifiers as targets for cancer therapy.....	25
1.7.1. <i>Histone methyltransferases as targets for cancer therapy</i>	28
1.8. Research focus of this thesis.....	31
1.9. Hypothesis	32
1.10. Homology modeling.....	33
1.10.1. <i>Why do we need homology modeling?</i>	33
1.10.2. <i>Steps involved in homology modeling</i>	36
1.10.3. <i>Homology modeling using SWISS-MODEL server</i>	45
1.11. Energy minimization and molecular dynamics (MD) simulations.....	50
1.12. Molecular docking.....	51
1.12.1. <i>Rigid docking</i>	52
1.12.2. <i>Flexible docking</i>	53
1.13. Determination of potential ligand binding sites by blind docking	53

1.14. Small molecule databases.....	53
1.15. Virtual screening.....	56
1.15.1. <i>AutoDock</i>	57
1.15.2. <i>AutoDock vina</i>	57
2. Chapter 2 – Materials and Methods.....	59
2.1. Cloning of <i>SUV39h1</i>	59
2.2. Mutagenesis of h <i>SUV39h1</i>	68
2.3. Expression of human Suv39h1.....	75
2.3.1. <i>Bacterial expression system</i>	75
2.3.2. <i>Mammalian expression system</i>	78
2.4. Protein purification.....	78
2.5. Sodium Dodecyl Sulphate – Polyacrylamide Gel Electrophoresis (SDS-PAGE)..	79
2.6. Western blot.....	80
2.7. Reagent recipes.....	81
2.8. Homology modeling of Suv39h1.....	83
2.8.1. <i>Identification of template for Suv39h1 homology modeling</i>	83
2.8.2. <i>Homology modeling using Swiss model server</i>	83
2.8.3. <i>Quality estimation of the predicted models</i>	84
2.9. Energy minimizations and molecular dynamics simulations.....	84
2.10. Molecular docking to determine potential ligand binding sites.....	85
2.10.1. <i>Determination of binding site of S-Adenosyl methionine</i>	85
2.10.2. <i>Determination of binding site of chaetocin</i>	86
2.10.3. <i>Determination of binding site of H3K9 peptide</i>	86
2.11. Small molecule inhibitor libraries.....	87
2.12. Virtual screening.....	88

2.13.	Analysis and clustering of the top hits obtained from virtual screening	89
2.14.	<i>In vitro</i> histone methyltransferase assay	89
2.14.1.	<i>Fluorescent in vitro histone methyltransferase assay</i>	89
2.14.2.	<i>Radioactive in vitro histone methyltransferase assay</i>	93
3.	Chapter 3 – Results	97
3.1.	Expression of recombinant Suv39h1	97
3.1.1.	<i>Bacterial protein expression system</i>	97
3.1.2.	<i>Mammalian protein expression system</i>	99
3.2.	Homology modeling of Suv39h1	101
3.2.1.	<i>Identification of template structure</i>	101
3.2.2.	<i>Homology model prediction using the Swiss model server</i>	104
3.2.3.	<i>Quality assessment of predicted model</i>	106
3.2.4.	<i>Energy minimization and MD simulation studies</i>	113
3.3.	Calculation of RMSD of the simulated structures	116
3.4.	Clustering of the simulated structures	118
3.5.	Determination of potential ligand binding sites	120
3.5.1.	<i>SAM binding site</i>	122
3.5.2.	<i>Chaetocin binding site</i>	126
3.5.3.	<i>Histone H3 N-terminal binding site</i>	129
3.6.	Small molecule inhibitors obtained from virtual screening	133
3.7.	Testing of the small molecule inhibitors by <i>in vitro</i> assay	149
3.7.1	Fluorescent histone methyltransferase assay	149
3.7.2	Validation of the radioactive <i>in vitro</i> histone methyltransferase assay	151
3.7.3	Testing of the small molecule inhibitors by radioactive <i>in vitro</i> assay	154

3.7.3.1 . <i>In vitro</i> methyltransferase assay results for the NCI diversity set 2 compounds	155
3.7.3.2 . <i>In vitro</i> methyltransferase assay results for the DrugBank, ZINC and ZDD compounds	163
4. Chapter 4 – Discussion and Future Directions.....	171
4.1. Expression of human Suv39h1	171
4.2. Homology modeling of human Suv39h1 and virtual screening.....	173
4.3. Development of an <i>in vitro</i> histone methyltransferase assay	175
4.4. Validation of the small molecule inhibitors by the <i>in vitro</i> histone methyltransferase assay.....	177
4.5. Future directions	178
Bibliography	180
Appendix.....	190
Structure function analysis of CENP-35/PHF2 during mitosis	190
1. Chapter 1 – Introduction.....	193
1.1. CENP-35.....	193
2. Chapter 2 - Materials and methods.....	195
2.1. Cloning and mutagenesis of CENP-35.....	195
2.2. Cell Culture.....	200
2.3. PEI transfection	200
2.4. HaloTag technology	201
2.5. Fluorescence microscopy	201
2.6. Live cell imaging.....	202

3. Chapter 3 – Results	204
3.1. Cloning and Mutagenesis of CENP-35	204
3.2. CENP-35 localization during the cell cycle	207
3.3. Cell cycle stage quantification of CENP-35 transfected cells.....	214
3.4. Live cell imaging	216
3.5. Expression of CENP-35 truncation fragment proteins.....	218
4. Chapter 4 –Discussions and Future Directions	220
Bibliography:	224

List Of Tables

Table 1.1: Sites of histone lysine methylation and their respective methyltransferases and demethylases	13
Table 1.2: Histone arginine methylation sites and their respective methyltransferases and demethylases	14
Table 1.3: Histone lysine methyltransferases and link to diseases	15
Table 1.4: Histone arginine methyltransferases and their links to disease	16
Table 1.5: FDA approved drugs that inhibit epigenetic mechanisms involved in cancer	27
Table 1.6: Structural representation of the small molecule inhibitors of various histone methyltransferases	30
Table 1.7: List of servers and software available to perform sequence alignment, homology modeling and model evaluation	39
Table 2.1: List of vectors used for cloning <i>hSUV39h1</i> and their properties	64
Table 2.2: List of primers used for cloning full-length <i>hSUV39h1</i>	65
Table 2.3: List of the sequencing primers used to sequence different vectors as shown in the table.	66
Table 2.4: List of expression vectors prepared containing full-length <i>hSUV39h1</i>	67
Table 2.5: List of primers used to generate various truncation mutants shown in Table 2.6.....	71
Table 2.6: List of <i>hSUV39h1</i> truncation mutants prepared.....	72
Table 2.7: List of primers used to create site directed mutants of <i>hSUV39h1</i>	73
Table 2.8: Site directed mutant constructs of full-length <i>hSUV39h1</i>	74
Table 2.9: List of the bacterial expression vectors and parameters used for testing the expression of human Suv39h1 fusion proteins.	77
Table 2.10: List of the small molecule inhibitor libraries screened and the number of small molecules in each of them.....	87
Table 2.11: Dimensions of the grid boxes and their centers at different binding pockets used for virtual screening.....	88

Table 2.12: List of samples used to test the fluorescent histone methyltransferase assay	92
Table 2.13: Experimental plan with all the necessary controls to validate the methyltransferase assay.....	95
Table 3.1: Bacterial expression vectors, <i>E.coli</i> strains, and expression parameters used to express SUV39h1	98
Table 3.2: List of small molecules obtained from virtual screening that bind to Suv39h1 with high affinity.	148
Table 3.3 Experimental plan and results of validation of the <i>in vitro</i> histone methyltransferase assay with necessary control.....	152
Table 3.4: Results of the inhibition assay for NCI Diversity set 2 compounds (Batch 1) at 5 μ M inhibitor concentration	156
Table 3.5: Results of the inhibition assay for NCI Diversity set 2 compounds (Batch 2) at 5 μ M inhibitor concentration	158
Table 3.6: Results of the inhibition assay for all NCI Diversity set 2 compounds tested at 1 μ M concentration.....	161
Table 3.7: Results of the inhibition assay for DrugBank, ZINC, ZDD compounds at 5 μ M inhibitor concentration (Batch 1)	164
Table 3.8: Results of the inhibition assay for DrugBank, ZINC, ZDD compounds at 5 μ M inhibitor concentration (Batch 2)	166
Table 3.9: Results of the inhibition assay for DrugBank, ZINC, ZDD compounds at 1 μ M inhibitor concentration	168

List Of Figures

Figure 1.1: Organization of DNA inside the nucleus.....	2
Figure 1.2: Mechanism of action of arginine methylation by PRMTs	8
Figure 1.3: Mechanism of action of lysine methylation by HMTs.....	10
Figure 1.4: Organization of Histone H3 modifications at the centromere of a metaphase chromosome	22
Figure 1.5: Yearly growth in the number of protein structures solved and submitted to the protein databank.	35
Figure 1.6: Steps involved in homology modeling.....	40
Figure 1.7: Representation of the Phi and Psi angles and Ramachandran plot.....	49
Figure 2.1: Schematic diagrams of the expression vectors used for cloning <i>hSUV39h1</i>	62
Figure 2.2: Schematic representation of the steps involved in fluorescence histone methyltransferase assay.....	91
Figure 3.1: Western blot showing expression of human Suv39h1 in the mammalian expression system.....	100
Figure 3.2: PSI-BLAST search result showing sequence identity between the sequence of Suv39h1 corresponding to the sequence of Suv39h2 crystal structure.....	102
Figure 3.3: Sequence alignment of the SET domains of human Suv39h1 and human Suv39h2.....	103
Figure 3.4: Homology model obtained from the Swiss model homology-modeling server.	105
Figure 3.5: ANOLEA plot for the homology model of Suv39h1 obtained from Swiss model server.....	107
Figure 3.6: QMEAN plot for the homology model obtained from Swiss model server.....	109
Figure 3.7: The QMEAN Z-score for the homology model of Suv39h1 obtained from Swiss model server.....	110

Figure 3.8: The Ramachandran plot for the homology model of Suv39h1 obtained from Swiss model.....	112
Figure 3.9: Plot of the potential energy of the system over a period of 23ns MD simulation.....	114
Figure 3.10: Superposition of the original homology model with the model after MD simulation.	115
Figure 3.11: A plot of the root mean square deviation of the C- α backbone of the model structures over a period of 23ns	117
Figure 3.12: Visualization of the structures at different time points during MD simulations.	119
Figure 3.13: Visualization of the blind docking method.	121
Figure 3.14: Clustering of binding conformations of SAM on Suv39h1 based on binding energies upon blind docking.....	123
Figure 3.15: Comparison of the binding conformation of SAM in SUV39h2 crystal structure and Suv39h1 homology model.....	124
Figure 3.16: Interaction map of SAM binding to various residues of Suv39h1 in its binding pocket.....	125
Figure 3.17: Cluster analysis of blind docking results of chaetocin	127
Figure 3.18: Comparison of the binding pockets of chaetocin and SAM.....	128
Figure 3.19: Sequence alignment of SET domain of Suv39h1 and EHMT1 shows 68% sequence similarity	130
Figure 3.20: 2D interaction map showing the interaction of mono-methylated H3K9 peptide with EHMT1.....	131
Figure 3.21: Superposition of EHMT1 and Suv39h1 to determine the binding pocket of H3K9 pocket.....	132
Figure 3.22: Results of fluorescent methyltransferase assay	150
Figure 3.23: Results of the validation of <i>in vitro</i> histone methyltransferase assay	153
Figure 3.24: Radioactive counts per minute (CPM) of NCI Diversity set compounds (Batch 1) assayed at 5 μ M concentration	157

Figure 3.25: Radioactive counts per minute (CPM) of NCI Diversity set compounds (Batch 2) assayed at 5 μ M concentration	159
Figure 3.26: Radioactive counts per minute (CPM) of NCI Diversity set compounds assayed at 1 μ M concentration	162
Figure 3.27: Radioactive counts per minute (CPM) of DrugBank, ZINC, ZDD compounds assayed at 5 μ M inhibitor concentration (Batch 1)	165
Figure 3.28: Radioactive counts per minute (CPM) of DrugBank, ZINC, ZDD compounds assayed at 5 μ M inhibitor concentration (Batch 2)	167
Figure 3.29: Radioactive counts per minute (CPM) of DrugBank, ZINC, ZDD compounds assayed at 1 μ M inhibitor concentration.	169

List Of Symbols And Abbreviations

aa	Amino acid
ADMET	Absorption, Distribution, Metabolism, Elimination, Toxicity
AdoMet	S-Adenosyl Methionine
ALA	Alanine
AMBER	Assisted Model Building with Energy Refinement
AML	Acute Myeloid Leukemia
ARG	Arginine
ASN	Asparagine
ASP	Aspartic acid
BLAST	Basic Local Alignment Search Tool
CPM	Counts Per Minute
CYS	Cysteine
DAPI	4',6-diamidino-2-phenylindole
DMSO	Dimethyl Sulfoxide
DNMT	DNA methyltransferase
ELISA	Enzyme Linked Immuno Sorbant Assay
EZH2	Enhancer of Zeste homolog 2
FDA	Food and Drug Administration
GFP	Green Fluorescent Protein
GST	Glutathione S-Transferase
HDAC	Histone deacetylase
HIS/HIE	Histidine
HMM	Hidden Markov Model
HMT	Histone methyltransferase
HP1	Heterochromatin Protein 1
HRP	Horse Radish Peroxide
HTS	High Throughput Screening
IC ₅₀	Half maximal inhibitory concentration
IPTG	Isopropyl β -D-thiogalactoside
logP	Partition co-efficient
LSD1	Lysine Specific Demethylase 1
LYS	Lysine
MBP	Maltose Binding Protein
MD	Molecular Dynamics
MEF	Mouse Embryonic Fibroblasts
NMR	Nuclear Magnetic Resonance
ORF	Open Reading Frame
PAGE	Polyacrylamide Gel Electrophoresis
PBS	Phosphate Buffer Saline

PCR	Polymerase Chain Reaction
PDB	Protein Data Bank
PEI	Polyethyleneimine
PEV	Position Effect variegation
PHE	Phenyl Alanine
PKMT	Protein Lysine Methyltransferase
PRMT	Protein Arginine Methyltransferase
PSI-BLAST	Position Specific Iterated Basic Local Alignment Search Tool
PTM	Post translation modification
Rb	Retinoblastoma protein (pRB)
RMSD	Root Mean Square Deviation
SAH	S-Adenosyl Homocysteine
SAM	S-Adenosyl Methionine
SBP	Streptavidin Binding Protein
SDS	Sodium Dodecyl Sulphate
SET	Suppressor of variegation, Enhancer of Zeste and Trithorax
SMI	Small molecule inhibitor
Su(var)	Suppressor of variegation
TAP	Tandem Affinity Purification
THR	Threonine
TRP	Tryptophan
TYR	Tyrosine
ZDD	ZINC Drug Database

UNITS OF MEASUREMENT

°C	Degree Celsius
Å	Angstrom
Ci	Curie
kCal/mol	Kilocalorie/mole
L	Liter
m	Meter
ml	Milliliter
µg	Microgram
µl	Microliter
µM	Micromolar
ns	nanosecond

Chapter 1 – Introduction

1. Chapter 1 – Introduction

1.1. Histones and nucleosomes

The eukaryotic genome is organized as a nucleoprotein complex called chromatin. Chromatin is a polymer of individual units, called nucleosomes, which are composed of 147 bases of DNA wrapped around an octameric scaffold of histone proteins (Luger, Mader et al. 1997). There are five different types of histone proteins: H1, H2A, H2B, H3 and H4. Each nucleosome is composed of dimers of histones H2A and H2B, and a tetramer of H3 and H4 histone proteins. The nucleosomes are linked to each other via histone H1, and this structure consisting of the nucleosome and histone H1 is called a chromatosome (Figure 1.1). Histone H1 also stabilizes the nucleosomal arrays and assists their condensation into chromatin fibers. These chromatin fibers are thought to condense further and form loops leading to the formation of higher order chromatin structure. The mechanism by which these chromatin fibers ultimately form chromosomes is still unclear and under investigation.

The nucleosomes are the fundamental units of chromatin (Kornberg 1974). The structure of each histone is highly conserved, consisting of a globular core domain and an unstructured N-terminal tail that protrudes from the surface (Luger, Mader et al. 1997). These N-terminal tails are also the sites of several post-translational modifications (PTMs) such as acetylation, methylation, phosphorylation that regulate the higher order organization of chromatin (Luger and Richmond 1998).

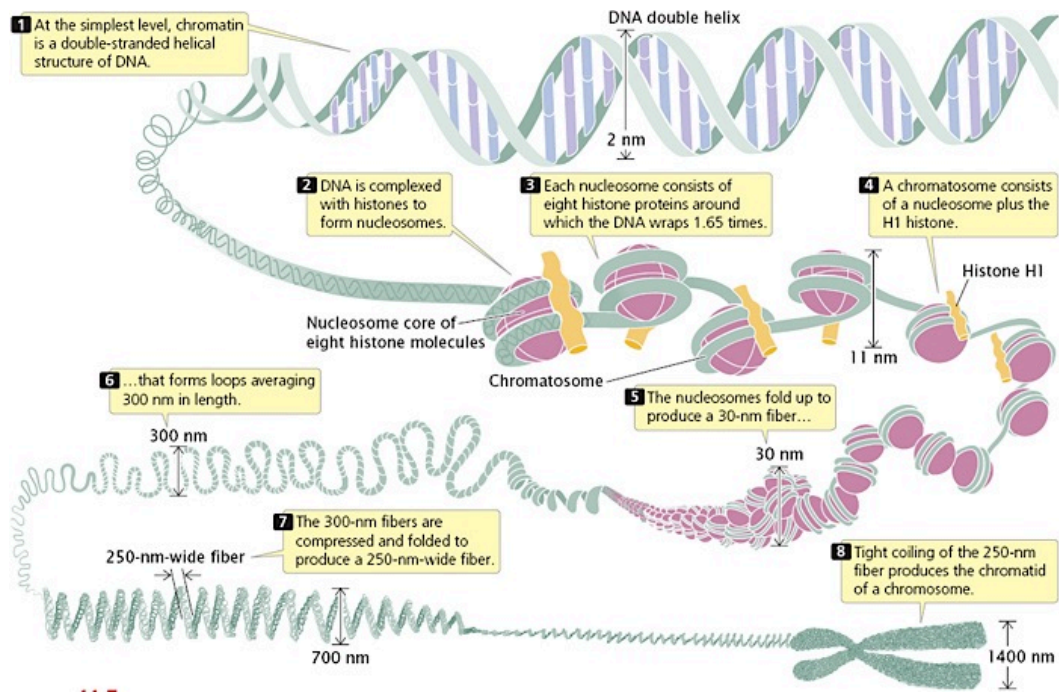


Figure 1.1: Organization of DNA inside the nucleus.

DNA double helix wraps around the histone proteins to form the nucleosomes. It is speculated that the array of nucleosomes condenses further to form the 30nm fiber of chromatin, which condenses further to form higher order chromatin ultimately condensing to form the chromosome. Figure adapted from © 2005 [W. H. Freeman](#) Pierce, Benjamin. *Genetics: A Conceptual Approach*, 2nd ed. (New York: W. H. Freeman and Company), 292.

1.2. Epigenetics and histone modifications

Epigenetics is the study of heritable changes in gene expression without any changes to the underlying DNA sequence itself. Although DNA is the material of choice for storage and transfer of genetic information from one generation to the next, there is another layer of information added to it by means of post-translational modifications of histone tails. Although all the cells in multicellular organisms contain identical genetic information, each cell type varies in its phenotype and performs distinct functions. This variation in the function of different cell types occurs because of the modulation of gene expression by various epigenetic mechanisms. There are three kinds of epigenetic mechanisms: post-translational histone modifications; DNA methylation and micro RNA mediated modulation of gene transcription. These epigenetic mechanisms control the gene expression profiles of each cell type, thus determining their function.

The N-terminal tails of histones are modified in numerous ways such as acetylation, methylation, phosphorylation, sumoylation, ribosylation and ubiquitination (Kouzarides 2007). Enzymes that are specific to individual sites carry out these epigenetic modifications. These post-translational modifications regulate the interaction of histones with other proteins, which bind to the modified sites and in turn control the chromatin organization and therefore alter gene expression (Strahl and Allis 2000; Rando 2007). Apart from gene transcription, epigenetic mechanisms also control other cellular processes such as DNA replication and DNA damage repair.

Each type of modification results in either activation or repression of gene expression based on the residue being modified and in some cases the extent to which it is being modified, i.e. one, two or three groups being added to the amino acid as in the case of methylation of lysine residues. For example, mono methylation of histone 3 lysine 4 (H3K4) is predominantly found in active euchromatin regions and is associated with activation of transcription, whereas trimethylated histone 3 lysine 9 (H3K9) is enriched in the silent heterochromatin regions and is associated with repression of gene transcription (Rice, Briggs et al. 2003). In general, acetylation of all histone substrates is associated with activation of gene transcription; whereas methylation is associated with repression and activation of gene transcription (Rice and Allis 2001; Kouzarides 2007). These differential effects of post-translational histone modifications on the structure and function of chromatin may be explained by the “histone code hypothesis” (Strahl and Allis 2000; Jenuwein and Allis 2001). This hypothesis states that epigenetic modifications at different sites on the histone tails act in a combinatorial manner leading to distinct functional outputs. These outputs are brought about by the recruitment of downstream effector proteins that alter the structure and function of chromatin, and the outcome of each modification depends on the function of the protein binding to that site and its interaction with other putative binding proteins at other post-translationally modified sites (Jenuwein and Allis 2001).

Epigenetic effects of histone modifications occur by three important steps involving three different classes of proteins. First, histone PTM “writers” that modify the amino acid residues of histones, such as, histone methyltransferases;

second, histone PTM “readers” that recognize the histone PTMs, for example, proteins containing a chromodomain or plant homeo domain (PHD); and third, histone PTM “erasers” that remove the PTMs, such as, histone demethylases. The proper orchestration of the functioning of these three classes of proteins is essential for establishing and maintaining the correct gene expression patterns in the cell. Any misregulation in expression and/or function of these proteins could lead to cellular dysfunction and contribute to disease (initiation and/or progression).

1.3. Histone methylation

Histone methylation occurs on the lysine and arginine residues of histone tails. Lysine can be tri-methylated where as arginine can be di-methylated. The enzymes that methylate these residues are called histone methyltransferases (HMT) and they use S-Adenosyl Methionine (SAM)/AdoMet as the methyl group donor which is converted to S-Adenosyl homocysteine (SAH/AdoHcy) (Figure 1.2 and Figure 1.3). Histone methylation is involved in transcription regulation, X-chromosome inactivation, heterochromatin formation and maintenance, as well as in DNA repair and genomic imprinting (Zhang and Reinberg 2001; Martin and Zhang 2005). Histone methylation influences the structure and function of chromatin without changing the charge of the histone tails. Methylation of lysine and arginine residues on the N-terminal tails of histones H3 and H4 leads to the organization of chromatin into heterochromatin and euchromatin depending on the residue being modified and the extent of its modification i.e. mono-, di- or tri-

methylation (Rice, Briggs et al. 2003). Recently, non-histone protein substrates of these methyltransferases, which were thought to methylate histone tails only, have also been identified (Huang and Berger 2008). Consequently, these methyltransferases have been reclassified as protein lysine methyltransferases (PKMTs) and protein arginine methyltransferases (PRMTs) to account for the non-histone substrates.

Due to the low turnover rates of the methyl groups which is comparable to the half life of histones, methylation of histone tails was considered to be stable compared to other post translational modifications such as acetylation until the discovery of histone demethylases (Shepherd, Hardin et al. 1971; Byvoet, Shepherd et al. 1972; Rice and Allis 2001). Two different types of histone demethylases have been discovered to date, lysine specific demethylase 1 (LSD1) and Jumonji domain containing proteins (JmjC) (Shi, Lan et al. 2004; Tsukada, Fang et al. 2006). LSD1 demethylates mono and di- methylated lysine residues only whereas the JmjC proteins can demethylate trimethylated lysine and mono- and di-methylated arginine residues as well. Although, these two classes of demethylases have different mechanism of action, they have been shown to coordinate functionally in demethylating their substrates (Wissmann, Yin et al. 2007).

All the lysine and arginine methylation sites and the respective methyltransferases and demethylases that have been characterized to date are tabulated in Table 1.1 and Table 1.2 (adapted from *histome.org*: the histome infobase)

1.3.1. Arginine methylation

Arginine methylation is carried out by PRMTs, which are classified into two classes: type 1 methyltransferases (PRMTs 1-4, 6 and 8) that catalyze the asymmetrical methylation of arginine residues (one methyl group on each nitrogen atom) and type 2 methyltransferases (PRMTs 5, 7 and 9) that catalyze the symmetrical methylation of arginine residues (both methyl groups on the same nitrogen atom) as shown in Figure 1.2 (Wysocka, Allis et al. 2006). Several PRMTs have been shown to be involved in proper functioning of different cellular processes such as transcriptional activation (PRMT1), activation of nuclear androgen receptors (PRMTs 2 and 5) (Qi, Chang et al. 2002; Bedford and Richard 2005), regulation of p53 transcription (PRMT 5) (Jansson, Durant et al. 2008), RNA maturation (Cheng, Cote et al. 2007) and DNA repair (Higashimoto, Kuhn et al. 2007).

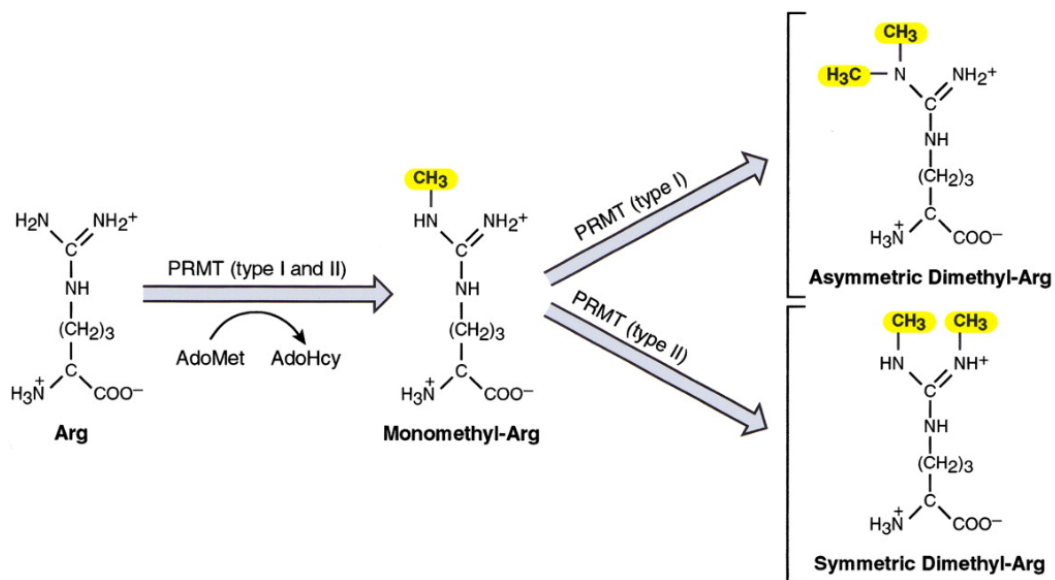


Figure 1.2: Mechanism of action of arginine methylation by PRMTs

PRMTs (Type 1 and 2) catalyze the transfer of methyl group (CH₃) from AdoMet (SAM) onto the NH₂⁺ group of the arginine residue. Type 1 PRMTs catalyze mono-methylation and the asymmetric di-methylation of arginine, whereas type 2 PRMTs catalyze mono-methylation and symmetric-dimethylation of arginine.

Figure adapted from (Zhang and Reinberg 2001)

1.3.2. Lysine methylation

Lysine can undergo be mono-, di- and tri-methylation (Figure 1.3). The most studied histone lysine methylation events include H3K4, H3K9, H3K27, H3K36, H3K79 and H4K20. Each of the methylation states, mono-, di and tri-methylation is associated with a distinct transcriptional readout (Santos-Rosa, Schneider et al. 2002; Wang, An et al. 2003). Histone methylation marks are highly specific; based on the residue and extent of its methylation different downstream effector proteins are recruited which lead to either relaxation or compaction of chromatin. For example, trimethylation of H3K4, H3K36 and H3K79 is shown to be associated with transcriptionally active sites whereas methylation of H3K9, H3K27 and H4K20 is associated with transcriptionally repressed sites on the chromosomes (Santos-Rosa, Schneider et al. 2002; Schotta, Lachner et al. 2004; Martin and Zhang 2005). This specificity of the methylation can help in identifying euchromatin and heterochromatin regions on the chromosome.

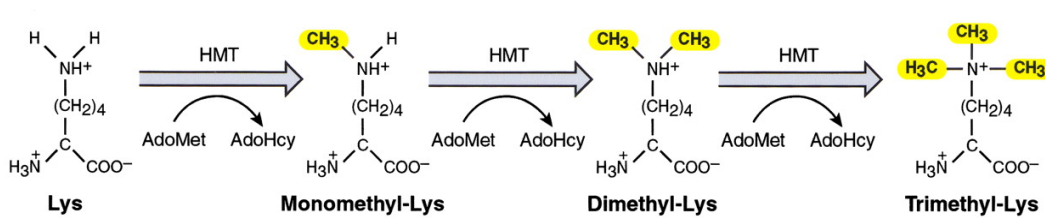


Figure 1.3: Mechanism of action of lysine methylation by HMTs

HMTs catalyze the transfer of methyl group (CH₃) group from AdoMet (SAM) to the NH₃ group at the amino terminal of lysine. The figure shows step-by-step mechanism of mono-, di- and trimethylation of lysine. Figure adapted from (Zhang and Reinberg 2001)

Sites of histone lysine methylation		
Site of modification	Writers	Erasers
H1K186me1	Histone-lysine N-methyltransferase, H3 lysine-9 specific 3, Histone-lysine N-methyltransferase, H3 lysine-9 specific 5	
H1K25me1	EZH2, Histone-lysine N-methyltransferase, H3 lysine-9 specific 3, Histone-lysine N-methyltransferase, H3 lysine-9 specific 5	Lysine-specific demethylase 4D
H2BK5me1		
H3K27me1	EZH1 EZH2, Histone-lysine N-methyltransferase, H3 lysine-9 specific 3, Histone-lysine N-methyltransferase, H3 lysine-9 specific 5	
H3K27me2	Histone-lysine N-methyltransferase EZH1, Histone-lysine N-methyltransferase EZH2, Histone-lysine N-methyltransferase NSD3	Histone lysine demethylase PHF8, Lysine-specific demethylase 6B, Lysine-specific demethylase 7
H3K27me3	EZH2 NSD3	Lysine-specific demethylase 6A, Lysine-specific demethylase 6B
H3K36me1	Probable histone-lysine N-methyltransferase ASH1L	
H3K36me2	SETMAR, Histone-lysine N-methyltransferase, H3 lysine-36 and H4 lysine-20 specific, N-lysine methyltransferase SMYD2, Probable histone-lysine N-methyltransferase ASH1L	Lysine-specific demethylase 2A, Lysine-specific demethylase 2B, Lysine-specific demethylase 8
H3K36me3	SETD2 Probable histone-lysine N-methyltransferase NSD2	Lysine-specific demethylase 4A, Lysine-specific demethylase NO66
H3K4me1	SETD7	Lysine-specific histone demethylase 1A
H3K4me2	NSD3	Lysine-specific demethylase 5A, Lysine-specific demethylase 5D, Lysine-specific histone demethylase

		1A
H3K4me3	MLL, MLL3, MLL4, PRDM9, SETD1A, SETD1B, SET and MYND domain-containing protein 3	Histone lysine demethylase PHF8, Lysine-specific demethylase 2B, Lysine- specific demethylase 5A, Lysine-specific demethylase 5B, Lysine-specific demethylase 5C, Lysine- specific demethylase 5D, Lysine-specific demethylase NO66
H3K79me1	Histone-lysine N- methyltransferase, H3 lysine-79 specific	
H3K79me2	Histone-lysine N- methyltransferase, H3 lysine-79 specific	
H3K79me3	Histone-lysine N- methyltransferase, H3 lysine-79 specific	
H3K9me1	Histone-lysine N- methyltransferase, H3 lysine-9 specific 3	Lysine-specific demethylase 3A, Lysine-specific demethylase 3B, Lysine- specific histone demethylase 1A
H3K9me2	Histone-lysine N- methyltransferase, H3 lysine-9 specific 3, Histone-lysine N- methyltransferase, H3 lysine-9 specific 5, PR domain zinc finger protein 2	Histone lysine demethylase PHF8, Lysine-specific demethylase 3A, Lysine- specific demethylase 3B, Lysine-specific demethylase 4C, Lysine-specific demethylase 4D, Lysine- specific demethylase 7, Lysine-specific histone demethylase 1A, Lysine- specific histone demethylase 1B
H3K9me3	SETDB1 SETDB2 SUV39H1 SUV39H2	Lysine-specific demethylase 4A, Lysine-specific demethylase 4B, Lysine- specific demethylase 4C, Lysine-specific demethylase 4D

H4K20me1	SETD8, Probable histone-lysine N-methyltransferase NSD2	Histone lysine demethylase PHF8
H4K20me2	SETD8, SUV420H1, SUV420H2, H3 lysine-36 and H4 lysine-20 specific	
H4K20me3	Histone-lysine N-methyltransferase SUV420H1, Histone-lysine N-methyltransferase SUV420H2, Probable histone-lysine N-methyltransferase NSD2	

Table 1.1: Sites of histone lysine methylation and their respective methyltransferases and demethylases

Sites of histone arginine methylation		
Site of modification	Writers	Erasers
H2AR3me2	Protein arginine N-methyltransferase 6	
H3R17me1	CARM1	
H3R17me2	CARM1	
H3R26me1	CARM1	
H3R2me1	CARM1	
H3R2me2	Protein arginine N-methyltransferase 6	Bifunctional arginine demethylase and lysyl-hydroxylase JMJD6
H3R8me2	Protein arginine N-methyltransferase 5	
H4R3me1	Protein arginine N-methyltransferase 1	Bifunctional arginine demethylase and lysyl-hydroxylase JMJD6
H4R3me2	Protein arginine N-methyltransferase 1, Protein arginine N-methyltransferase 5, Protein arginine N-methyltransferase 6	Bifunctional arginine demethylase and lysyl-hydroxylase JMJD6

Table 1.2: Histone arginine methylation sites and their respective methyltransferases and demethylases

Histone lysine methyltransferases and their links to diseases		
Lysine methyltransferase	Links to disease	Histone target sites
Suv39h1 (KMT1A)	Elevated mRNA level in colon cancer tissue samples ⁹⁴	H3K9
G9a (KMT1C)	Suppressor gene silencing	H3K9
Eu-HMTase1 (KMT1D)	Overexpression in gland tumors	H3K9
MLL1 (KMT2A)	Rearrangement/amplification blocks hematopoietic differentiation	H3K4
MLL4 (KMT2D)	Involved in hepatitis B virus dependent liver carcinogenesis	H3K4
SMYD2 (KMT3C)	Suppression of p53 transcriptional activity	H3K56, p53K370
SMYD3	Overexpression and enhanced tumor cell growth in breast cancer Overexpression in colon cancer and hepatocellular carcinoma	H3K450
DOT1L (KMT4)	Enzymatic activity crucial for leukaemogenesis	H3K79
EZH2 (KMT6)	Amplification and overexpression in multiple cancer types	H3K27
	Marker for precancerous state in breast carcinogenesis	H3K27
	Marker for aggressive breast cancer	H3K27
	Promotes proliferation and invasiveness of prostate cancer cells	H3K27
	Useful as a biomarker for poor prostate cancer prognosis	H3K27
SET7/9 (KMT7)	Hyperglycemia induces SET7/9 causing increased p65 gene expression	H3K4, p53K37226
SETDB1/ESET (KMT1E)	Cooperates with DNA methyltransferase in promoter silencing in tumors	
SET8/PR-SET7 (KMT5A)	Suppresses p53 dependent transcription	

Table 1.3: Histone lysine methyltransferases and link to diseases

Histone Arginine methyltransferases and their links to diseases		
Arginine methyltransferase	Links to disease	Histone target sites
PRMT1	Essential component of MLL oncogenic transcriptional complex Coactivator of hormone receptors in hormone dependent cells	H4R3
PRMT2	Coactivator of the androgen receptor	PRMT2
PRMT4 (CARM1)	Essential for estrogen-induced cell cycle progression in breast cancer cells Aberrant expression in prostate tumors Methylation of CBP contributes to coactivation CARM1 knockdown blocks androgen receptor signaling	H3R2 H3R17 H3R26
PRMT5.	Downregulates the expression of tumor suppressor genes in fibroblasts	H3R8 H4R3
PRMT6	Overexpression diminished viral Tat activity	H3R2 H4R3 H2AR3
PRMT7	Downregulation sensitizes cancer cells to camptothecin treatment	

Table 1.4: Histone arginine methyltransferases and their links to disease

1.4. H3K9 methylation and its significance

H3K9 methylation is performed by Suv39h1, Suv39h2, G9a (EHMT1), GLP (G9a like protein) (EHMT2), and ESET1/SETDB1 in humans. Orthologues of these genes in other organisms such as Su(var)3-9, dG9a and dSETDB1 in *Drosophila* (Tachibana, Sugimoto et al. 2001; Schotta, Ebert et al. 2002; Tachibana, Ueda et al. 2005; Fritsch, Robin et al. 2010) and Clr4 in yeast also methylate H3K9. These methyltransferases have different affinities for unmethylated and monomethylated H3K9 and produce different H3K9 methylated states which have a distinct downstream effect. Suv39h1, Suv39h2 and SETDB1 are trimethylases, whereas G9a and GLP are involved in mono- and di-methylation of H3K9 (Tachibana, Sugimoto et al. 2001).

H3K9 methylation has been implicated in regulation of gene expression, heterochromatin formation, X-chromosome inactivation (Heard, Rougeulle et al. 2001), proper functioning of the cell cycle and chromosome segregation (Melcher, Schmid et al. 2000; Heit, Rattner et al. 2009). Trimethylated H3K9 is a hallmark of constitutive heterochromatin at pericentromeric and telomeric regions, whereas mono and di-methylated H3K9 are primarily found in euchromatin regions showing that each type of methylation mark occupies a distinct position in the genome and is involved in distinct functions (Rea, Eisenhaber et al. 2000; Peters, Kubicek et al. 2003). Methylation of H3K9 regulates the higher order organization of chromatin in the genome. In particular, H3K9me3 regulates the chromatin structure by creating a binding site for heterochromatin protein 1 (HP1), a structural protein enriched in heterochromatin,

which binds to the methylated site via its chromodomain (Aagaard, Laible et al. 1999; Bannister, Zegerman et al. 2001; Lachner, O'Carroll et al. 2001).

Heterochromatin assembly at the centromere via the H3K9me3 mediated pathway has been shown to be essential for the proper kinetochore formation and sister chromatid cohesion (Bernard and Allshire 2002). Cells lacking H3K9me3 at the pericentric heterochromatin exhibit a wide range of mitotic defects such as misalignment of chromosomes at metaphase, non-disjunction of chromosomes during anaphase and lagging chromosomes during cytokinesis (Peters, O'Carroll et al. 2001; McManus, Biron et al. 2006).

1.5. *SUV39h1* (Suppressor of variegation 3-9 homolog 1)

SUV39h1 is a human homolog of the *Drosophila* suppressor of variegation, *SU(VAR)3-9*, which is a suppressor of position effect variegation (PEV) phenomenon. PEV is a phenomenon in which variegation is caused by the inactivation of a gene in some cells through its abnormal juxtaposition with heterochromatin. The classical example of PEV is the *Drosophila* white-mottled-4 translocation in which an inversion on the X chromosome places the white gene that is responsible for the red eye phenotype next to the pericentric heterochromatin. This leads to mottled pattern color in the *Drosophila* eye. *SUV39h1* was found to be a suppressor of this phenomenon in the genetic screens carried out in *Drosophila* to identify the modifiers of PEV (Reuter and Spierer 1992). *Su(var)3-9* protein in *Drosophila* was shown to localize strongly to the heterochromatic regions on the chromosome and is associated with repressive

chromatin, however its function was not known (Jenuwein, Laible et al. 1998). Subsequent studies led to the characterization of Su(var)3-9, Suv39h1 and Clr4 (homologue of Su(var)3-9 in yeast) as histone methyltransferases with high specificity towards H3K9 (Rea, Eisenhaber et al. 2000). Suv39h1 catalyzes the transfer of a methyl group from S-Adenosyl methionine (SAM), which is the principal methyl donor in cells, to the H3K9 residue (Rea, Eisenhaber et al. 2000) (Loenen 2006).

Suv39h1 is a 48kDa protein that is located on the X-chromosome. It contains a N-terminal chromodomain and a C-terminal SET domain (Su(var), Enhancer of Zeste and Trithorax) that is flanked by pre-SET domain and cysteine rich post-SET domain. The SET domain was shown to be the catalytic domain responsible for the methyltransferase activity in Suv39h1 and other SET domain containing methyltransferases (Rea, Eisenhaber et al. 2000; Trievel, Beach et al. 2002; Zhang, Tamaru et al. 2002; Wu, Min et al. 2010). The pre-SET and post-SET domains are required for its methyltransferase activity and stable binding to heterochromatin protein 1 (Krouwels, Wiesmeijer et al. 2005).

Mammals have two homologues of the Su(var)3-9 gene, Suv39h1 and Suv39h2, which show overlapping expression profiles during embryogenesis. In adults however, Suv39h1 is expressed ubiquitously while Suv39h2 is expressed only in testes (O'Carroll, Scherthan et al. 2000). Both homologues have been shown to be involved in H3K9 methylation, leading to the conclusion that they may have a redundant role especially during embryonic development. This is further strengthened by the observation that knocking out either Suv39h1 or

Suv39h2 does not affect viability or fertility in mice. However, knocking out both the Suv39 homologues in mice severely impaired their viability, with only 33% survival rates, genomic instability, increased risk of tumourigenesis and a significant reduction in the size of the mice compared to wild type (Peters, O'Carroll et al. 2001). Mouse embryonic fibroblasts derived from these mice show a very high rate of chromosome missegregation leading to genomic instability, which is a hallmark of cancer. This demonstrates the importance of Suv39h enzymes for proper segregation of chromosomes to the daughter cells in mitosis.

1.5.1. Suv39h1 and its role in mitosis

The localization of Suv39h1 is dynamic during mitosis. It localizes along the arms of the chromosomes in prophase and pro-metaphase and is located only at the centromere at metaphase; it disappears from the centromere at the onset of anaphase (Aagaard, Schmid et al. 2000). In a normal cell, the centromere is surrounded by heterochromatin that is enriched in H3K9me3 that is critical for the assembly of kinetochore during mitosis (refer Figure 1.4).

The kinetochore is the most important protein structure responsible for proper functioning of mitosis. The kinetochores are multi-proteinaceous complexes and are estimated to contain around 200 proteins (Ohta, Bukowski-Wills et al. 2010) that form the point of attachment to the microtubules emanating from the spindle poles during mitosis. The sister chromatids attain bipolarity by attaching to the microtubules emanating from the spindle poles during late prophase and align themselves at the metaphase plate during metaphase. Once all

the chromosomes are aligned at the metaphase plate, the sister chromatids separate and migrate to the opposite poles during anaphase. To prevent the premature separation of the sister chromatids that leads to missegregation of chromosomes and aneuploidy, the cell has a surveillance mechanism in place, called the mitotic checkpoint, which is on until all the chromosomes attain bipolarity and align at the metaphase plate (Hartwell 1992; Musacchio and Salmon 2007). However, cells lacking Suv39h1 do not contain H3K9me3 at the pericentric heterochromatin, and therefore are prone to improper kinetochore assembly leading to missegregation of chromosomes and aneuploidy, which is a hallmark of cancer.

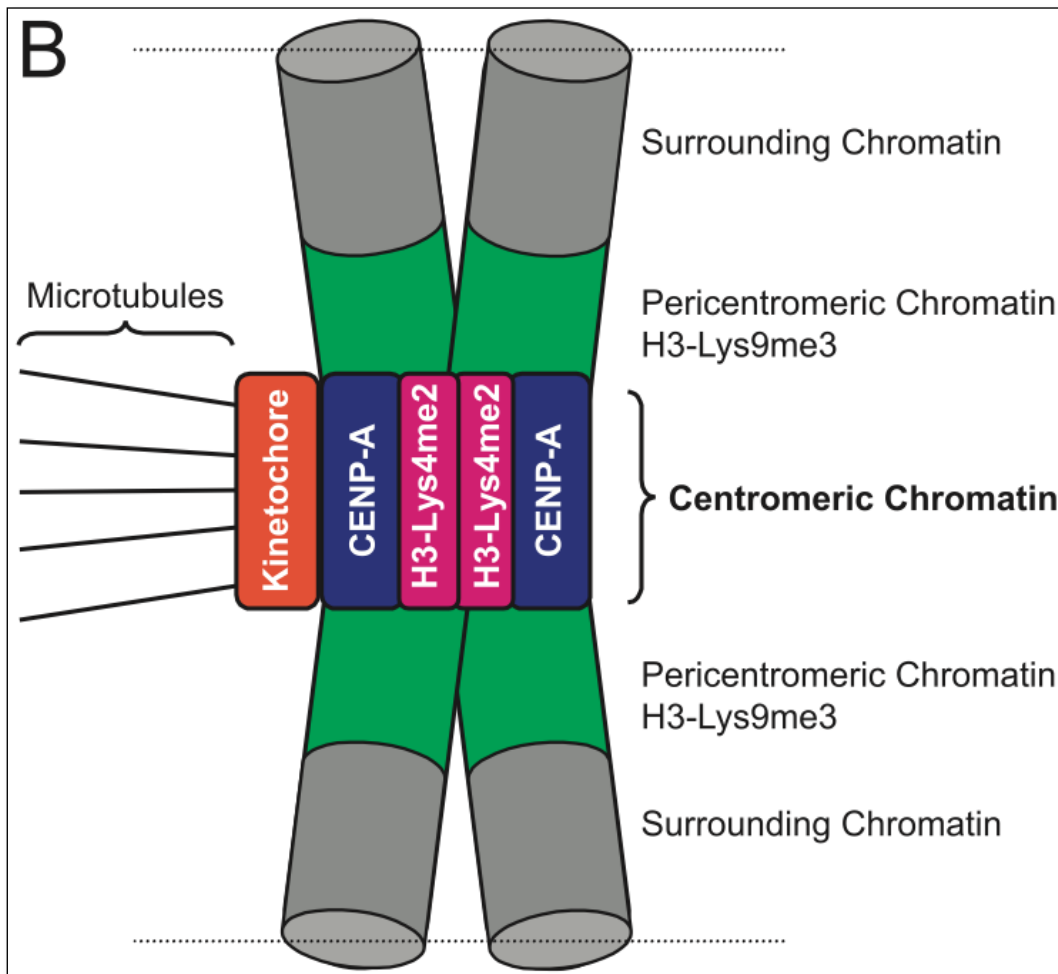


Figure 1.4: Organization of Histone H3 modifications at the centromere of a metaphase chromosome

In metaphase chromosomes, the H3K9me3 is enriched at the regions surrounding the centromere (pericentromeric heterochromatin). H3K4me2 is located at the inner centromere region and CENP-A nucleosomes are located at the outer surface facing towards the spindle poles. The kinetochore assembles on the CENP-A nucleosomes and forms the point of contact for microtubules to attach to the sister chromatids. Figure adapted from (Vos, Famulski et al. 2006)

1.5.2. Suv39h1 and cancer

Fibroblasts derived from Suv39h1/h2 double knockout mice embryos lack H3K9 methylation at the pericentric regions and are genetically unstable and have higher numbers of chromosomes compared to wild type fibroblasts. This seems to suggest that trimethylation of H3K9 by Suv39h1 protects cells from genomic instability, which is a critical process in tumorigenesis and also a hallmark of cancer. Consistently, the double null mice were shown to develop late onset B-cell lymphoma that is characterized by a similar phenotype to that of non-Hodgkin lymphoma in humans (Peters, O'Carroll et al. 2001).

In contrast to the earlier knock out experiments, overexpression of Suv39h1 in HeLa cells was shown to cause ectopic heterochromatin formation and severe chromosome segregation defects such as lagging chromosomes and formation of chromosomal bridges during mitosis (Melcher, Schmid et al. 2000). These cells also showed a high incidence of poly- and micronuclei formation compared to the control cells. Both, knock out and overexpression of Suv39h1 results in defects in mitosis, thus leading the cells to a cancerous phenotype, suggesting that it is necessary to maintain a delicate balance in the levels of Suv39h1 in the cell in order for the cell to function properly. Misregulation of Suv39h1 has been implicated in the formation of B-cell lymphoma in mice (Peters, O'Carroll et al. 2001) colo-rectal cancer (Kang, Lee et al. 2007) and Acute Myeloid Leukemia (AML) (Chakraborty, Sinha et al. 2003) in humans.

Suv39h1 has also been implicated in tumorigenesis due to its interaction with Retinoblastoma protein (Rb) during the cell cycle (Vandel, Nicolas et al.

2001). It has been shown that Suv39h1 may interfere in the proper functioning of the cell cycle by mediating Rb associated repression of cell cycle control genes. Suv39h1 and HP1 have been implicated in the repression of E2F regulated transcription of cyclin E by methylating its promoters (Nielsen, Schneider et al. 2001).

1.6. Histone methyltransferases and cancer

Of the approximately 48 histone methyltransferases that methylate different lysine and arginine residues at varying degrees that have been identified and characterized in human cells to date, 22 of them have been implicated in cancer or other diseases in humans or in mouse models (Table 1.3 and Table 1.4) (Qian and Zhou 2006; Wolf 2009; Albert and Helin 2010). The HMTs are highly specific in methylating the residues on the histone tails, thus reflecting the importance of these epigenetic marks (Huang and Berger 2008).

With the discovery of histone demethylases, the transient nature of the methyl marks was confirmed (Shi, Lan et al. 2004). There exists a delicate balance between the methylated and demethylated states of histone tails that is essential for the proper functioning of cellular processes that each epigenetic mark is responsible for. Therefore it is essential that expression, activity and recruitment of these methyltransferases and demethylases be tightly regulated as any misregulation, such as overexpression of the methyltransferases leading to methylation and thereby repression of tumor suppressor genes, or overexpression of demethylases leading to demethylation and thereby activation of oncogenes

that are usually repressed could lead to cancer (Nguyen, Weisenberger et al. 2002; Kim and Huang 2003).

Although many of the histone methyltransferases have been implicated in cancer, the mechanisms by which the histone methyltransferases are involved in cancer initiation or progression are still not understood completely and is currently a very exciting field of research.

1.7. Epigenetic modifiers as targets for cancer therapy

Epigenetic modifications are dynamic and reversible. Therefore epigenetic errors unlike genetic mutations can be reversed. This dynamic nature of epigenetic modifications can be exploited by developing drugs that can inhibit the enzymes that are responsible for the abnormal epigenetic modifications in a diseased condition. Consistent with this idea several epigenetic modifiers implicated in cancer have been targeted to discover drugs for cancer therapy. To date, four drugs targeting epigenetic modifiers have been approved by the Food and Drug Administration (FDA) and are currently being used for cancer therapy; two of them targeting DNA methyltransferases (DNMTs) and two targeting histone deacetylases (HDACs) (Table 1.5). Even though it is known that these inhibitors function by potentially restoring the normal functioning of the tumor suppressor genes that are silenced due to hypermethylation by DNMTs and deacetylation by HDACs respectively, the precise molecular mechanism of action of these drugs is still being investigated (Karagiannis and El-Osta 2006; Yang, Doshi et al. 2006; Carew, Giles et al. 2008). Several other new and improved inhibitors targeting

other epigenetic modifiers are currently under investigation and are undergoing clinical trials.

Drug	Epigenetic target	Type of cancer	Reference
Virinostat	HDAC	Advanced cutaneous T-cell lymphoma	(Duvic and Vu 2007)
Depsipeptide	HDAC	Cutaneous T-cell lymphoma	(Piekarz, Frye et al. 2009)
Decitabine	DNMT	Myelodysplastic syndromes, Acute Myeloid Leukemia	(Wijermans, Lubbert et al. 2005)
Azacitidine	DNMT	Myelodysplastic syndromes	(Fenaux, Mufti et al. 2009)

Table 1.5: FDA approved drugs that inhibit epigenetic mechanisms involved in cancer

Virinostat and depsipeptide are histone deacetylase inhibitors that have been approved by the FDA for the treatment of cutaneous T-cell lymphoma. Decitabine and azacitidine are DNA methyltransferase inhibitors that have been approved by the FDA for the treatment of myelodysplastic syndromes.

1.7.1. Histone methyltransferases as targets for cancer therapy

As shown in Table 1.3 and Table 1.4 HMTs are involved in several different types of cancers. In cases where overexpression of HMTs leads to silencing of tumor suppressor genes, HMT inhibitors could be potentially used to inhibit the activity of these enzymes and restore the normalcy to the cancerous cell, in a way much similar to DNMT inhibitors and HDAC inhibitors. The small molecule inhibitors that have been discovered to inhibit lysine HMT activity are briefly described below and their structures are shown in Table 1.6

1.7.1.1. Chaetocin

Chaetocin is a fungal mycotoxin that was found to inhibit Su(var)3-9 selectively with an IC_{50} (half maximal inhibitory concentration) of $0.6\mu\text{M}$, human Suv39h1 at $IC_{50}=0.8\mu\text{M}$; G9a and DIM5 were inhibited with an IC_{50} of $2.5\mu\text{M}$ and $3\mu\text{M}$ (Greiner, Bonaldi et al. 2005). It was shown to be rather selective towards Suv39h1 when tested against other methyltransferases such as EZH2 ($IC_{50}>90\mu\text{M}$) and SET7/9 ($IC_{50}>180\mu\text{M}$). Chaetocin was shown to be a competitive inhibitor suggesting that it binds to the same binding pocket of SAM. A major drawback of chaetocin is that it displays high cellular toxicity irrespective of its methyltransferase inhibitory activity. However, the specificity of chaetocin towards Suv39h1 has been challenged recently and it has been shown to be a general-purpose histone methyltransferase inhibitor (Cherblanc, Chapman et al. 2013).

1.7.1.2. BIX-01294

BIX-01294 is a diazepin-quinazolinamine derivative that was discovered as a selective inhibitor towards G9a and generation of H3K9me2 (Kubicek, O'Sullivan et al. 2007). It inhibits G9a with an IC₅₀ value of 1.7µM and was shown to be a non competitive inhibitor of SAM. It was found to be a good inhibitor of H3K9me2 at lower concentrations but toxic at concentrations higher than 4.1µM.

1.7.1.3. SAM analogues

Since methyltransferases use SAM as a methyl donor, analogues of SAM were one of the methyltransferase inhibitors to be used against both, DNMTs and HMTs. These include methylthioadenosine, S-adenosyl homocysteine and the bacterial metabolite sinefungin. Although, these molecules have been shown to inhibit methylation, by their nature, they inhibit all SAM-dependent enzymes and therefore cannot be used for therapeutic purposes.

1.7.1.4. Deazaneplanocin (DZNep)

DZNep is a cyclopentenyl analog of 3-deazaadenosine. It was shown to reduce H3K9 trimethylation by specifically targeting EZH2 levels in primary AML cells by inducing apoptosis (Tan, Yang et al. 2007). However, the specificity of DZNep has been challenged in further studies (Miranda, Cortez et al. 2009).

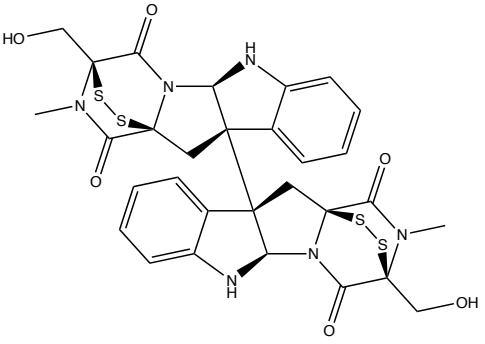
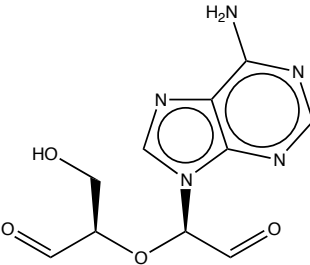
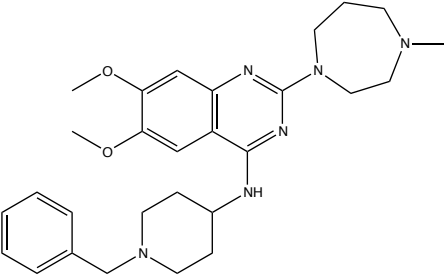
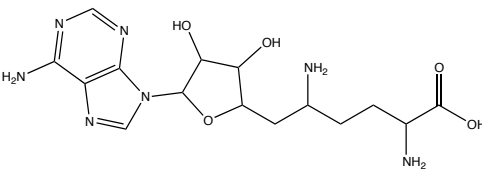
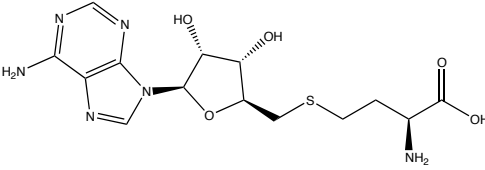
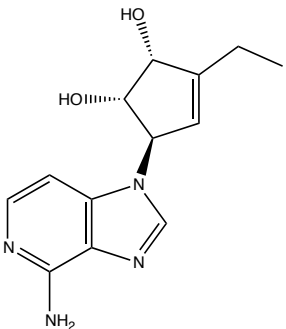
 <p>Chaetocin</p>	 <p>Adox</p>
 <p>BIX-01294</p>	 <p>Sinefungin</p>
 <p>S-Adenosyl Homocysteine</p>	 <p>DZNep</p>

Table 1.6: Structural representation of the small molecule inhibitors of various histone methyltransferases

1.8. Research focus of this thesis

Suv39h1 is a H3K9 methyltransferase that plays a crucial role in mitosis and its misregulation has been implicated in cancer as described earlier. Based on its involvement in different types of cancer, it is an attractive target to develop small molecule inhibitors. Apart from being used as therapeutic agents, small molecule inhibitors are excellent research reagents that can be used to decipher mechanism of action of Suv39h1 in the cell. The effects of inhibition of its activity on a signaling pathway in real time and/or identification of its interaction partners independent of its catalytic activity can be determined by using small molecule inhibitors.

The major focus of this thesis is to discover small molecule inhibitors (SMIs) that are specific to H3K9 trimethylase Suv39h1 and validate them *in vitro*. These SMIs can be used to study the effects of inhibiting Suv39h1 activity in the various pathways that it has been implicated in during mitosis; and also be used as lead molecules that can be developed into therapeutic drugs for cancer therapy. During the course of this project, since the crystal structure of Suv39h1 was not yet solved, a homology model of Suv39h1 was constructed, optimized and validated. The binding sites of three ligands that are known to bind to Suv39h1, i.e. chaetocin, S-Adenosyl methionine and the N-terminal of the histone H3 peptide on the homology model were determined by blind docking and structural analysis. These sites were used for screening several small molecule libraries by virtual screening to identify small molecules that bind with high binding affinity to these sites and could potentially inhibit the methyltransferase activity of

Suv39h1. The top hits from virtual screening were tested by an *in vitro* methyltransferase assay.

1.9. Hypothesis

A homology model of H3K9 methyltransferase Suv39h1 can be built; and it can be used to discover small molecule inhibitors to inhibit its activity. The inhibitors that are discovered will be used as anti-mitotics in combination with other anti-mitotic drugs available. They could also prove to be excellent lead compounds that can be used to develop novel drugs. These drugs can be used in combination with other epigenetic therapies to reactivate epigenetically silenced genes involved in growth control.

1.10. Homology modeling

Homology modeling is the process of predicting the 3-D structure of a protein based on its sequence similarity with other proteins. This is based on the general observation that proteins with similar amino acid sequences, usually evolutionarily related, have similar 3-D structures as well (Chothia and Lesk 1986). Therefore, by using this characteristic feature of related proteins, the 3-D model of a protein of interest (target protein) can be designed based on the known structures of related proteins treated as templates. This is especially useful in the case of a family of proteins, where if the structure of at least one member of the family is determined experimentally, then the structures of the others can be modeled based on their alignment to the known structure.

1.10.1. Why do we need homology modeling?

Having the knowledge of the three-dimensional (3-D) structure of a protein is critical for answering several biological questions. Traditionally protein structures are determined by X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy. However, determination of 3-D structures of proteins using these experimental methods is a tedious and time-consuming task and cannot be applied in some cases for a variety of reasons. Therefore, the number of proteins with 3-D structures remains extremely low compared to the number of unique annotated protein sequences. With the development of modern techniques there has been a gradual increase in the number of 3-D structures determined, but progress in this field is quite slow compared to the pace at which new genes and genomes are

being sequenced. As of January 2013, the number of manually annotated and reviewed protein sequences stood at 538849, whereas the number of structures in the Protein Data Bank (PDB) is only 87838 as shown in Figure 1.5.

Considering the technical difficulties and time associated with experimental structure determination, it is imperative to find alternative methods, which are easier, reliable and able to generate structural data comparable to experimentally derived structures. Computational methods offer a relatively easier and useful option for the determination of protein structures in cases where structure determination by existing techniques fail. Protein structures can be computationally predicted either from their sequence alone (*ab-initio*) or by comparative methods (homology modeling) that rely on the protein structure databases that comprise known protein structures (Baker and Sali 2001; Bonneau and Baker 2001). The *Ab-initio* method is a template free protein structure prediction system whereas homology modeling is a template based protein structure prediction system. Relatively speaking, homology modeling has been shown to yield better and more reliable models compared to *ab-initio* methods, mainly due to the limitations in our current understanding of protein folding which is crucial for achieving high accuracy with *ab-initio* methods (Jauch, Yeo et al. 2007).

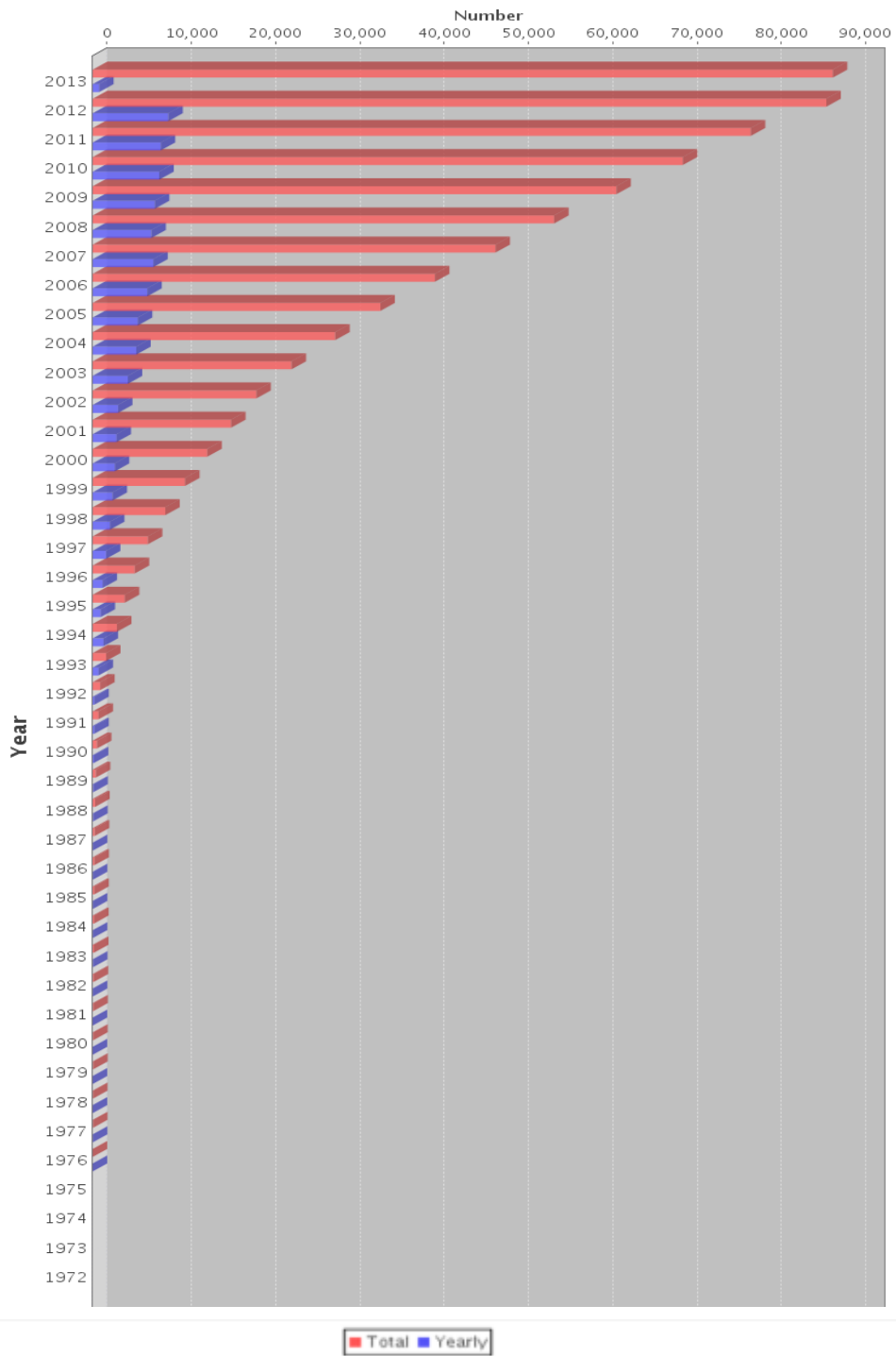


Figure 1.5: Yearly growth in the number of protein structures solved and submitted to the protein databank.

Figure adapted from: www.pdb.org/pdb/statistics/

Homology modeling provides us with an alternative strategy to carry ahead the research on a target if it is stalled due to difficulties in structure determination, especially in the fields such as inhibitor discovery and identification of binding partners. Drug discovery against a protein by virtual screening uses structural knowledge about the protein of interest, making it is imperative to have the 3D structure of the protein to perform structure based virtual screening.

1.10.2. Steps involved in homology modeling

Apart from the commercial software packages, the advent of several web servers that require only a target sequence as an input to predict homology models has made the process of homology modeling considerably easier. Although this might seem very simple at the outset, the results obtained from this type of software and webservers must be analyzed critically before proceeding ahead to use it for any purpose. It is important to note that each web server uses different algorithms and the results may vary especially in cases where the target and template share very low sequence identity. A list of software and webservers that are currently available for building homology models are tabulated in Table 1.7.

The process of homology modeling involves three steps: a) Identification of homologues/templates with known structure by sequence alignment; b) Initial back bone modeling by using the target-template sequence alignment; c) Building a 3D model including loops and side chains. Once the side chains are packed, the model is then evaluated for probable modeling errors in any of the previously

mentioned steps. These steps can be repeated until a satisfactory model of the target protein is obtained (Figure 1.6).

Sequence alignment tools	
FFAS03	ffas.ljcrf.edu/ffas-cgi/cgi/ffas.pl
BLAST/PSI-BLAST	http://blast.ncbi.nlm.nih.gov/Blast.cgi
PROMALS3D	http://prodata.swmed.edu/promals3d/promals3d.php
T-COFFEE	http://tcoffee.crg.cat/apps/tcoffee/index.html
PROBCONS	http://toolkit.tuebingen.mpg.de/probcons
MUSCLE	http://www.drive5.com/muscle/
ClustalW2	http://www.ebi.ac.uk/Tools/msa/clustalw2/
FastA/SSEARCH	http://www.ebi.ac.uk/Tools/sss/fasta/
Fold recognition by threading	
PHYRE	http://www.sbg.bio.ic.ac.uk/~phyre/
FUGUE	http://tardis.nibio.go.jp/fugue/prfsearch.html
LOMETS	http://zhanglab.ccmb.med.umich.edu/LOMETS/
LOOPP	cbsuapps.tc.cornell.edu/
MUSTER	http://zhanglab.ccmb.med.umich.edu/MUSTER/
PSIPRED	http://bioinf.cs.ucl.ac.uk/psipred/
SAM-T08	http://compbio.soe.ucsc.edu/SAM_T08/T08-query.html
Comparative modeling, loop and side chain modeling	
PUDGE	http://bhapp.c2b2.columbia.edu/pudge/cgi-bin/pipe_int.cgi
3D-JIGSAW	http://bmm.cancerresearchuk.org/~3djigsaw/
M4T	http://manaslu.aecom.yu.edu/M4T/
MMM	http://manaslu.aecom.yu.edu/MMM/
RAPPER	http://mordred.bioc.cam.ac.uk/~rapper/
WHATIF	http://swift.cmbi.ru.nl/whatif/
SWISS-MODEL	http://swissmodel.expasy.org/workspace
ESYPRED3D	http://www.fundp.ac.be/sciences/biologie/urbm/bioinfo/esypred/
I-TASSER	http://zhanglab.ccmb.med.umich.edu/I-TASSER/
MODWEB	modbase.compbio.ucsf.edu/ModWeb20-html/modweb.html
PCONS	pcons.net
HHPRED	toolkit.tuebingen.mpg.de/hhpred
CPH-MODELS	www.cbs.dtu.dk/services/CPHmodels/
MODELLER	www.salilab.org/modeller/modeller.html
Loop modeling	
ARCHPRED	http://manaslu.aecom.yu.edu/loopred/
MODLOOP	http://modbase.compbio.ucsf.edu/modloop/

Side chain modeling	
IRECS	http://irecs.bioinf.mpi-inf.mpg.de/index.php
SCWRL4	http://dunbrack.fccc.edu/scwrl4/index.php
Model evaluation	
ANOLEA	http://protein.bio.puc.cl/anolea/index.html
PROCHECK	http://www.ebi.ac.uk/thornton-srv/software/PROCHECK/
PROQ	http://www.sbc.su.se/~bjornw/ProQ/ProQ.html
Prosa	https://prosa.services.came.sbg.ac.at/prosa.php
Sub-AQUA	http://kiharalab.org/SubAqua/
VERIFY3D	http://nihserver.mbi.ucla.edu/Verify_3D/
WHATCHECK	http://swift.cmbi.ru.nl/gv/whatcheck/

Table 1.7: List of servers and software available to perform sequence alignment, homology modeling and model evaluation.

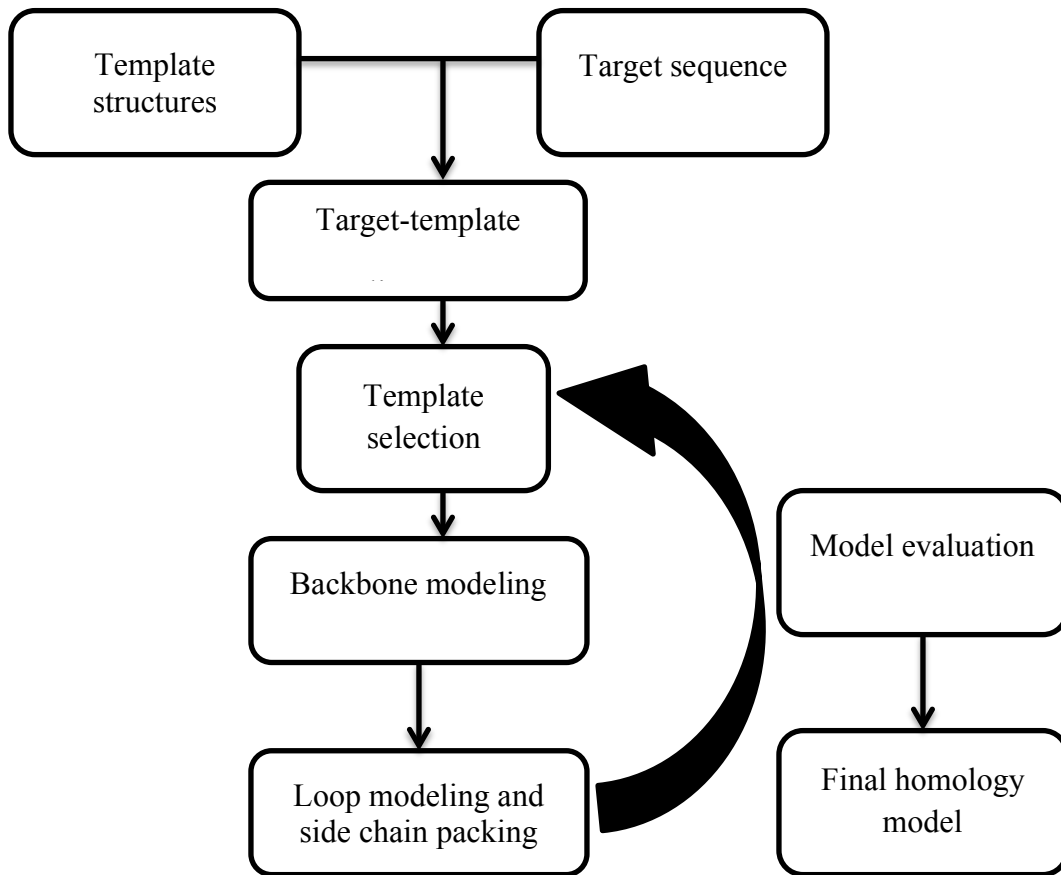


Figure 1.6: Steps involved in homology modeling

1.10.2.1. Template identification / Sequence alignment

Homology model prediction depends on the similarity of structures between homologous proteins; therefore it is crucial to identify a good template structure based on sequence identity of the target sequence with proteins, which have known structures. Template identification is essentially done by comparing and correctly aligning the target protein sequence with other sequences in protein databases. The accuracy of the predicted model depends on the accuracy of the sequence alignment. Incorrect alignment of protein sequences is one of the major reasons for false structure predictions. Several sequence alignments programs such as BLAST (Basic Local Alignment Search Tool), PSI-BLAST (Position Specific Iterative-BLAST), CLUSTAL-W, FASTA and T-COFFEE (Tree-based Consistency Objective Function For alignment Evaluation) that use different algorithms to compare and align sequences are used for this purpose (Altschul, Gish et al. 1990; Altschul, Madden et al. 1997; Notredame, Higgins et al. 2000; Chenna, Sugawara et al. 2003). These programs vary in their ability to perform alignment. BLAST and FASTA, which make use of a substitution matrix and assign penalties for gap initiation and extension, are the most commonly used programs and are able to identify sequences with more than 30% sequence identity (Pearson and Lipman 1988; Altschul, Gish et al. 1990). The advantage of these types of alignment programs is the speed, which allows users to search and compare with hundreds of thousands of sequences very quickly. Identification of more remote homologues is done with PSI-BLAST and CLUSTAL-W, which use multiple sequence alignment and Hidden Markov Models (HMM) to make a

more accurate prediction but at a higher computational expense (more computing power and longer duration) (Edgar and Batzoglou 2006). These methods are very sensitive and extremely useful when structures homologous to the target sequence are not easily available in the database. As a rule of thumb in the homology modeling field, target-template sequence identity of >50% is considered good and the models derived from such templates have a root mean square deviation (RMSD) of $\sim 1\text{\AA}$ compared to experimentally determined structures (Chothia and Lesk 1986). Models predicted from target-template sequence alignments with <30% sequence identity fall in the “twilight zone” and are generally not reliable due to significant alignment errors and lack of templates with sufficiently close homology to the target sequence (Rost 1999; Jauch, Yeo et al. 2007).

1.10.2.2. Backbone modeling

Once the sequence alignment is done correctly, the next step in constructing a homology model is to build a backbone structure of the model. This is done by using the alpha-carbon (C- α) co-ordinates of template residues that align perfectly with the target sequence upon sequence alignment. Co-ordinates of secondary structure elements of the model, such as α -helices and β -sheets are established during this step. This method is known as rigid body replacement (Wallner and Elofsson 2005). The regions built by this method are usually the structurally conserved regions within a protein family and are important for their function. The backbone structure is easy to build when there is a high degree of sequence identity between the template and target sequence. In the case of lower sequence

identity, gaps in the backbone structure have to be filled by other methods, which leads to a less reliable homology model.

1.10.2.3. Loop modeling and side chain packing

With the determination of the backbone structure based on the aligned regions, the next step is to fill the unaligned and gap regions of the target-template alignment. These unaligned regions usually represent the loops present in the protein structure. Loops are the flexible regions that connect well-defined regions such as α -helices and β -sheets to each other. Although these regions are usually not well conserved and are difficult to be determined by X-ray crystallography, they are necessary for protein function, protein-protein interactions, and formation of ligand binding sites on the protein and are therefore important for structure based drug design (Fernandez-Fuentes, Querol et al. 2005). Loops can be modeled either by database sampling or *de-novo* methods. The database sampling method involves selecting a loop fragment based on its conformation and its sequence similarity with the target sequence (Michalsky, Goede et al. 2003; Fernandez-Fuentes, Oliva et al. 2006). The advantage of this method is that it allows rapid generation of models with physically reasonable conformations; however, it is limited by the lack of relevant loop structures from known protein structures, especially for longer loop fragments (Peng and Yang 2007). The *de-novo* method generates many random conformations of the loop sequence that fit the geometric constraints by extensively searching the conformational space. While this method is accurate for loops with 5 to 7 residues, it becomes less reliable with increasing

loop length. Increase in the length of the loop to be generated leads to a decrease in accuracy and increase in computational time, as longer sequences require sampling of larger conformational space (Rufino, Donate et al. 1996; Jacobson, Pincus et al. 2004).

After completing the backbone construction by adding loops, the next step is to arrange the side chains in their proper conformation on the modeled backbone. Proper side chain conformations are crucial for the overall 3-D structure of the protein especially for applications such as ligand docking and drug designing (Al-Lazikani, Jung et al. 2001). Each amino acid has many possible side chain conformations called rotamers, therefore side chain packing is a difficult task due to its combinatorial complexity (Vasquez 1996). The major challenge in this step is to pack the side chains in such a manner as to avoid atomic clashes between them. Using rotamer libraries that consist of empirically determined side chain conformations among known proteins solves this problem. The advantage of using these libraries is that it circumvents the combinatorial problem and greatly increases computational efficiency.

1.10.2.4. Evaluation of homology models

A homology model generated by employing the aforementioned steps may include varying degrees of errors in geometry in them. Hence, it is crucial to evaluate the initial structure obtained by the prediction software. If the model is grossly erroneous, it can be discarded by mere visual inspection. However, with the increasing sophistication of the prediction algorithms, such gross errors rarely

occur unless there is absolutely no homologous structure available to be used as a template and a forced target-template alignment has been made. If that is the case, then as per the old adage: “garbage in, garbage out”, one can expect useless model predictions.

For a reasonable model with good sequence alignment, most errors occur in the loop modeling and side chain packing steps. This is mainly because protein loops are inherently structurally unstable unlike the α -helices and β -sheets making it difficult to predict their structure. Accurate prediction of the stereochemistry of the side chains is not feasible because of the sheer number of possible positions that the side chains can be packed into on the backbone structure and high probability of clashes that can occur in between them (Fiser, Do et al. 2000; Xiang, Steinbach et al. 2007). Regardless of the scale of the errors, it is important that they are identified and rectified in order to obtain a homology model that could be most similar to its native protein structure.

1.10.3. Homology modeling using SWISS-MODEL server

The SWISS-MODEL homology modeling web server provides a workspace that integrates all the different steps in homology modeling by incorporating all the different programs required for performing each step of modeling (Peitsch 1996). SWISS-MODEL was the first automated homology-modeling server made available publicly. It can be freely accessed at <http://swissmodel.expasy.org/>. SWISS-MODEL workspace provides two modes to perform homology modeling i.e. an automated mode and an alignment mode.

The automated mode is suitable for cases with high target-template sequence identity and experimentally determined highly similar template structures are readily available (Schwede, Kopp et al. 2003; Kopp and Schwede 2004). This mode requires minimal user intervention and only the target sequence is input into the server in FASTA format or by using its UniProt accession code. The inbuilt algorithms of the program will automatically select suitable templates on the basis of a BLAST search or a HMM comparison protocol (Soding, Biegert et al. 2005).

The alignment mode is used for more distantly related target and template sequences as the number of possible errors in automated sequence alignments increases considerably in such cases (Rost 1999). In this case, an alignment of the template and target sequences is fed into the server as input. This mode allows for a more flexible approach to modeling as the user can perform and test several alternative sequence alignments based on the current theoretical and experimental knowledge of the family of proteins to which the target and template belong.

1.10.3.1. Evaluation of homology model by Swiss model server

To evaluate the quality of a predicted model, the Swiss model server calculates the empirical potential energy of the model by its inbuilt functions such as Anolea (mean force potential), QMEAN (qualitative model energy analysis), DFire and Verify3D programs (Zhou and Zhou 2002; Laskowski, Chistyakov et al. 2005). Apart from these methods, physical geometry of the model is also evaluated by other programs such as Procheck and Whatcheck, which are also part of the server

(Laskowski 1993; Hooft, Vriend et al. 1996). These functions will be described in detail below.

Anolea is a program that performs energy calculations on a protein chain to assess packing quality of the models (Melo and Feytmans 1998).

QMEAN is a composite scoring function for both the estimation of the global quality of the entire model as well as for the local per-residue analysis of different regions within a model (Benkert, Tosatto et al. 2008). QMEAN4 is an adaptation of QMEAN used by the Swiss model workspace to evaluate the global quality of the model. It provides a reliability score for the whole model that is used in order to compare and rank alternative models of the same target. The quality estimate ranges between 0 and 1 with higher values for better models. QMEAN4 is a linear combination of four structural descriptors using statistical potentials: torsion angle potential to analyze local geometry, two distance dependent interaction potentials to assess long-range interactions and a solvation potential to assess the burial status of the residues (Benkert, Biasini et al. 2011).

DFire is used to assess non-bonded atomic interactions in the protein model. It is a knowledge based scoring function that uses *Distance-scale Finite Ideal-gas Reference* state rather than a statistically averaged state. It evaluates the theoretical pseudo energy of the entire model, which reflects the quality of the model and can be used for ranking alternative predictions of the same target. A lower energy indicates that a model is closer to the native conformation (Zhou and Zhou 2002).

Procheck is a suite of programs that assess the “stereochemical quality” of a protein structure. It is used to assess the geometry of residues of a given protein by comparing it with the stereochemical parameters derived from well-refined, high-resolution structures (Laskowski 1993). One of the main components of it is the Ramachandran plot, which uses the Phi (Φ) and Psi (Ψ) rotation angles of the amino acids of a protein to distribute them into allowed and disallowed regions based on the phi-psi rotation angles as shown in Figure 1.7 (Ramachandran, Ramakrishnan et al. 1963). The Phi angle represents the angle between the nitrogen atom of the amide group and alpha-carbon atom (N-C α) and the Psi angle represents the angle between the alpha-carbon atom and carbon atom of the carbonyl group (C-C α) (Figure 1.7a). These angles are restricted to certain values as some angles result in steric clashes between main chain and side chain atoms of the amino acids in a protein. Based on the permitted Phi and Psi angles the Ramachandran assigns certain regions on the plot as stereochemically most favorable (Red), allowed (dark yellow) and generously allowed regions (light yellow) as seen in Figure 1.7b. A good protein structure will have all its residues in the stereochemically favourable and allowed regions of the Ramachandran plot.

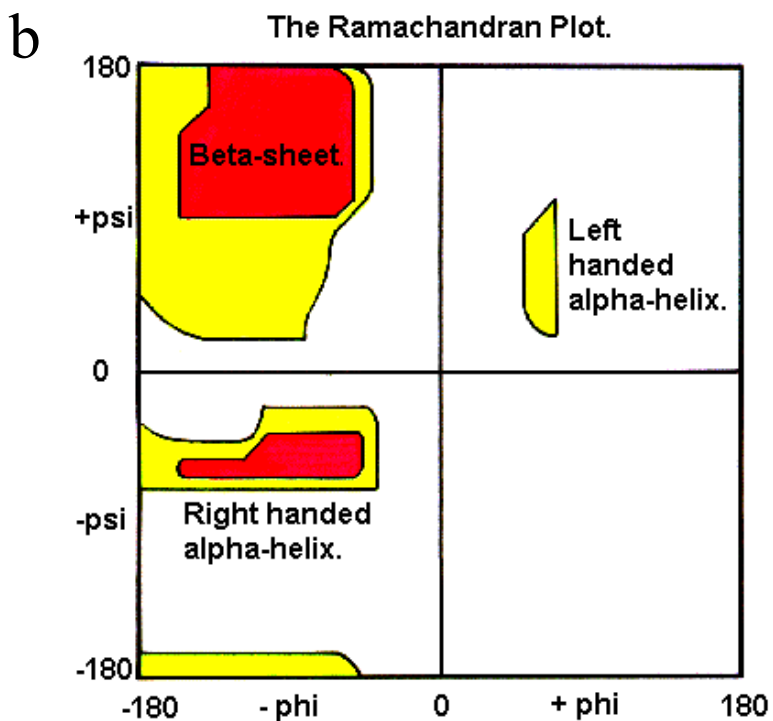
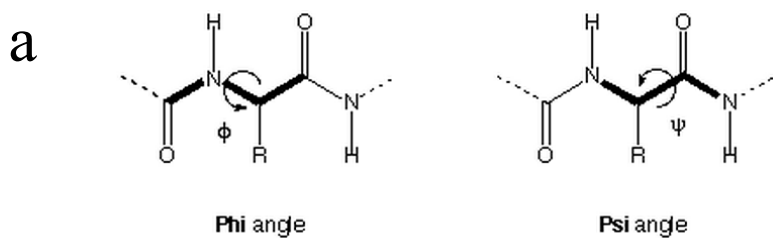


Figure 1.7: Representation of the Phi and Psi angles and Ramachandran plot

- a) Phi and Psi angles of an amino acid. Phi is the angle between the carbon- α atom and the nitrogen atom and Psi the angle between the carbon- α and carbon atom of the carbonyl group.
- b) Ramachandran plot showing the stereochemically allowed and disallowed regions. The colored regions represent the allowed secondary structure conformations as indicated in the figure and the white regions of the plot represent the disallowed regions.

1.11. Energy minimization and molecular dynamics (MD) simulations

The guiding principle behind homology modeling is to predict a protein model that is nearest to its native structure. Given an amino acid sequence, the number of putative three-dimensional structures that it can form is infinitesimal. However, from a thermodynamics perspective, the native state of a protein corresponds to its lowest free energy state in solution conditions as seen inside cells. The energy landscape of protein structures may be considered uneven with several local minima representing several non-native states of the protein and a global energy minimum representing the lowest energy state. Hence it is essential to ensure that the homology model achieves this lowest energy state. This is done by the process of energy minimization and MD simulation of the homology model.

Energy minimization is used to relax the structure by repairing distorted geometries and releasing internal constraints of amino acid side chains and does not affect the backbone structure of the protein. It is done by placing the structure in a solvent environment under the influence of a force field. Energy minimization will help the structure achieve the nearest local energy minimum, however it will not pass through high energy barriers which may be required to achieve the native structure.

MD simulation is an extension of energy minimization process. The main objective of this process is to correct any serious defects in the backbone structure of the model to achieve a thermodynamically stable energy state of the structure. It is important to know that the change in backbone structure due to MD

simulation is minimum, if a good quality homology model is obtained. The force fields that are used implement a set of empirical potential energy functions to mimic the behavior of atoms and molecules in a given solvent. Hence, MD simulation of a structure will allow us to model the behavior of the protein structure over the simulation time period. The structures over the time period are then sampled to obtain a representative structure at each time point, which are then clustered based on root mean square deviation in between structures.

1.12. Molecular docking

Molecular docking is the most critical step in the virtual screening procedure. It predicts the potential binding conformations and affinity of two macromolecules or usually a macromolecule (receptor) and a small molecule (ligand). Docking of small molecules is mainly based on the principle of molecular recognition whereby small molecules bind to the active site of a protein based on their interaction with certain key residues that anchor it in the binding pocket. The ultimate goal of molecular docking is to identify the best possible receptor-ligand conformation and the docking programs try to answer this question by employing two steps: searching the conformational space of the ligand to identify all possible binding configurations of the ligand in the binding pocket, and selection of the optimal binding configuration by scoring and ranking all suggested binding configurations based on their interactions and predicted binding energy. The number of configurations that can be achieved within a given search space is virtually infinite. To account for this, the docking programs should be accurate

and effective in sampling the various possible modes very quickly and identify the binding modes that are closest to the native conformation. There are over 30 different free and commercial docking programs available today which employ different sampling and scoring functions. The most commonly used programs are AutoDock (Goodsell, Morris et al. 1996), AutoDock vina (Trott and Olson 2010), GOLD (Verdonk, Cole et al. 2003), Glide (Friesner, Banks et al. 2004; Halgren, Murphy et al. 2004), DOCK (Ewing, Makino et al. 2001), FlexX (Claussen, Buning et al. 2001) and ICM (Totrov and Abagyan 1997). Based on the flexibility of receptor and ligand, docking can be classified into rigid docking and flexible docking.

1.12.1. Rigid docking

If the bond angles, lengths and torsional angles of the target protein and the ligand are kept constant (i.e. not flexible), then this type of docking is called rigid docking. This limits the degrees of freedom of rotation available to the components and is therefore computationally faster (C, Subramanian et al. 2009). This method is useful for docking small or very rigid molecules. Sometimes, the ligand could be allowed to be flexible in rigid docking, however, the protein structure remains rigid.

1.12.2. Flexible docking

In contrast to rigid docking, flexible docking allows certain residues of the protein to be flexible. The ligand may or not be flexible. Making some residues flexible allows for conformational change upon the binding of the ligand, thus simulating an environment similar to real protein-ligand binding.

1.13. Determination of potential ligand binding sites by blind docking

Blind docking is done to determine the potential binding sites of ligand on the protein structure (Hetenyi and van der Spoel 2002). This is especially useful to determine the binding sites of inhibitors and ligands that have been known to bind to the protein previously, but lack specific information about their binding sites on the protein surface. In this method the whole protein is encapsulated in a 3-D grid box and the surface of the protein is probed with a ligand to find its potential binding sites.

1.14. Small molecule databases

If the protein structure is considered as a lock, then the small molecule database is akin to a set of keys that may include keys that could potentially open that lock. In this project the following small molecule libraries were used to screen for inhibitors of Suv39h1 by virtual screening:

- National Cancer Institute diversity set II (NCIDS II),
- DrugBank database

- La Chimiotheque Nationale (CN) library
- Zinc clean fragments library
- Zinc Drug Database
- Princeton library of natural compounds
- HKMT library

DrugBank, Zinc clean fragments library, Princeton library of natural compounds, database obtained from the small molecule database ZINC (<http://zinc.docking.org/>) (Irwin, Sterling et al. 2012). The CN library was available with the Tuszynski lab as part of collaboration for other projects in the lab and was obtained from <http://chimiotheque-nationale.enscm.fr/>. There are around 21 million compounds in the ZINC repository that are available in several 3D formats suitable for screening with different docking programs in use.

The NCIDS II is a library of 2,044 compounds representing scaffolds derived from almost 140,000 compounds that are available for distribution from the National Cancer Institute. The major advantage of these compounds is that each molecule is distinct from another and they are the cluster representatives of their parent family.

The DrugBank database has 6,712 drug molecules including the 1448 FDA approved small molecule drugs, 131 FDA-approved biotech (protein/peptide) drugs, 85 nutraceuticals and 5,080 experimental drugs. It also incorporates detailed drug data with comprehensive target information (Wishart, Knox et al. 2006). The advantage of using this database is that the small molecules in this database are already clinically proven and FDA approved;

therefore any small molecule that is found to be active against our target of interest can be repurposed thus avoiding the tedious and very expensive task of clinical trials and FDA approval.

The CN library is a collection of synthetic compounds, natural compounds and natural extracts available in French academic institutions. The library has 87,512 compounds, which are divided into two main categories. The first category contains synthetic products and their associated chemical information and the second category contains the natural products and their extracts. This library was not obtained from the ZINC repository, hence the compounds had to be cleaned by assigning proper protonation states and conformational states before being used for screening.

Apart from being a repository of small molecule databases, ZINC also has many other useful features. An important feature available in ZINC is the ability to create customized subsets of compounds based on cut-off values of a specific set of parameters such as molecular weight, water-octanol partition co-efficient ($\log P$), functional groups etc. Consistent with this feature, a subset of ZINC database called as the clean fragments consisting of 72,861 compounds was used for screening.

ZINC Drug Database (ZDD) is a collection of all drugs that have been approved for human use somewhere in the world and are commercially available as pure compounds.

The Princeton library of natural compounds is a commercially available collection of compounds from Princeton BioMolecular Research Company. This library contains around 60,000 compounds.

The HKMT library is a database of small molecules prepared based on piperidine, lactone and lactam scaffolds. This library was prepared specifically to use for screening against histone lysine methyltransferases, based on the assumption that they could be synthesized in Dr. Hall's laboratory at the University of Alberta's Department of Chemistry.

1.15. Virtual screening

Historically drug discovery has been done by high throughput screening (HTS) of small molecules against a particular drug target. However, this method is extremely expensive and relatively time consuming even with the advances in automation of the methodology, and has been plagued by a very low success rate (Macarron, Banks et al. 2011). An alternative to this method is virtual screening (VS), where small molecules are screened against targets *in-silico* and the binding energy of the small molecules to the target structure is predicted. Often far more compounds exist or can be designed by combinatorial methods than can be tested affordably by HTS. One of the major advantages of VS is that the compounds to be tested need not be synthesized. Only the 3-D file of the molecule is docked into the binding site and its binding affinity is predicted. Based on the binding affinity, some top hits can be recommended that can then be tested by HTS. These top hits can then be further experimentally tested for their inhibitory capacity, which will

significantly reduce the effort, expenses and time involved in the drug discovery pipeline (Good, Krystek et al. 2000).

1.15.1. AutoDock

AutoDock is one of the most extensively used docking programs. Several search methods such as genetic algorithms, simulated annealing and local search are available in AutoDock (Morris, Goodsell et al. 1998). It implements an empirical free energy force field to predict free binding energies of the structure-ligand complex. This software was used to determine the binding pocket of SAM and chaetocin in this project.

1.15.2. AutoDock vina

AutoDock Vina is a new molecular docking and virtual screening program. The main advantage of this program is that it is at least twice as fast and shows higher accuracy in prediction of binding modes (Trott and Olson 2010). This software was used to for all virtual screening experiments performed in this project.

Chapter II – Materials and Methods

2. Chapter 2 – Materials and Methods

2.1. Cloning of *SUV39h1*

The DNA coding sequence of full-length human *SUV39h1* was amplified from the human fetal brain cDNA library (Clontech) by PCR using primers specific to the N-terminal and C-terminal end of human *SUV39h1* along with partial *attB* recombination site linkers (Table 2.2). The partial *attB* site linkers were extended by a second PCR reaction using *attB* adapter primers and then the amplified gene was recombined into the pDONR221 entry vector by BP recombination reaction (Gateway cloning system, Invitrogen). 5µl of the reaction product was used to transform into *E.coli* DH5α and plated onto LB (Luria Broth) agar plates containing 50µg/ml Kanamycin. Plasmids containing the h*SUV39h1*-Ch855/Ch856 gene fragment were screened by restriction digestion using HincII restriction enzyme and then analyzed by agarose gel electrophoresis by looking at band migration, which represent the size of the DNA fragments. The positive clones confirmed by restriction digestion were then sequenced to determine the exact sequence and location of the gene insert using BigDye Terminator v3.1 and an ABI PRISM 310 capillary sequencer (Applied Biosciences). The sequencing primers used to confirm the entry and destination clones are listed in Table 2.3. The melting temperature of the primers was calculated by using the MacVector program.

The human *SUV39h1* (h*SUV39h1*) gene was then recombined into Gateway destination vectors by LR reaction (Gateway cloning system,

Invitrogen). The destination vectors used were pDEST-N112-MBP, pDEST-N112-GST, pDEST-N110 (These plasmids were a generous gift from Dr. Michael Dyson, The Wellcome Trust Sanger Institute, UK) and pCEMM-NTAP-GS-(GW) (a generous gift from Dr. Tilmann Burckstummer, Research center for molecular medicine, Vienna, Austria) (Dyson, Shadbolt et al. 2004; Burckstummer, Bennett et al. 2006). The pDEST-N110 and pDEST-N112 vectors were used for bacterial expression and the pCEMM-NTAP-GS(GW) vector was used for mammalian expression of h*SUV39h1*.

The pDEST-N110 vector consists of the T7 RNA polymerase promoter, the lac operator, the Shine Dalgarno (SD) sequence and 10-Histidine (H10) residues in frame with the open reading frame (ORF) of the protein of interest (recombinant protein), which is flanked by *attB1* and *attB2* sites as shown in Figure 2.1. The *attB* sites are the recombination sites that can be used to shuttle the gene of the interest from one vector to others that are gateway compatible. The pDEST-N112 vectors have been fused with small protein tags such as Maltose Binding Protein (MBP), Glutathione S-Transferase (GST) in addition to the other elements present in pDEST-N110 vector. The fusion of these tags with recombinant proteins is shown to increase the solubility of protein expression in *E.coli* (Dyson, Shadbolt et al. 2004). The pCEMM-NTAP-GS-(GW) vector contains a Tandem Affinity Purification (TAP) tag which is a fusion protein consisting of protein G, Streptavidin Binding Protein (SBP) and Tobacco Etch Virus (TEV) protease site, fused at the N-terminal of the recombinant protein as

shown in Figure 2.1. The recombinant protein can be purified by utilizing the affinity of protein G with IgG and affinity of streptavidin with biotin.

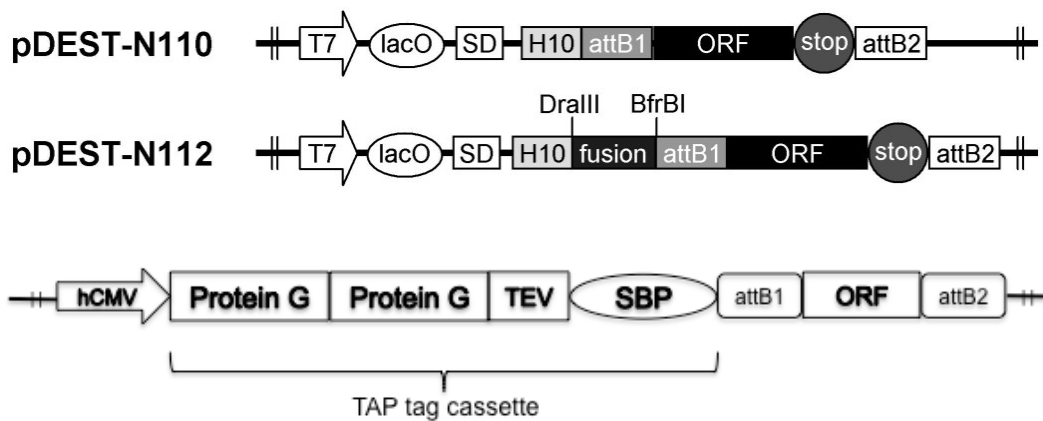


Figure 2.1: Schematic diagrams of the expression vectors used for cloning *hSUV39h1*.

The pDEST-N110 vector contains the T7 RNA polymerase promoter, lac operator, Shine Dalgarno (SD) sequence, 10-Histidine (H10) in frame with the ORF of the protein of interest (recombinant protein), which is flanked by *attB1* and *attB2* sites. The *attB* sites are the recombination sites that can be used to shuttle the gene of the interest from one vector to other that are gateway compatible. The pDEST-N112 vectors contain a fusion protein such as MBP and GST along with the elements present in the pDEST-N110 vector. The pCEMM-NTAP-GS(GW) vector consists of a TAP tag cassette containing protein G, TEV and SBP at the N-terminal followed by the *attB* recombination sites that can be used to insert the gene of interest.

The LR reaction product was transformed into *E. coli* DH5 α and plated onto LB agar plate containing 50 μ g/ml ampicillin. The destination clones were then confirmed by the same protocol as the entry vector. A list of the expression clones consisting full-length h*SUV39h1* is given in Table 2.4.

The h*SUV39h1* coding sequence was also cloned into the pET-SUMO vector (Invitrogen). h*SUV39h1* was amplified from pDONR221-h*SUV39h1*-Ch855/Ch856 using primers Ch1018 (N-terminal primer) and Ch1019 (C-terminal primer) (Table 2.2) and then cloned into the pET-SUMO vector by TA cloning. TA cloning is done by utilizing the extra deoxyadenosine (A) added to the 3' end of a PCR product by *Taq* polymerase, which is then ligated into a linearized empty pET-SUMO vector which has a single deoxythymidine (T) residue. 5 μ l of the ligated reaction product was used to transform One Shot Mach1-T1 *E.coli* bacteria (Invitrogen) and plated on LB agar plate containing 50 μ g/ml kanamycin. The colonies on the plate were screened for positive clones by the previously described protocol and confirmed by sequencing.

Vector	Fusion tag	Gateway	Antibiotic Resistance	Source of the vector
pDONR221	None	Yes	Kanamycin	Invitrogen
pDEST-N112-MBP	Maltose Binding protein and 6-His	Yes	Ampicillin	(Dyson, Shadbolt et al. 2004)
pDEST-N112-GST	Glutathione S-transferase and 6-His	Yes	Ampicillin	
pDEST-N110	10-His and 6-His	Yes	Ampicillin	
pET-SUMO	SUMO protein	No	Kanamycin	Invitrogen
pCEMM-NTAP-GS(GW)	Protein G and streptavidin	Yes	Ampicillin and Chloramphenicol	(Burckstummer, Bennett et al. 2006)

Table 2.1: List of vectors used for cloning *hSUV39h1* and their properties

The pDONR221 is an entry vector. The pDEST-N110, pDEST-N112-MBP, pDEST-N112-GST and pET-SUMO are bacterial expression vector and pCEMM-NTAP-GS(GW) is a mammalian expression vectors. All the vectors except the pET-SUMO vector are Gateway (Invitrogen) compatible vectors. The fusion tags of each vector and their antibiotic resistance are listed above.

Human Suv39h1, 412 amino acids

attB1 forward:5'-AA AAA GCA GGC TNN-template-specific sequences-3'

attB2 reverse:5'-A GAA AGC TGG GTN-template-specific sequences-3'

Ch855 hSuv39h1 N-terminal *attB1* linker primer 34mer

Tm: 63.1°C

5' - AA AAA GCA GGC TCG-ATG GCG GAA AAT TTA AAA GG -3'
 K A G S M A E N L K

Ch856 hSuv39h1 C-terminal *attB2* linker primer 33mer

Tm: 63.0°C

5' -A GAA AGC TGG GT-CTA TCC ACG CCA TTT CAC C- 3'
 * G R W K V

Ch374 *attB1* adapter sequence 29mer

Tm:72.7°C

5' - G GGG ACA AGT TTG TAC AAA AAA GCA GGC T -3'
 G T S L Y K K A G

Ch375 *attB2* adapter sequence 29mer

Tm:74.5°C

5' - GGG GAC CAC TTT GTA CAA GAA AGC TGG GT -3'
 G D H F V Q E S W

Ch1018 pETSUMO-hSuv39h1 N-terminal primer 24mer

Tm: 64.6°C

5' - TCG ATG GCG GAA AAT TTA AAA GGC -3'
 S M A E N L K G

Ch1019 pETSUMO-hSuv39h1 C-terminal primer 23mer

Tm: 66.7°C

5' - CTA GAA GAG GTA TTT GCG GCA GG -3'
 * F L Y K R C

Table 2.2: List of primers used for cloning full-length hSUV39h1.

The Tm of the primers was calculated using the MacVector program

Ch93	M13 reverse (-29) pGEM-T	17mer	Tm: 56.1°C
	5' - CAG GAA ACA GCT ATG AC -3'		
	Q E T A M		
Ch94	M13 forward (-40) pGEM-T	17mer	Tm: 56.1°C
	5' - GT TTT CCC AGT CAC GAC -3'		
	F P S H D		
Ch1016	hSuv39h1 5' sequencing primer (402-419)	18mer	Tm: 59.8°C
	5'- G CAG AAG GCC AAG CAG AG -3'		
	Q K A K Q		
Ch1017	hSuv39h1 3' sequencing primer (847-829)	20mer	Tm: 58.0°C
	5'- T GTT CTT GCG AAT CTT CTC C -3'		
	N K R I K E		
pET-SUMO forward sequencing primer		23mer	Tm: 57.9°C
	5' - A GAT TCT TGT ACG ACG GTA TTA G -3'		
	D S C T T V L		
T7 Reverse sequencing primer		20mer	Tm: 62.7°C
	5'- T AGT TAT TGC TCA GCG GTG G -3'		
	S Y C S A V		

Table 2.3: List of the sequencing primers used to sequence different vectors as shown in the table.

Designation	Gene insert	Length of the insert (amino acids)
pDEST-N112-MBP-h <i>SUV39h1</i> - Ch855/Ch856	Full length human <i>SUV39h1</i>	1-412
pDEST-N112-GST-h <i>SUV39h1</i> - Ch855/Ch856	Full length human <i>SUV39h1</i>	1-412
pDEST-N110-h <i>SUV39h1</i> - Ch855/Ch856	Full length human <i>SUV39h1</i>	1-412
pCEMM-NTAP-(GS)-GW- h <i>SUV39h1</i> -Ch855/856	Full length human <i>SUV39h1</i>	1-412
pET-SUMO-h <i>SUV39h1</i> - Ch1018/Ch1019	Full length human <i>SUV39h1</i>	1-412

Table 2.4: List of expression vectors prepared containing full-length h*SUV39h1*.

2.2. Mutagenesis of hSUV39h1

pDONR221-hSUV39h1-855/856 was used as a template for all mutagenesis reactions. Truncation mutants were generated by PCR with gene specific primers with partial *attB* recombination site linkers at specific sites (Table 2.5). The recombination linkers were extended by a second PCR reaction and then recombined into pDONR221 vector by BP reaction (Invitrogen). The truncation mutants were recombined into expression vectors by LR reaction (Invitrogen). A list of the hSUV39h1 truncation mutants that were prepared is shown in Table 2.6. The site directed mutagenesis of SUV39h1 was done by using QuikChange site-directed mutagenesis kit (Stratagene). The site directed mutagenesis primers were designed by using the PrimerX software with the recommended specifications as per the mutagenesis kit ($T_m \geq 78^\circ\text{C}$, GC% = $\geq 40\%$ and a minimum of 10 bp of matched sequence on each side) (Table 2.7). Two site directed mutant constructs were prepared by using the pCEMM-NTAP-GS(GW)-hSUV39h1-855/856 construct as a template; a substitution mutation of the histidine 320 residue to arginine (H320R) in order to express hyperactive human Suv39h1 and a substitution mutation of the histidine 324 residue to lysine (H324L) to express methyltransferase activity dead hSUV39h1 (Table 2.8) (Rea, Eisenhaber et al. 2000). The mutation was introduced into the sequence by PCR with PfuI Turbo DNA polymerase. The PCR reaction product was digested with DpnI for 2 hours at 37°C to remove the parental plasmid. The reaction product is transformed into *E.coli* and screened for positive clones by restriction digestion as explained in section 2.1. The mutant sequences were confirmed by sequencing using BigDye

terminators v3.1 and an ABI PRISM 310 capillary sequencer (Applied Biosciences).

Ch855 hSuv39h1 N-terminal *attB1* linker primer 34mer

T_m: 63.1°C

5'- AA AAA GCA GGC TCG-ATG GCG GAA AAT TTA AAA GG -3'
 K A G S M A E N L K

Ch856 hSuv39h1 C-terminal *attB2* linker primer 33mer

T_m = 63.0°C

5'-A GAA AGC TGG GT-TA GAA GAG GTA TTT GCG GCA G 3'
 * F L Y K R C

Ch977 hSUV39h1 3' *attB2* linker primer with STOP codon
31mer, upto **aa66** T_m: 62.0°C

5'-A GAA AGC TGG GT-CTA TCC ACG CCA TTT CAC C- 3'
 * G R W K V

Ch978 hSUV39h1 3' *attB2* linker primer with STOP codon
33mer, upto **aa92** T_m: 61.9°C

5'-A GAA AGC TGG GT-TA GTC CTT GTG GAA CTG CTT G-3'
 * D K H F Q K

Ch1036 hSUV39h1 – SET domain N-terminal primer

31mer starts at **aa100** T_m: 64.4°C

5'- AA AAA GCA GGC TC-G CAC CAC CGG TCA AAG AC-3'
 K A G S H H R S K

Ch1037 hSUV39h1 – SET domain N-Terminal primer

33mer starts at **aa82** T_m: 63.3°C

5'- AA AAA GCA GGC TCG-TGT GTG CGT ATC CTC AAG C-3'
 K A G S C V R I L K

Ch1050 hSUV39h1 SET domain N – terminal primer for pET
SUMO, 22mer starts at **aa82** T_m: 69.4°C

5' – TCG TGT GTG CGT ATC CTC AAG C - 3'
 S C V R I L K

Ch1051 hSUV39h1 SET domain N – terminal primer for pET
SUMO, 23mer starts at **aa100** T_m: 68.5°C

5' – TCG CAC CAC CGG TCA AAG AC -3'
 S H H R S K

Ch1019 pETSUMO-hSuv39h1 C-terminal primer 23mer

Tm: 66.7°C

5' - CTA GAA GAG GTA TTT GCG GCA GG -3'
* F L Y K R C

Table 2.5: List of primers used to generate various truncation mutants shown in Table 2.6

Designation	Truncation (Amino acid)	Vectors used	Description
<i>hSUV39h1</i> - Ch855/977	1-66	pDEST-N112-MBP pDEST-N112-GST pDEST-N110	Bacterial expression vectors
<i>hSUV39h1</i> - Ch855/978	1-92	pDEST-N112-MBP pDEST-N112-GST pDEST-N110	Bacterial expression vectors
<i>hSUV39h1</i> - Ch1036/856	100-412	pCEMM-NTAP- GS(GW)	Mammalian expression vector
<i>hSUV39h1</i> - Ch1037/856	82-412	pCEMM-NTAP- GS(GW)	Mammalian expression vector
<i>hSUV39h1</i> - Ch1050/1019	82-412	pET-SUMO	Bacterial expression vector
<i>hSUV39h1</i> - Ch1051/1019	100-412	pET-SUMO	Bacterial expression vector

Table 2.6: List of *hSUV39h1* truncation mutants prepared

hSUV39h1-Ch855/977 and *hSUV39h1*-Ch855/978 are the N-terminal truncations and were made in order to synthesize protein fragments that could be used for antibody production. *hSUV39h1*-Ch1036/856 and *hSUV39h1*-Ch1037/856 contain the SET domain only and were cloned into mammalian expression vector. *hSUV39h1*-Ch1050/1019 and *hSUV39h1*-Ch1051/1019 also contain the SET domain only and were cloned into bacterial expression vector.

Ch1144: hSUV39h1 **H320R** 5' primer 36mer Tm: 80.2°C
5'-C TAT GGC AAC ATC TCC CGC TTT GTC AAC CAC AGT TG-3'
Y G N I S **R** F V N H S

Ch1145: hSUV39h1 **H320R** 3' primer 36mer Tm: 80.2°C
5'-CA ACT GTG GTT GAC AAA GCG GGA GAT GTT GCC ATA G-3'
S H N V F **R** S I N G Y

Ch1146: hSUV39h1 **H324L** 5' primer 34mer Tm: 80.4°C
5'-C TCC CAC TTT GTC AAC CTC AGT TGT GAC CCC AAC-3'
S H F V N **L** S C D P N

Ch1147: hSUV39h1 **H324L** 3' primer 34mer Tm: 80.4°C
5'-GTT GGG GTC ACA ACT GAG GTT GAC AAA GTG GGA G-3'
N P D C S **L** N V F H S

Table 2.7: List of primers used to create site directed mutants of hSUV39h1

Designation	Mutant aminoacid	Description
pCEMM-NTAP-(GS)-GW-h <i>SUV39h1</i> -H320R-Ch1144/1145	H320R	Hyperactive methyltransferase activity
pCEMM-NTAP-(GS)-GW-h <i>SUV39h1</i> -H324L-Ch1145/1146	H324L	Methyltransferase activity dead

Table 2.8: Site directed mutant constructs of full-length h*SUV39h1*

2.3. Expression of human Suv39h1

2.3.1. Bacterial expression system

Recombinant full-length and truncated SUV39h1 protein fused with MBP, GST and 6-His (6-histidine) tags were induced in *E.coli* strains JM109, Rosetta and BL21 DE3 codon+. An overview of the different expression vectors, bacterial strains and expression conditions used to express human Suv39h1 is given in Table 2.9. The expression of the proteins was induced with 0.1 mM, 0.5 mM, 1.0 mM and 2.0 mM isopropyl β -D-thiogalactoside (IPTG) (Invitrogen) for 3 hours at 37°C and overnight (O/N) at 16 °C and 25 °C in 2 ml of LB media (mini-inductions) with appropriate antibiotics (refer to Table 2.1) to determine the solubility and expression of the fusion proteins. The overexpressed protein was extracted by lysing the bacterial cells using BugBuster reagent (Novagen) and separating the soluble and insoluble fractions by centrifugation at 16000 \times g at 4 °C for 20 minutes. The soluble and insoluble fractions were analyzed by sodium-dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE) and staining the gel with commassie blue stain. The expression and solubility of the fusion protein was determined by analyzing the gel and looking for the presence of the correct size protein band on the polyacrylamide gel in the lanes containing either the soluble or insoluble fraction.

Large volume inductions (400 ml of LB media each) were also done to produce large quantities of protein. Single colony of antibiotic resistant bacteria was used to inoculate 5ml of LB media with appropriate antibiotic (refer to Table 2.1) and the culture was grown O/N. 3 ml of the O/N culture was subcultured into

400 ml of LB media with appropriate antibiotic and the culture was grown until it reached $OD_{600} = 0.7$. The overexpressed protein was extracted from the cells using lysis buffer containing 1% NP40 followed by sonication. During sonication the sample was placed in a small plastic beaker on ice and sonicated in short pulses of 30 seconds for 5 times. The bacterial extract was centrifuged at 4 °C at $3830\times g$ for 10 minutes and the soluble and insoluble fractions were analyzed by SDS-PAGE.

Lactose induction was done to encourage protein expression in soluble fraction. Lactose replaces IPTG as an inducer of the lac operon and is shown to express high levels of soluble protein expression compared to IPTG induction (Tian, Tang et al. 2011). The protein was expressed in Rosetta, JM109 and BL21-DE3 codon+ bacterial strains for 12 hours at 20 °C and extracted and analyzed as described previously for mini-inductions.

Expression Vectors	pDEST-N112-MBP		pDEST-N112-GST		pDEST-N110		pET-SUMO	
	Gene fragment	Full length	SET domain	Full length	SET domain	Full length	SET domain	Full length
Parameter	Full length	SET domain	Full length	SET domain	Full length	SET domain	Full length	SET domain
IPTG(mM) (0.5,1.0,2.0)	✓	✓	✓	✓	✓	✓	✓	✓
Temp (°C) (16, 25, 30)	✓	✓	✓	✓	✓	✓	✓	✓
Time (4 and 16 hours)	✓	✓	✓	✓	✓	✓	✓	✓
Lactose induction (Overnight at 20°C)	✓	✓	✓	✓	✓	✓	✓	✓
JM109	✓	✓	✓	✓	✓	✓	✓	✓
Rosetta	✓	✓	✓	✓	✓	✓	✓	✓
BL21-DE3 Codon+	✓	✓	✓	✓	✓	✓	✓	✓

Table 2.9: List of the bacterial expression vectors and parameters used for testing the expression of human Suv39h1 fusion proteins.

Expression vectors containing only the SET domain and full-length *hSUV39h1* were transformed into JM109, Rosetta, BL21-DE3 codon+ strains of *E.coli*. The transformation mix was plated onto LB agar medium with appropriate antibiotic (refer to Table 2.1) and single colonies were picked and grown in LB medium overnight with appropriate antibiotic. 50 µl of the overnight bacterial culture was inoculated to fresh 2ml LB media and was propagated until the culture reached an OD₆₀₀ of 0.7. Recombinant protein was expressed by adding 0.1 mM, 0.5 mM, 1.0 mM and 2.0 mM of IPTG (inducer of lac operon) and growing the cells for 4 hours at 30 °C and overnight at 16 °C and 25 °C. Lactose induction system was also used for all these constructs at 20 °C overnight.

2.3.2. Mammalian expression system

The *hSUV39h1* gene was cloned into the pCEMM-NTAP-GS-(GW) vector, which contains protein G and Streptavidin Binding Protein (SBP) tags that can be used to purify the recombinant protein. The *hSUV39h1* gene was cloned in frame with the TAP tag cassette in the vector and was used for protein expression. HEK293T cells were seeded at a density of 1×10^5 cells/ml (2 ml) in 35 mm dishes and were transiently transfected with pCEMM-NTAP-GS-(GW)-*hSUV39h1*-855/856 plasmid using polyethylenimine (PEI) and the protein was transiently expressed for 24, 36 and 48 hours. Transfection efficiency was confirmed by observing the GFP expression levels under the microscope. Cells were trypsinized and harvested after 24, 36 and 48 hours later; they were pelleted at 1000RPM for 8minutes, washed with PBS, repelleted and lysed in 1% NP40 buffer with full protease and phosphatase inhibitors. The lysate was incubated on ice for 10 minutes and then transferred to a cold 1.6ml microfuge tube and the soluble and insoluble fractions were separated by centrifugation at 12000g for 10 minutes. The 3X SDS loading buffer was added to the soluble fraction to a final concentration of 1X and the protein expression was confirmed by SDS-PAGE and western blot.

2.4. Protein purification

The expressed recombinant protein was purified by affinity chromatography. pDEST-N112-MBP, pDEST-N112-GST and pDEST-N110 vectors contain a N-terminal fusion of 6-Histidine protein and their respective fusion tag proteins (Table 2.4). The 6-Histidine tag exhibits high affinity and selectivity to nickel charged affinity resin. In this project the ProBond Nickel-Chelating Resin

(Invitrogen) was used to purify the 6-His tagged human Suv39h1 proteins from the bacterial cell lysate containing overexpressed protein following manufacturers protocol. This resin allows for harsh buffer conditions to be used in order to elute insoluble protein under denaturing conditions. The protein was eluted in different fractions and its purity was checked by SDS-PAGE.

2.5. Sodium Dodecyl Sulphate – Polyacrylamide Gel Electrophoresis (SDS-PAGE)

SDS-PAGE was performed to separate proteins based on their size. A 12.5% polyacrylamide resolving gel was used usually unless otherwise mentioned (BioRad mini gel system). The protein samples were boiled for 5 minutes in a final concentration of 1X SDS loading buffer and were electrophoresed for 1 to 1.5 hours at 150 volts. SDS-6H2 ladder (Sigma) was used for gels stained with Coomassie and Prestained PageRuler Plus ladder (Thermo scientific) was used for western blots. The gels were stained with Coomassie blue for 10-15 minutes, washed with nanopure water and destained overnight with destaining solution. The separated protein bands were visualized and imaged.

2.6. Western blot

The protein samples were electrophoresed on a polyacrylamide gel by SDS-PAGE as described previously and then transferred onto Immobulin-P membrane (Millipore) for 1.5 hours at 400mA in 1X western transfer buffer. The blot was then carefully separated from the gel and was blocked overnight with Odyssey blocking buffer (Li-Cor Biosciences). In the case of mammalian expression of human Suv39h1 the membrane was probed with Alexa 680 goat anti-rabbit IgG antibody (Invitrogen). The antibody was diluted in 10% blocking buffer in PBS (1:10000, Invitrogen) and incubations were done for 1 hour followed by three washes of 5 minutes each with PBS with 0.1% Tween20 and one last wash for 5 minutes with PBS alone. The blots were scanned with Odyssey infrared imager system (Li-Cor Biosciences).

2.7. Reagent recipes

1% NP40 lysis buffer

1%NP40
150mM NaCl
50mM Tris pH8
Protease inhibitors

12% SDS resolving gel stock (Total volume = 200ml)

83.4ml of 30% acrylamide stock (37.5:1 acrylamide:bis)
1ml of 20% SDS
0.2ml of 0.5M EDTA
50ml of 1.5M Tris pH8.8
ddH₂O up to 200ml

15% SDS resolving gel stock (Total volume = 200ml)

100ml of 30% acrylamide stock (37.5:1 acrylamide:bis)
1ml of 20% SDS
0.2ml of 0.5M EDTA
50ml of 1.5M Tris pH8.8
ddH₂O up to 200ml

Stacking gel (Total volume = 200ml)

26.7ml of 30% acrylamide stock (37.5:1 acrylamide:bis)
1ml of 20% SDS
0.2ml of 0.5M EDTA
50ml of 0.5M Tris pH6.8
ddH₂O upto 200ml

SDS-PAGE running buffer

30.3g of Tris
1441.7g of glycine
10g of SDS or 50ml of 20% SDS (0.1% SDS final concentration)
Upto 10L with ddH₂O

Coomassie blue stain

1.25g Coomassie blue (0.25% Coomassie blue)
250ml of methanol (10% Acetic acid)
50ml glacial acetic acid
Upto 500ml with ddH₂O

SDS-PAGE destaining solution

10% Acetic acid
10% Methanol

1X western transfer buffer

12.12g of Tris

57.68g of glycine

4g SDS or 20ml of 20% SDS (0.1% SDS final concentration)

800ml methanol

Upto 4L with ddH₂O

2.8. Homology modeling of Suv39h1

2.8.1. Identification of template for Suv39h1 homology modeling

The protein sequence of human Suv39h1 (1-412 amino acids) was obtained from the NCBI protein database with accession id: NP_003164. A protein similarity search done with PSI-BLAST revealed that there were no experimentally determined crystal structures of Suv39h1 or a complete protein crystal structure of any of its homologues available. The PSI-BLAST search revealed, however, that the nearest homologous crystal structure was that of the methyltransferase domain of human SUV39h2 from amino acid residues 111-411 with the PDB id 2R3A (Wu, Min et al. 2010). This alignment was submitted to the alignment mode of the Swiss model homology-modeling server to model the structure (<http://swissmodel.expasy.org/workspace>) (Arnold, Bordoli et al. 2006; Bordoli, Kiefer et al. 2009).

2.8.2. Homology modeling using Swiss model server

Both the alignment mode and the automated mode of the Swiss model server were used to obtain the homology model. As described earlier the sequence alignment was input into the server using the alignment mode and in the automated mode only the primary amino acid sequence of Suv39h1 was input into the server.

2.8.3. Quality estimation of the predicted models

The quality of the predicted models was assessed by using the ANOLEA, QMEAN, DFire and PROCHECK analysis modules in the Swiss model server (Hooft, Vriend et al. 1996; Laskowski, Chistyakov et al. 2005).

2.9. Energy minimizations and molecular dynamics simulations

The predicted models obtained from the Swiss model server were energy minimized and subjected to molecular dynamics (MD) simulations using AMBER11 software. AMBER (Assisted Model Building with Energy Refinement) is a set of software that contains molecular mechanical force fields for the simulation of biomolecules and a package of molecular simulation programs. In both energy minimization and MD, AMBER force field ff03.r1 was used (Cornell, Cieplak et al. 1995). The potential cutoff was set to 15 Å; temperature was 300 K; and explicit water model TIP3P was used. The charge of the entire system was neutralized by adding counter ions to the environment. The size of the periodic box for simulation was set to 71.61 Å.

Energy minimization was done in two stages. The first stage was the restrained energy minimization, which was done using the steepest descent method for the first ten steps and then switched to the conjugate gradient method for the next 1000 steps. The value of potential restraint weights used for this stage was 100 kcal/mol×Å. The second stage was unrestrained energy minimization, which was done for 2000 steps in the conjugate gradient method as well.

The simulation time step for MD simulations of the model was set to 2 femtosecond (fs). The Langevin dynamics method with a collision frequency of 1

ps⁻¹ was used to regulate temperature of targets and the surrounding environment. Bonds involving hydrogen were constrained using the SHAKE algorithm, constant pressure was used to equilibrate the density of solution and the isotropic position scaling method was used to regulate pressure. The weight of the potential restraints used in the third stage was 10 kcal/mol×Å². MD simulation was performed for 23ns.

2.10. Molecular docking to determine potential ligand binding sites

Docking was done using AutoDock software to determine the potential binding site of SAM, chaetocin and the histone H3 N-terminal peptide on the surface of the predicted model as described below.

2.10.1. Determination of binding site of S-Adenosyl methionine

The binding site of SAM on the surface of Suv39h1 was determined by blind docking using AutoDock software. The refined Suv39h1 model was encapsulated in a grid box and the entire surface of the model was searched to find the putative binding sites of SAM. The genetic algorithm – local hybrid search (part of AutoDock) along with its default parameters was used to perform this step. A grid box of dimensions 88Å × 82Å × 104Å, with the grid points separated by 0.61Å was set up around the Suv39h1 model.

2.10.2. Determination of binding site of chaetocin

Blind docking was done to determine the potential binding pocket of chaetocin on the surface of the predicted model as per the protocol described above for SAM.

2.10.3. Determination of binding site of H3K9 peptide

The structure of the H3K9 peptide was obtained from the crystal structure of the catalytic domain of human euchromatic histone methyltransferase 1 (hEHMT1) from PDB (PDB id: 3HNA), which has been crystallized with the peptide bound to it (Wu, Min et al. 2010). The peptide was treated as a flexible entity by specifying each amino acid to be completely rotatable along its root axis. Blind docking using AutoDock with the peptide and the model was performed as described earlier. The protein-peptide docking server GRAMMX was also used to determine the potential binding site of H3K9 peptide.

A sequence and structure comparison was done between the crystal structure of EHMT1 consisting the mono-methylated H3K9 peptide and the homology model of Suv39h1 to identify the binding pocket of the peptide on Suv39h1.

2.11. Small molecule inhibitor libraries

The small molecule libraries used for virtual screening are listed in Table 2.10.

Small molecule library	Number of small molecules
NCIDS II	2044
DrugBank	2225
La Chimiotheque Nationale library	87512
Zinc clean fragments	72861
Princeton library of natural compounds	61056
HKMT library	49572

Table 2.10: List of the small molecule inhibitor libraries screened and the number of small molecules in each of them.

2.12. Virtual screening

AutoDock Vina 1.0.2 was used for virtual screening of all the small molecule libraries (Trott and Olson 2010). The Suv39h1 PDB file was processed and converted to PDBQT format by using AutoDock tools. This process involves adding hydrogen atoms and adding partial atomic charges to the protein structure by the Gasteiger-Marsili method (Gasteiger and Marsili 1978). The binding sites of SAM, chaetocin and the histone H3 N-terminal peptide determined as explained earlier were used for this process. The sizes of the grid boxes containing the protein and the grid centers for the three binding pockets are as described in Table 2.11. The exhaustiveness parameter, which is a term for thoroughness was set to 8, and 10 binding modes were determined for each ligand per binding pocket.

All the small molecule libraries were screened on the pharmamatrix cluster consisting of 128 nodes with each node consisting of 8GB RAM.

Binding pocket	Grid size	Grid center
SAM binding pocket	26Å × 30Å × 26Å	54Å × 32Å × 68Å
Chaetocin binding pocket	18Å × 18Å × 18Å	72Å × 42Å × 63Å
H3K9 peptide binding pocket	26Å × 30Å × 26Å	57Å × 26.5Å × 88.5Å

Table 2.11: Dimensions of the grid boxes and their centers at different binding pockets used for virtual screening.

2.13. Analysis and clustering of the top hits obtained from virtual screening

The small molecules were ranked by AutoDock Vina based on the predicted binding energy. The top hits from each database were visually inspected to analyze their binding conformation. The top 50 hits from each database were clustered based on their structural similarities using chemmine tools in order to get structurally unique compounds (<http://chemmine.ucr.edu/>) (Backman, Cao et al. 2011). Finally, the clustered top hits from all the databases were combined and re-clustered to obtain a set of compounds that are distinct from each other.

2.14. *In vitro* histone methyltransferase assay

The next step after virtual screening was to validate the small molecules predicted from virtual screening by an *in vitro* assay. Fluorescent and radioactive histone methyltransferase assay methods were tried to study the inhibitory capacity of the small molecule inhibitors.

2.14.1. Fluorescent *in vitro* histone methyltransferase assay

The *EpiQuik*TM Histone Methyltransferase Activity/Inhibition Kit (H3-K9) (catalog# P-3003-1, Epigentek) was tried first to validate the predicted small molecules. It is an ELISA (Enzyme-Linked Immuno Sorbant Assay) based assay in which the histone substrate is stably captured on the strip wells of a 96 well plate. hSuv39h1 enzyme transfers a methyl group to histone H3 substrate from SAM to H3K9 residue and the methylated H3K9 is recognized by a high affinity

antibody. The ratio or the amount of methylated H3K9, which is directly proportional to enzyme activity, can be quantified through a Horse Radish Peroxidase (HRP) secondary antibody–colour development system. The colour developed was measured by using a microplate reader at 450nm. One of the reasons for considering this type of assay was its ease of use in a kit form and safety advantage over the radioactive assay. A schematic of the steps involved in the fluorescent methyltransferase assay is shown in Figure 2.2.

The fluorescent methyltransferase assay was performed using several controls and samples to optimize it for screening of the top hits obtained from virtual screening. All the samples were normalized before being used in the assay. The different controls and samples used to test this assay are listed in Table 2.12.

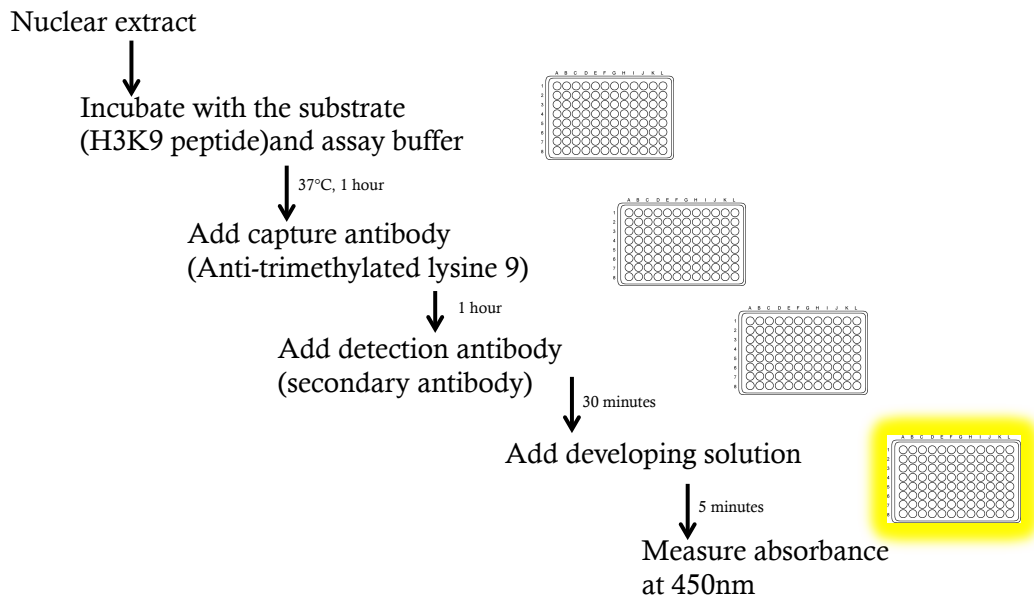


Figure 2.2: Schematic representation of the steps involved in fluorescence histone methyltransferase assay

Samples
HEK293T lysate transfected with pCEMM-NTAP-GS(GW)-h <i>SUV39h1</i>
Positive control (supplied with the kit)
HeLa cell lysate
HEK293T cell lysate
W8 - wild type Mouse Embryonic Fibroblast (MEF) lysate
D5 - Suv39h1/h2 double null MEF lysate
Suv39h1 transfected HEK293T lysate treated with 1.6 μ M chaetocin

Table 2.12: List of samples used to test the fluorescent histone methyltransferase assay

2.14.2. Radioactive *in vitro* histone methyltransferase assay

The identified top hits were tested for inhibitory activity against purified human Suv39h1 (Sigma). Biotinylated N-terminal fragments of histone H3 peptides (ARTKQTARKSTGGKAPRKQLA-GG-K(BIOTIN)-NH₂) either unmethylated, monomethylated at K9 and acetylated at K9 (Anaspec) were used as substrates for the reaction. The H3 peptide acetylated at K9 was used as a negative control in the reaction. The reaction setup with all the controls used to validate this assay is as shown in Table 2.13.

The assay was done by incubating 5 µg of the peptides with 1 µCi of tritiated S-adenosyl methionine (³H-SAM) and 0.1 µg of purified human Suv39h1 enzyme (Sigma) in 30µl of methylation activity buffer (50 mM Tris pH 8.5, 20 mM KCl, 5 mM MgCl₂, 5 mM β-Mercaptoethanol and 250 mM Sucrose) for 60 minutes at 37 °C. After 60 minutes, the reaction tubes were quickly spun to collect all the droplets and the reaction mix was spotted on to streptavidin membrane (SAM² biotin capture membrane, Promega). This membrane binds biotinylated molecules based on their affinity for streptavidin. The membranes were dried for 10 minutes at room temperature and then washed sequentially as per the following protocol in a beaker with 300 ml of washing solution. Wash 1 time with 2M NaCl for 30 seconds, 3 times with 2M NaCl for 2 minutes each, 4 times with 2M NaCl with 1% phosphoric acid for 2 minutes each and finally 2 times with deionized water for 30 seconds each. The washing procedure was done on an orbital shaker to prevent the membranes from settling to the bottom of the beaker. The membranes were then dried for 15 minutes at room temperature and

then immersed in plastic vials containing 5 ml of scintillation fluid (Ecolite(+)TM Liquid Scintillation Fluid, MP Biomedicals). The vials containing the membranes were stored for an hour for the counts to stabilize and then counted using liquid scintillation counter for 5 minutes per sample.

During the completion stages of the project, a slight variation in the assay protocol was employed. The membranes were not allowed to dry before being washed.

Sample description	H3-SAM (0.55μCi)	Suv39h1 (0.2μg)	H3K9 (5μg)	H3K9me (5μg)	H3K9Ac (5μg)	Chaetocin (1.6μM)	DMSO
H3K9	✓	✓	✓	×	×	×	×
H3K9me	✓	✓	×	✓	×	×	×
H3K9Ac	✓	✓	×	×	✓	×	×
No Suv39h1	✓	×	✓	×	×	×	×
No H3-SAM	×	✓	✓	×	×	×	×
No peptide	✓	✓	×	×	×	×	×
only H3-SAM	✓	×	×	×	×	×	×
Chaetocin	✓	✓	✓	×	×	✓	×
DMSO	✓	✓	✓	×	×	×	✓
Blank	×	×	×	×	×	×	×

Table 2.13: Experimental plan with all the necessary controls to validate the methyltransferase assay.

Chapter III – Results

3. Chapter 3 – Results

3.1. Expression of recombinant Suv39h1

Effort was put into the production of active Suv39h1 enzyme, so that it could be used for the *in vitro* validation of the small molecule inhibitors. Full length *SUV39h1* and only SET domain of *SUV39h1* were cloned into different vectors. These were then transformed into bacteria or HeLa cells to express recombinant *SUV39h1*.

3.1.1. Bacterial protein expression system

Several expression plasmids were made which were then used to express human *SUV39h1* in bacterial expression system as listed in Table 3.1. Despite trying different expression parameters such as varying the temperature of induction, IPTG concentration, time of protein induction, different strains of *E.coli* and different *SUV39h1* constructs, functionally active Suv39h1 was not expressed in the soluble fraction.

Expression Vectors	pDEST-N112-MBP		pDEST-N112-GST		pDEST-N110		pET-SUMO	
	Gene fragment	Full length	SET domain	Full length	SET domain	Full length	SET domain	Full length
Parameter	Full length	SET domain	Full length	SET domain	Full length	SET domain	Full length	SET domain
IPTG(mM) (0.5,1.0,2.0)	✓	✓	✓	✓	✓	✓	✓	✓
Temp (°C) (16, 25, 30)	✓	✓	✓	✓	✓	✓	✓	✓
Time (4 and 16 hrs)	✓	✓	✓	✓	✓	✓	✓	✓
Lactose induction (Overnight at 20°C)	✓	✓	✓	✓	✓	✓	✓	✓
JM109	✓	✓	✓	✓	✓	✓	✓	✓
Rosetta	✓	✓	✓	✓	✓	✓	✓	✓
BL21-DE3 Codon+	✓	✓	✓	✓	✓	✓	✓	✓
Expression in soluble fraction	×	×	×	×	×	×	×	×

Table 3.1: Bacterial expression vectors, *E.coli* strains, and expression parameters used to express SUV39h1

Constructs containing full length and SET domain only *SUV39h1* were used to express the protein in JM109, Rosetta and BL21 DE3 strains of *E.coli*. All the expression strategies employed did not result in soluble expression of Suv39h1.

3.1.2. Mammalian protein expression system

The attempts to express functionally active SUV39h1 using bacterial expression system were unsuccessful. Therefore a mammalian expression system was employed to express functionally active SUV39h1. The coding sequence of *SUV39h1* was cloned into the pCEMM-NTAP-GS-(GW) vector, which contains protein G and Streptavidin Binding Protein (SBP) tags that can be used to purify the recombinant protein. The pCEMM-NTAP-GS-(GW) vector also contains GFP. The transfection efficiency was confirmed by looking at the GFP expression levels under the Olympus microscope. The protein expression was confirmed by western blot as shown in Figure 3.1.

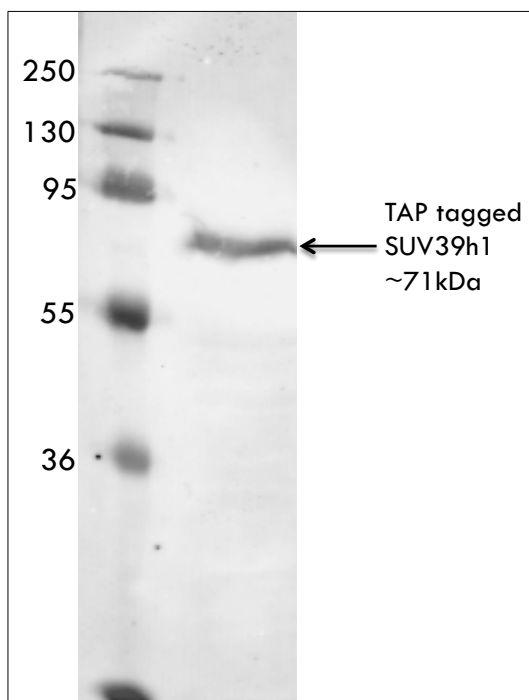


Figure 3.1: Western blot showing expression of human Suv39h1 in the mammalian expression system

Western blot of HEK293T cell lysate transfected with pCEMM-NTAP-GS(GW)-h*SUV39h1*-855/856 showing expression of full length human Suv39h1. The expected size of the protein is 70.97kDA. The western blot shows a band at the expected size, which can be determined by comparison with the molecular weight marker.

3.2. Homology modeling of Suv39h1

3.2.1. Identification of template structure

The methyltransferase domain of human Suv39h2 (PDB id: 2R3A) was identified as the template structure by a similarity search using the PSI-BLAST program. The 2R3A crystal structure does not contain the first 110 amino acids of Suv39h2. The PSI-BLAST's sequence alignment shows that there is a 62% sequence identity between the amino acid sequence of Suv39h1 and Suv39h2 in the region corresponding to the crystal structure of Suv39h2, i.e. from amino acid 111 to 411 (Figure 3.2). Suv39h2 was the only structure available with such high sequence similarity. The next available structure with high sequence similarity was the G9a-like histone methyltransferase (PDB id: 3FPD), with 42% sequence similarity. Further analysis showed that the SET domains of Suv39h1 and Suv39h2 share 77% sequence identity, which gives us enough confidence to select Suv39h2 as a template to model Suv39h1 structure (Figure 3.3).

Chain A, Methyltransferase Domain Of Human Suppressor Of Variegation 3-9 Homolog 2
 Sequence ID: [pdb|2R3A|A](#) Length: 300 Number of Matches: 1

Range 1: 7 to 299 [GenPept](#) [Graphics](#) ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
385 bits(990)	2e-129	Compositional matrix adjust.	188/305(62%)	225/305(73%)	13/305(4%)
Query 108	RHLDPSLANYL	VQKAKQRRALRRWEQELNAKRSHLGRITVENEVDLDGPPRAFVYINEYR	167		
Sbjct 7	+ L P++A Y+V+KAKQR AL+RW+ ELN +++H G I VEN VDL+GPP F YINEY+	KTLKPAIAEYIVKKAKQRIALQRWQDELNRRKNHKGMIFFVENTVDLEGPPSDFYIINEYK	66		
Query 168	VGEGITL-NQVAVGCECQDCLWAPTGGCCPGASLHKFAYNDQGQVRLRAGLPIYECNSRC	226			
Sbjct 67	GI+L N+ GC C DC + CCP + AYN Q+++ G PIYECNSRC	PAPGISLVNEATFGCSTDCFFQK---CCPAEAGVLLAYNKNQIQIKIPPGTPIYECNSRC	123		
Query 227	RCGYDCPNRVVQKGI RYDLCIFRTDDGRGWGVRTLEKIRKNSFVMEYVGEIITSEEAERR	286			
Sbjct 124	+CG DCPNR+VQKG +Y LCIFRT +GRGWG+TL KI++ SFVMEYVGE+ITSEEAERR	QCGPDCPNRIVQKGTQYSLCIFRTSNGRGWGVKTLVKIKRMSFVMEYVGEVITSEEAERR	183		
Query 287	GQIYDRQGATYLFDLDYVEDVYTVDAAYYGNISHFVNHS CDPNLQVYVNFIDNLDRLPR	346			
Sbjct 184	GQ YD +G TYLFDLDY D +TVDAA YGN+SHFVNHS CDPNLQV+NVFIDNLD RLPR	GQFYDNKGIT YLFDLDYSEDEFTVDAARYGNVSHFVNHS CDPNLQVFNVFIDNLDTRLPR	243		
Query 347	IAFFATRTIRAGEELTFDYNMQVDPVDMESTRMSNFGLAGLPGSPKKRVRIECKCGTES	406			
Sbjct 244	IA F+TRTI AGEELTFDY M+ D+ S +D + KKRVR CKCG +	IALFSTR TINAGEELTFDYQMK-GSGDISSDSIDHS-----PAKKRVRTVCKCGAVT	294		
Query 407	CRKYL 411				
Sbjct 295	CR YL	CRGYL 299			

Figure 3.2: PSI-BLAST search result showing sequence identity between the sequence of Suv39h1 corresponding to the sequence of Suv39h2 crystal structure

In the figure, query sequence is the Suv39h1 amino acid sequence from residues 111 to 411; the subject sequence is the amino acid sequence of Suv39h2 (PDB id: 2R3A). This alignment shows that there is a 62% sequence identity between the sequence of Suv39h1 corresponding to the amino acid sequence of the crystal structure of Suv39h2. The first 110 amino acids of Suv39h2 were not crystallized in the 2R3A structure, hence there is a 110 amino acid shift in the numbering of the the sequences during the sequence alignment.

```

hSUV39h2_SET 1 Q Y S L C I F R T S N G R G W G V K T L V K I K R M S F V M E Y V G E V 36
hSUV39h1_SET 1 R Y D L C I F R T D D G R G W G V R T L E K I R K N S F V M E Y V G E I 36
                Y L C I F R T G R G W G V . T L K I . . S F V M E Y V G E .

hSUV39h2_SET 37 I T S E E A E R R G Q F Y D N K G I T Y L F D L D Y E S D E F T V D A A 72
hSUV39h1_SET 37 I T S E E A E R R G Q I Y D R Q G A T Y L F D L D Y V E D V Y T V D A A 72
                I T S E E A E R R G Q Y D G . T Y L F D L D Y D . T V D A A

hSUV39h2_SET 73 R Y G N V S H F V N H S C D P N L Q V F N V F I D N L D T R L P R I A L 108
hSUV39h1_SET 73 Y Y G N I S H F V N H S C D P N L Q V Y N V F I D N L D E R L P R I A F 108
                Y G N . S H F V N H S C D P N L Q V . N V F I D N L D R L P R I A

hSUV39h2_SET 109 F S T R T I N A G E E L T F D Y Q M K G S 129
hSUV39h1_SET 109 F A T R T I R A G E E L T F D Y N M Q V D 129
                F T R T I A G E E L T F D Y . M .

```

Figure 3.3: Sequence alignment of the SET domains of human Suv39h1 and human Suv39h2

The SET domain is highly conserved among the histone lysine methyltransferase family. This figure shows that there is 77% sequence identity between the SET domains of human Suv39h1 and human Suv39h2.

3.2.2. Homology model prediction using the Swiss model server

Both the automated and alignment mode of the Swiss model server were used for homology modeling of Suv39h1. In the automated mode, the FASTA sequence of Suv39h1 was input into the server and in the alignment mode, the sequence alignment obtained from the previous step was input into the server.

The automated mode also chose 2R3A as the template structure, which was a validation for our previous step thus confirming that it is the best option available as a template for the modeling of Suv39h1. Since identical results were obtained from both modes, the automated result was chosen for further work (Figure 3.4).

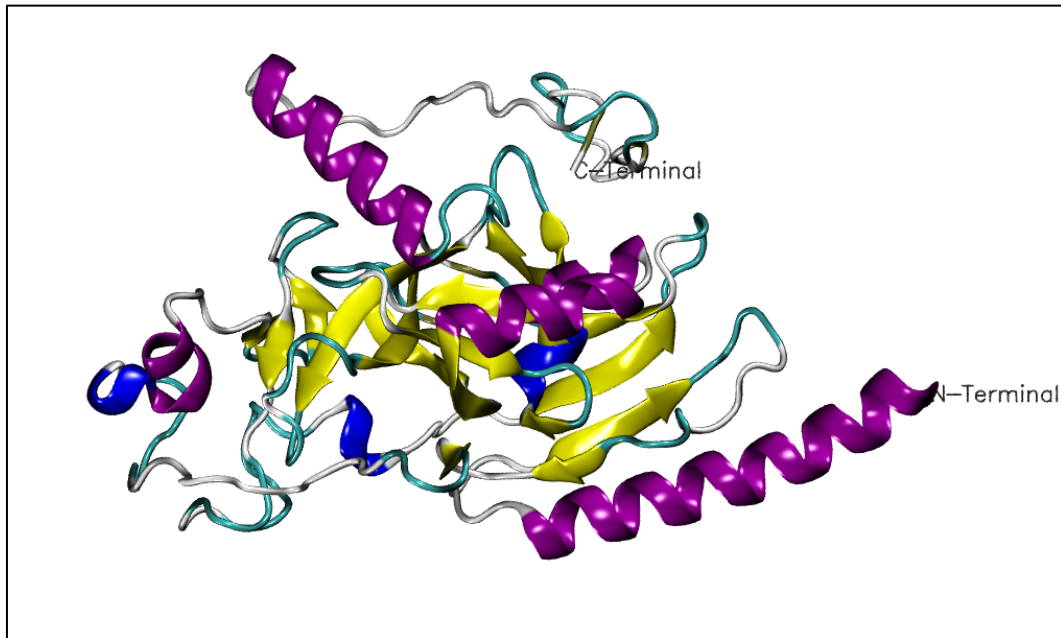


Figure 3.4: Homology model obtained from the Swiss model homology-modeling server.

The N-terminal begins with Alanine 115 (ALA115) and the C-terminal represents the last amino acid of the protein. The amino acids have been colored based on their secondary structure properties. The α -helices are represented in purple, β -sheets in yellow, loops are in white and the folds are in cyan.

3.2.3. Quality assessment of predicted model

The quality of the predicted model obtained from the Swiss model server was evaluated by the following criteria:

3.2.3.1. ANOLEA

Anolea evaluates the non-local environment of each heavy atom in the molecule. The y-axis of the plot represents the energy for each amino acid of the protein chain. The negative energy values (in green) represent a favourable energy environment whereas positive values (in red) represent an unfavourable energy environment for a given amino acid. For Suv39h1, majority of the residues are in favourable states except for a stretch of amino acids at the C-terminal (Figure 3.5).

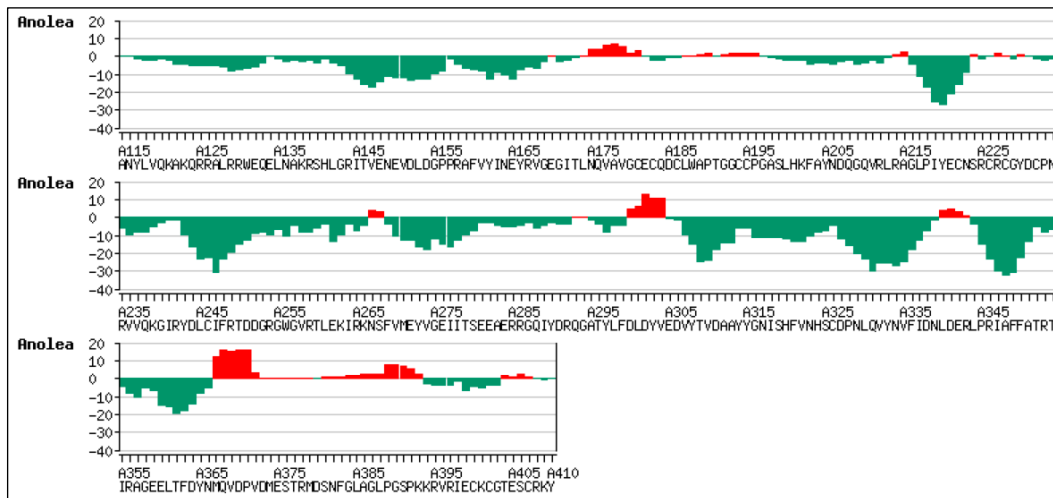


Figure 3.5: ANOLEA plot for the homology model of Suv39h1 obtained from Swiss model server.

The Y-axis represents the ANOLEA score and the amino acids of Suv39h1 are listed on the X-axis.

3.2.3.2. QMEAN

QMEAN estimates the global quality of the entire model as well as evaluates the local per-residue analysis of different regions within a model. The per residue analysis shown in Figure 3.6 supplements the ANOLEA analysis by showing the same residues being in an unfavourable environment (in red). Another term called the QMEAN Z-score, a measure of the absolute quality of the model is -2.5 (Figure 3.7).

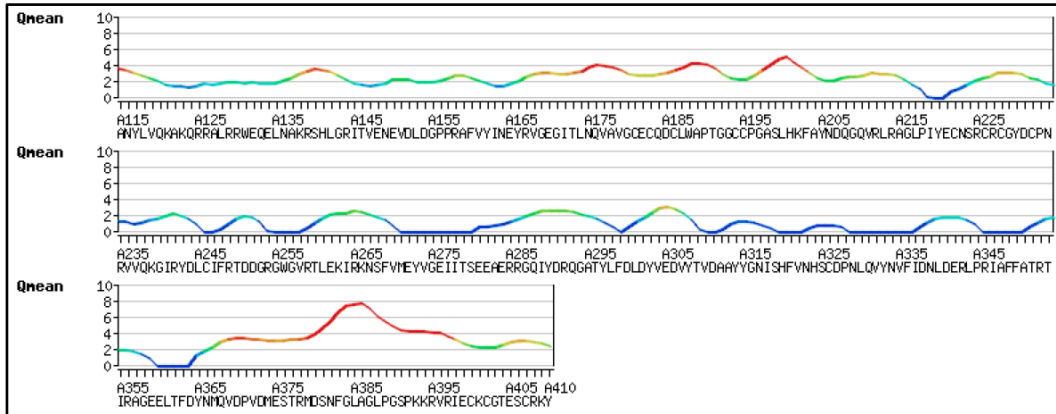


Figure 3.6: QMEAN plot for the homology model obtained from Swiss model server.

This model has a QMEAN4 score of 0.61 (estimates the global model quality based on scores between 0 to 1, the model is better if its closer to 1).

Comparison with non-redundant set of PDB structures

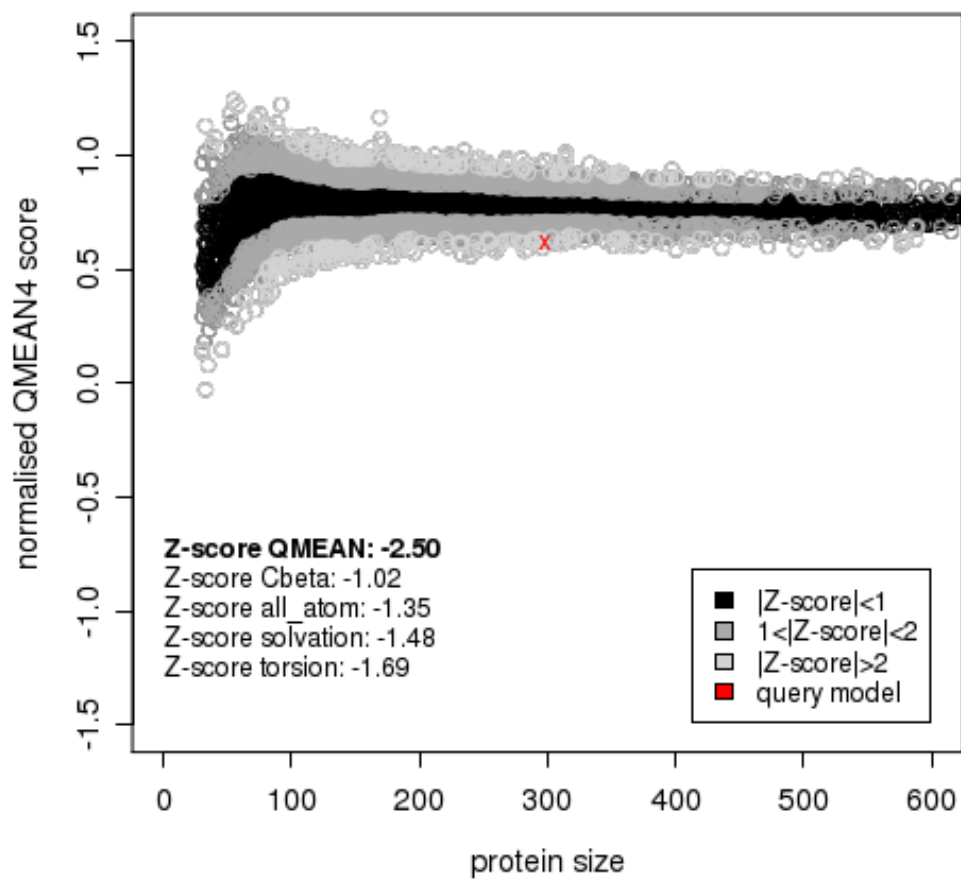


Figure 3.7: The QMEAN Z-score for the homology model of Suv39h1 obtained from Swiss model server

3.2.3.3. DFire

The predicted model has a DFire energy score of -390.65.

3.2.3.4. PROCHECK

The main component of the PROCHECK suite of programs is the Ramachandran plot. As per the Ramachandran plot, 99.6% of the residues were in the allowed regions as per the plot (83.5% in the most favored regions, 14.2% in allowed regions and 1.9% in generously allowed regions); whereas only 0.4% of the residues modeled were in the disallowed regions as shown in Figure 3.8.

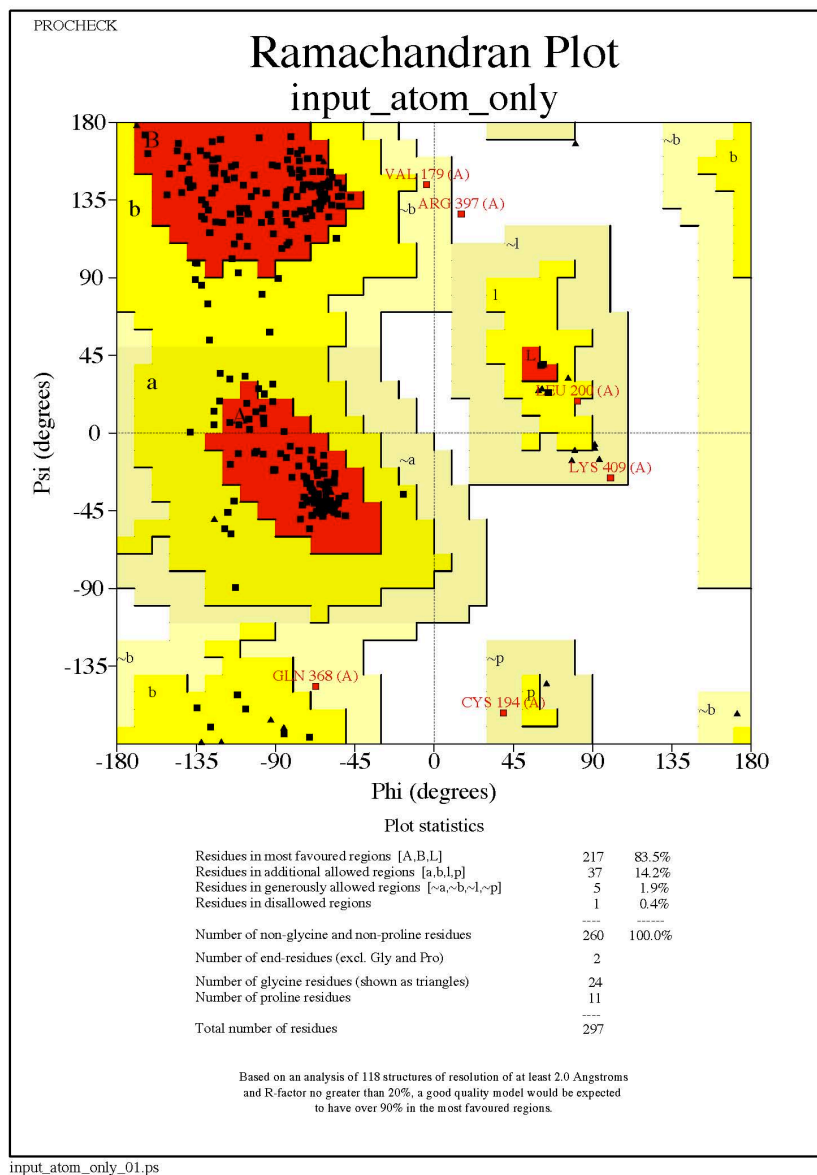


Figure 3.8: The Ramachandran plot for the homology model of Suv39h1 obtained from Swiss model.

This plot shows that 99.6% of the residues are in stereochemically allowed regions. Residues in the red regions correspond to the core alpha, beta and left-handed alpha and are in the most favored regions. The residues in the dark yellow regions are in the allowed regions and the residues in the light yellow region are in the generously allowed region. Except for ARG 397 (which is on the edge of the generously allowed regions), all residues are in the stereochemically allowed regions of the plot proving that this model is acceptable.

3.2.4. Energy minimization and MD simulation studies

The homology model was further refined by energy minimization and MD simulations. MD simulations were done for 23ns, when the total energy of the system was equilibrated and tried to reach its global energy minimum as shown in the plot below. The MD simulations were performed in four stages: for the first 0.05ns, restrained energy minimization of the whole system (target and solution) was performed, with target structures restrained with respect to the initial coordinates to remove water molecules surrounding the target that are too close to it. Second, from $t = 0.05\text{ns}$ to $t = 0.1\text{ns}$, energy minimization was done again without restrains on target structures to remove close contacts within the targets. Third, from $t = 0.1\text{ns}$ to $t = 0.2\text{ns}$, short MD simulation was performed on the whole system with the target structure restrained, in order to equilibrate the distribution of water molecules in the system. Finally, long MD simulation up to 23 ns was conducted without restrains to equilibrate the target structure.

MD simulations revealed that after the initial reduction from -1.5×10^5 to -1.55×10^5 , the potential energy of the model remained constant at this level up to 23ns (Figure 3.9). Although this energy minimum was reached by the 2ns time point, MD simulation was performed over a much longer time period to ensure that this was the global energy minimum and not a local energy minimum.

The achievement of the global energy minimum by the system so early during the simulation validates the good quality of the homology model. Except for slight adjustments in the positions of the side chains of the amino acids, there were no major changes to the homology model as shown in Figure 3.10.

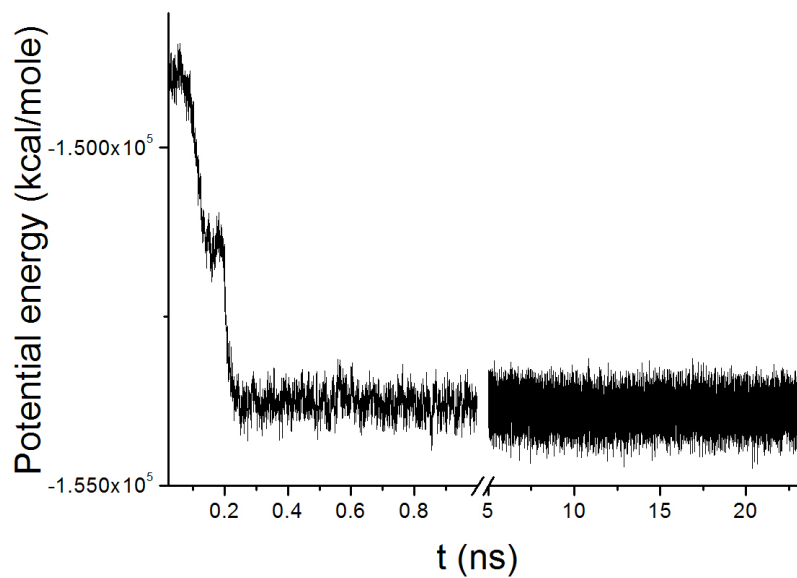


Figure 3.9: Plot of the potential energy of the system over a period of 23ns MD simulation.

The initial time period, from $t = 0$ to $t = 0.2$ ns represents restrained energy minimization of the system. The un-restrained MD simulation was performed from $t = 0.2$ ns to 23ns.

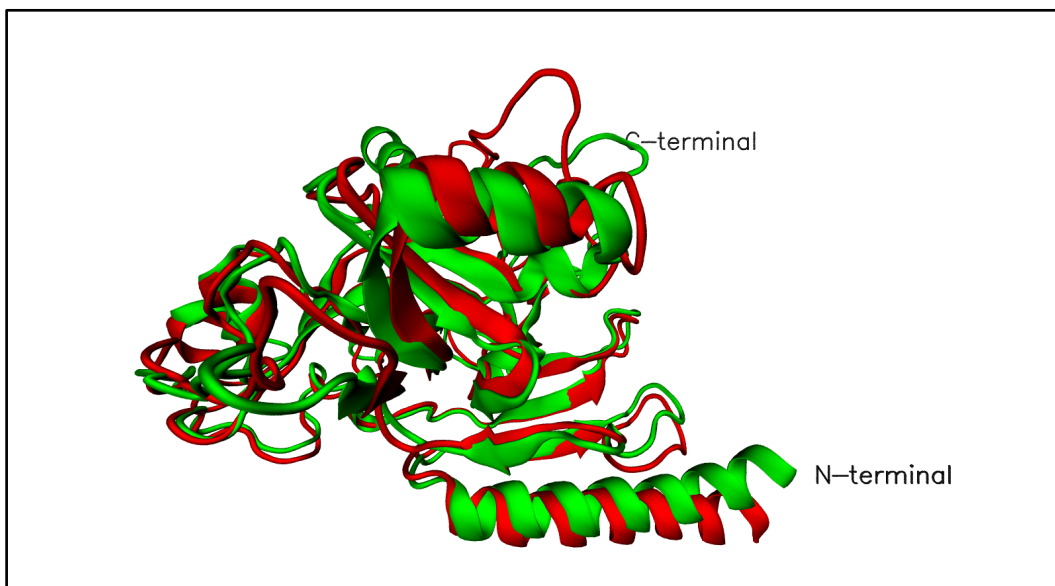


Figure 3.10: Superposition of the original homology model with the model after MD simulation.

The original model obtained from the Swiss model server is shown in green and the MD refined model is shown in red. There is no change in the α -helices and β -sheets, whereas there is a significant change in the positions of the loop structure due to MD simulations.

3.3. Calculation of RMSD of the simulated structures

The change in the overall shape of the model that is possible during MD simulations was calculated by plotting the co-ordinates of the carbon-alpha (C- α) atoms as a function of time, compared to the original homology model obtained from the Swiss model server.

As seen in the Figure 3.11, there is a rapid change in structure during the initial time period that corresponds to the removal of restraints during simulations. However, the structures stabilize around the 3.5-4 ns time point, and very little difference is seen in the structures from there on up to 23 ns. There is an overall change of around 4Å in the structures at the final time point compared to the original structure. Visual inspection of the structure shows that much of this change can be attributed to loop regions, and there is not much change in the co-ordinates of the α -helices and β -sheets as seen in the figure above.

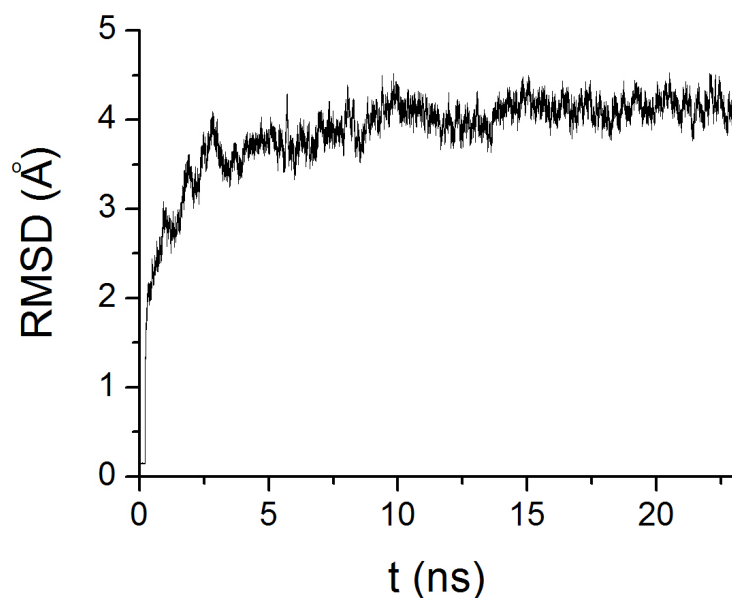


Figure 3.11: A plot of the root mean square deviation of the C- α backbone of the model structures over a period of 23ns

The RMSD at time point = 0 represents the co-ordinates of parent structure. There is a significant change in the RMSD of the C- α atoms during MD simulation from the beginning upto approximately 4ns. At this time point the structure seems to have reached a stable state and there is very little change in the structure until the end as seen in the plot.

3.4. Clustering of the simulated structures

The RMSD plot reveals that the structure stabilizes around the 15 ns time point and there is no more change seen up to 23ns. Therefore the simulated structures from 15 ns to 23 ns were sampled every 0.2 picoseconds (ps) using the ptraj module in the AMBER suite of programs generating a total of 4000 structures. These structures were clustered based on nearest neighbor clustering analysis by using the program MaxCluster (<http://www.sbg.bio.ic.ac.uk/~maxcluster/>). The centroid structure of the largest cluster was chosen as the representative structure and was used for further experiments.

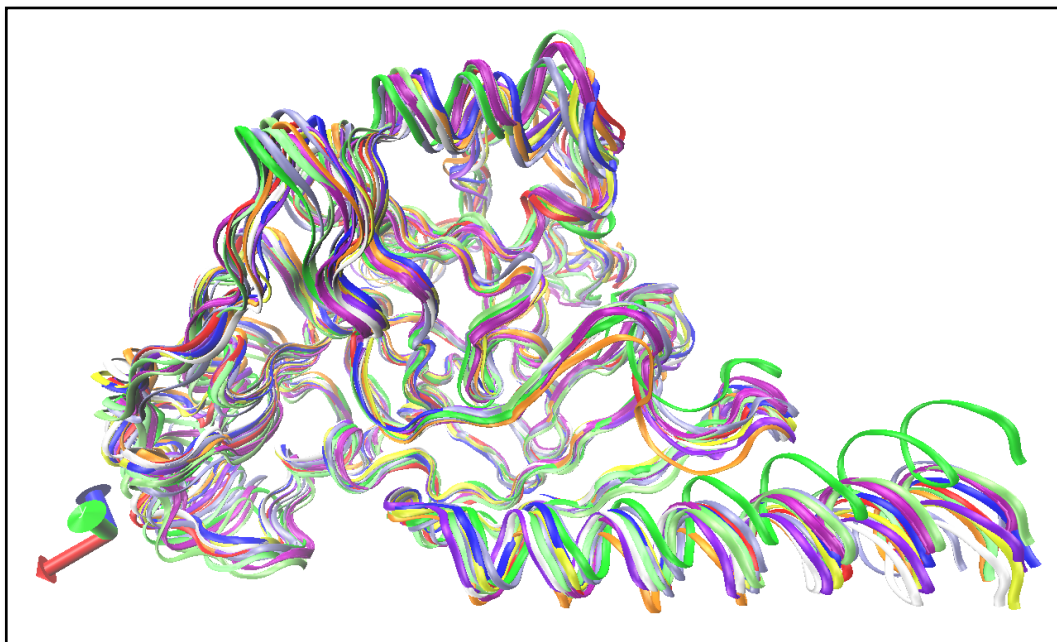


Figure 3.12: Visualization of the structures at different time points during MD simulations.

These structures were clustered and the centroid structure of the largest cluster was chosen a representative refined structure of Suv39h1.

3.5. Determination of potential ligand binding sites

Three ligands are known to bind to Suv39h1: the methyl group donor S-adenosyl methionine, N-terminal of histone H3 and chaetocin (small molecule inhibitor of Suv39h1). However, none of their binding sites on the surface of Suv39h1 are known. Therefore a blind docking experiment was performed with the homology model of Suv39h1 to determine the binding sites of each of these ligands.

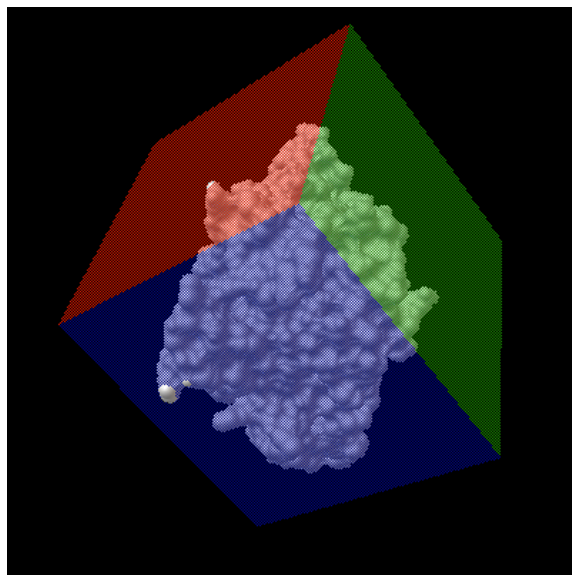


Figure 3.13: Visualization of the blind docking method.

Suv39h1 homology model is encapsulated in a three-dimensional grid box and the ligand is used to search for all possible binding sites on its surface. The binding poses of the ligand at the potential binding sites are then scored and clustered based on the predicted binding energies at each site.

3.5.1. SAM binding site

The SAM binding site was determined by a combination of blind docking and observation of the SAM binding site on other histone methyltransferases. After blind docking, the binding affinity of SAM at different sites was ranked and the binding conformations (poses) were clustered based on their binding energies within a RMSD of 2 Å as shown in

Figure 3.14. The cluster of poses with the highest binding energy was -6.5 kCal/mole. The clustering showed many clusters with only 1 or 2 conformations, two clusters with three conformations and only one cluster with four conformations. The highest ranking binding conformations and the cluster with the largest number of conformations were visually analyzed using AutoDock tools.

The SUV39h2 crystal structure that was used as a template to build the homology model was crystallized with SAM bound in its binding site. Since the homology model of Suv39h1 and the template structure share a very similar structure, the binding conformations predicted by blind docking and the binding conformation of SAM on SUV39h2 were compared as shown in Figure 3.15. This reveals that the third highest cluster with a binding energy of ~5.8 kCal/mole (Figure 3.14) is the conformation similar to the one crystallized with SUV39h2. Therefore 5.8 kCal/mol was chosen as the cut off point for determining small molecules with high binding potential.

Further analysis of this binding conformation reveals that SAM interacts with the following residues in its active site: ARG140, TRP142, TYR183,

HIS206, ASN209, HIS210, LYS287, PHE207, CYS212, CYS288 and THR290 as shown in

Figure 3.16. This binding pocket was used to screen for small molecule inhibitors.

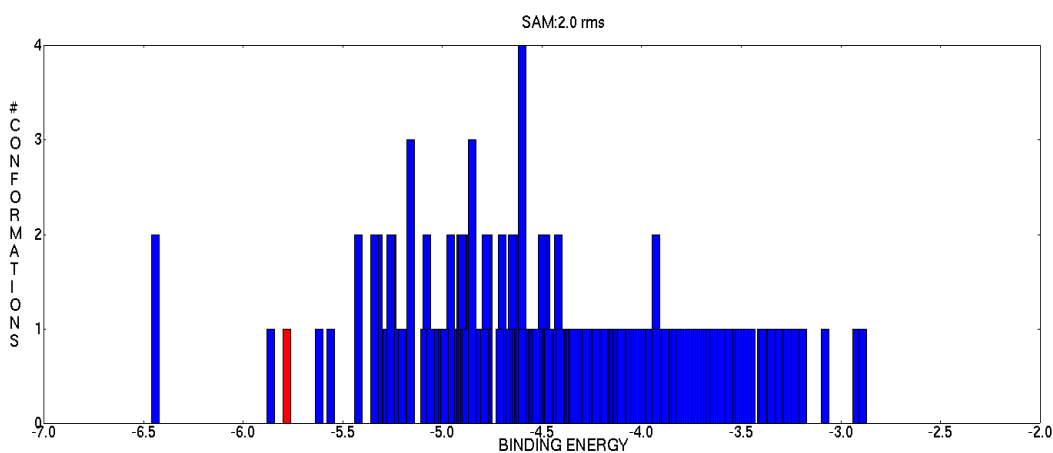


Figure 3.14: Clustering of binding conformations of SAM on Suv39h1 based on binding energies upon blind docking

The cluster in red color corresponds to the binding energy of the correct binding conformation of SAM on the Suv39h1 homology model upon blind docking as determined by comparison with the crystal structure SAM bound SUV39h2. Although there are two other clusters with higher predicted binding energies and 3 clusters with more conformations that bind at other sites, the correct binding site was chosen based on knowledge obtained from the crystallization data.

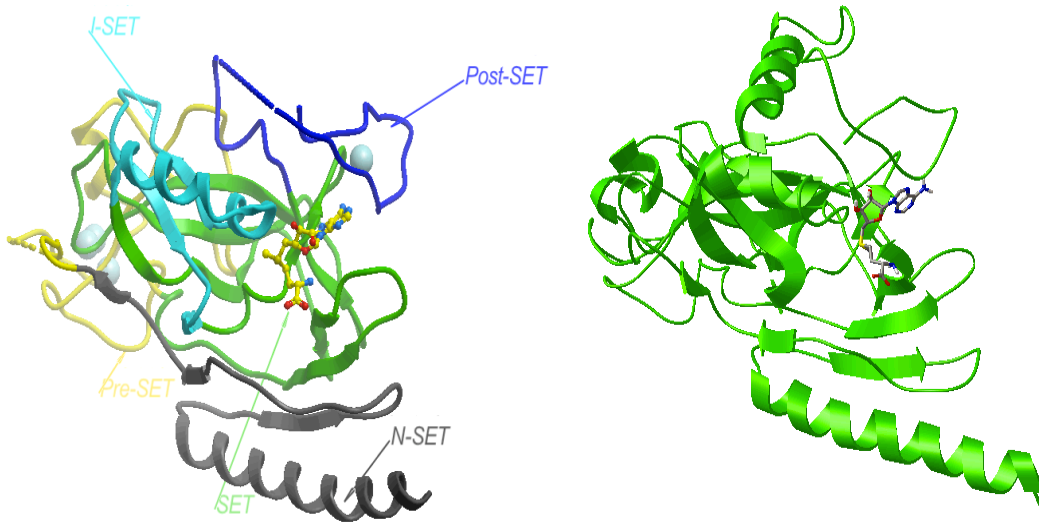


Figure 3.15: Comparison of the binding conformation of SAM in SUV39h2 crystal structure and Suv39h1 homology model.

The structure at left is the crystal structure of SUV39h2 co-crystallized with SAM (Wu, Min et al. 2010); the protein structure is coloured based on its secondary structure and SAM is in yellow. The structure at right is the homology model of Suv39h1 docked with SAM in its binding pocket. This binding conformation was chosen by blind docking and comparison with the structure of SUV39h2 at right.

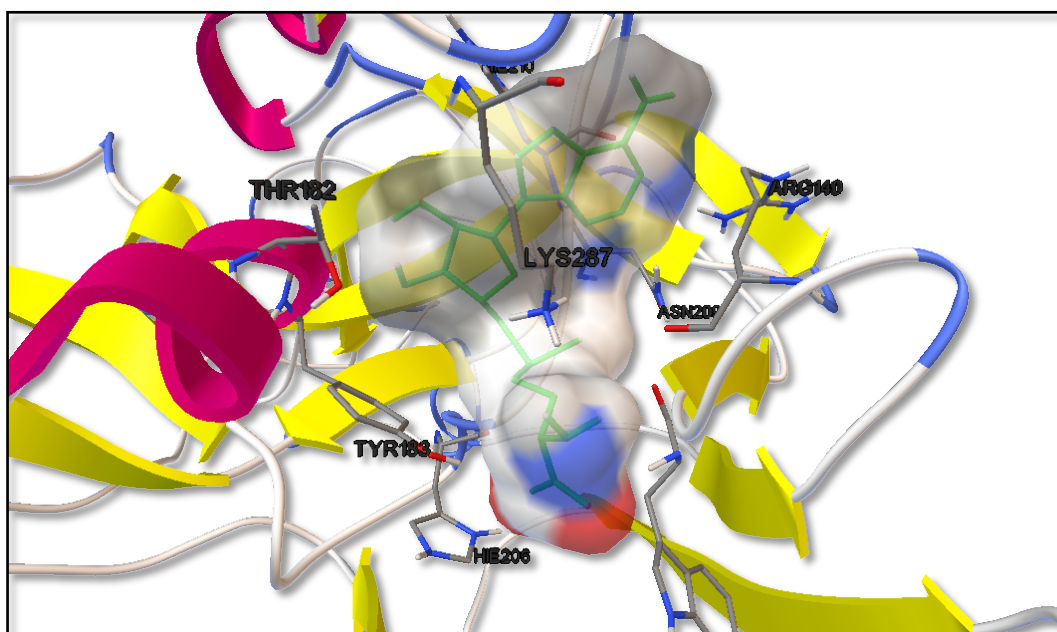


Figure 3.16: Interaction map of SAM binding to various residues of Suv39h1 in its binding pocket.

The residues making contacts are labeled and SAM (green) is shown in surface representation. ARG140, TRP142, TYR183, HIE206, ASN209, HIE210, LYS287, PHE207, CYS212, CYS288 and THR290 are the residues that interact with SAM in its binding pocket.

3.5.2. Chaetocin binding site

Chaetocin was docked blindly onto the homology model of Suv39h1 with the protocol described for SAM. The clustering of the docked conformations based on their binding energies revealed one binding site that was the most favored compared with more than 60 conformations at this site with a binding energy of -10.72 kCal/mole as shown in Figure 3.17. This is in contrast to the blind docking results of SAM, which did not show a large cluster favoring one binding spot (Figure 3.14). The binding site of chaetocin is distinct from the SAM binding site that was determined in the previous section as shown in Figure 3.18. This suggests that chaetocin could be a non-competitive inhibitor of human Suv39h1 and inhibit the methyltransferase activity by allosteric mechanism of inhibition, which is in contrast to the earlier report where it was suggested that chaetocin acts as a competitive inhibitor for SAM (Greiner, Bonaldi et al. 2005). This is in agreement with a recent report indicating that the inhibition of methyltransferase activity by chaetocin is not dependent on the availability of SAM (Cherblanc, Chapman et al. 2013).

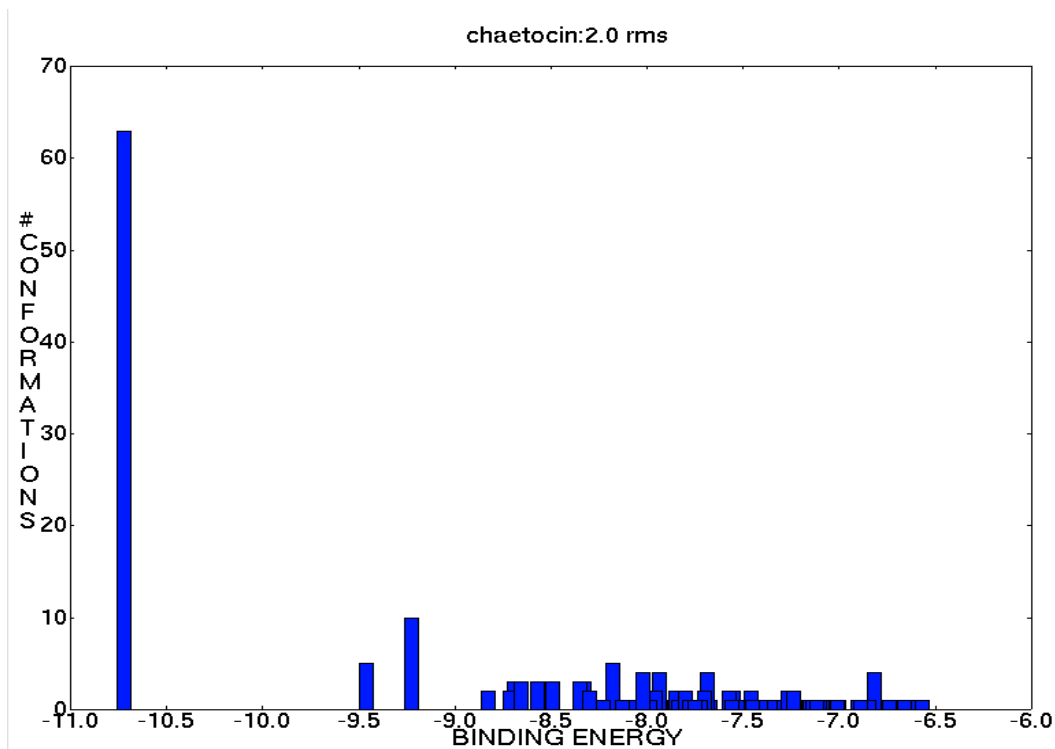
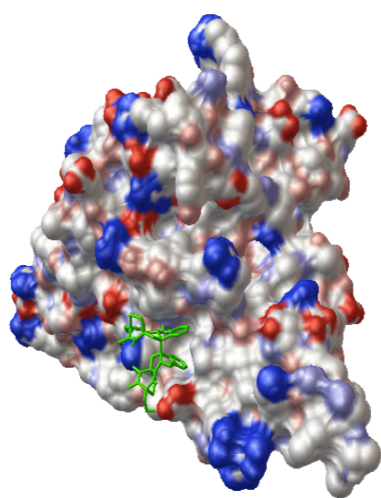
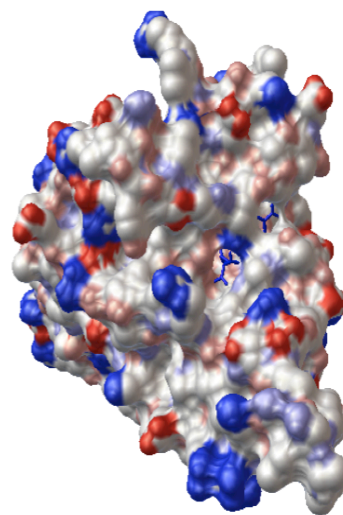


Figure 3.17: Cluster analysis of blind docking results of chaetocin

The plot of the binding conformations versus the binding energies at different binding sites upon blind docking shows a distinct cluster of bound conformations at one particular binding site with a binding energy of -10.72 kcal/mole



Chaetocin binding pocket



SAM binding pocket

Figure 3.18: Comparison of the binding pockets of chaetocin and SAM

SAM and chaetocin have been found to bind at different binding sites on Suv39h1 upon blind docking as seen in this figure. The structure at left represents chaetocin docked into Suv39h1 and the structure at right shows SAM docked in its binding pocket. Chaetocin and SAM are in green and the Suv39h1 is coloured based on the charge of the amino acids.

3.5.3. Histone H3 N-terminal binding site

The binding site of H3K9 peptide was determined by comparing the structures of the Suv39h1 homology model and the EHMT1 crystal structure (PDB id: 3HNA) that has been co-crystallized with histone H3 N-terminal peptide monomethylated at lysine 9 (Wu, Min et al. 2010). The structures of histone methyltransferases are conserved and adopt a typical fold consisting of a conserved SET domain flanked by Pre- and Post-SET domains and characterized by a distinct co-factor and substrate binding areas and a narrow substrate lysine docking channel (Cheng, Collins et al. 2005; Qian and Zhou 2006).

3.5.3.1. Histone H3 N-terminal binding site on Suv39h1 by structural analysis

EHMT1 is the closest methyltransferase to be crystallized with the N-terminal of histone H3 peptide bound to it. An amino acid sequence alignment of the SET domains of EHMT1 and Suv39h1 shows 68% sequence similarity (Figure 3.19). Wu et al. describe the binding of H3K9 peptide to EHMT1 in a narrow groove formed by the I-SET domain and post-SET domain (Wu, Min et al. 2010). An interaction map representing the interaction of H3K9 peptide with EHMT1 was generated using LigPlot+ software. This map shows that H3K9 peptide interacts with TYR1124, ASP1131, ALA1134, ASP1135, ASP1140, ASP 1145, ASN1148, TYR1211 and ARG1214 (Figure 3.20). The groove made by the residues corresponding to these residues on Suv39h1 was identified by visual inspection and used for virtual screening (Figure 3.21).

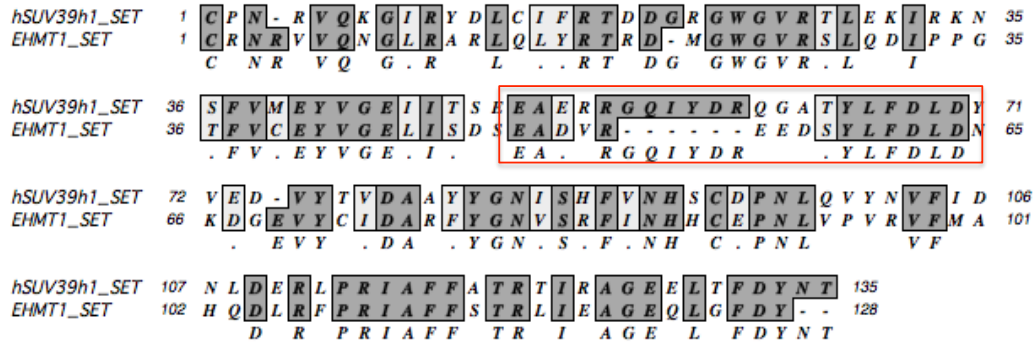


Figure 3.19: Sequence alignment of SET domain of Suv39h1 and EHMT1 shows 68% sequence similarity

The highlighted region shows the alignment of the residues that form the lysine docking channel.

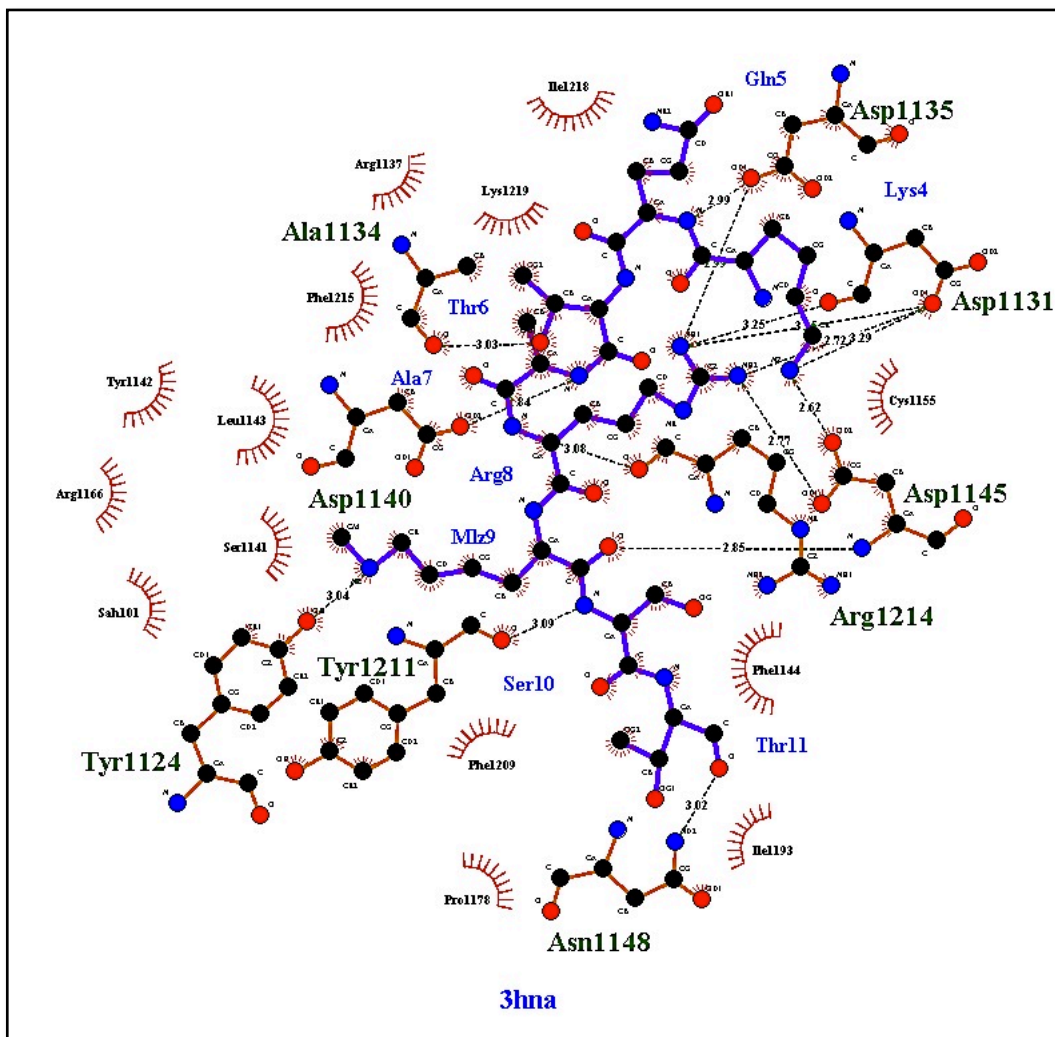


Figure 3.20: 2D interaction map showing the interaction of mono-methylated H3K9 peptide with EHMT1

This map shows that TYR1124, ASP1131, ALA1134, ASP1135, ASP1140, ASP 1145, ASN1148, TYR1211 and ARG1214 of EHMT1 interact with the mono-methylated peptide.

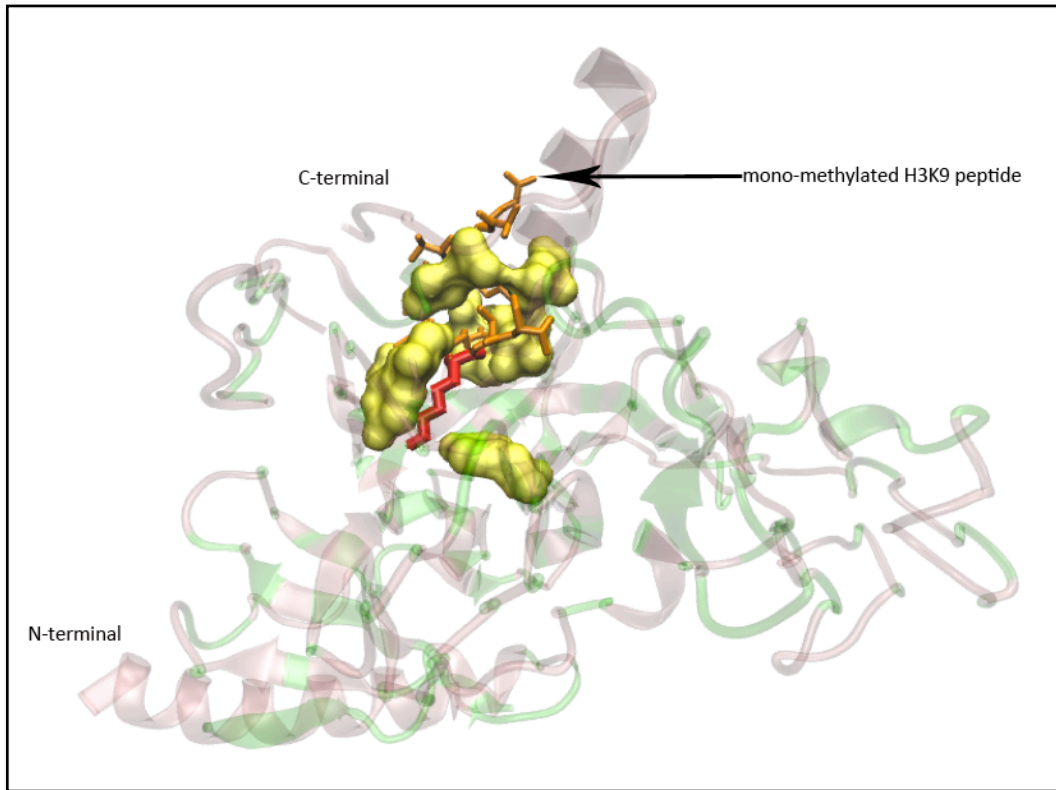


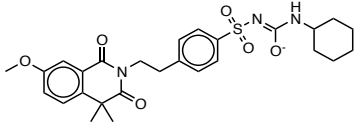
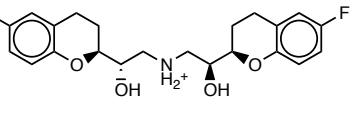
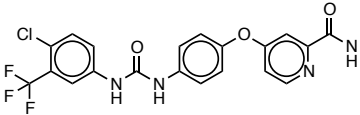
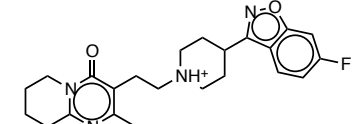
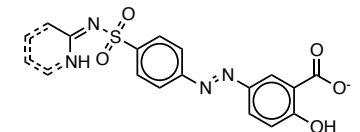
Figure 3.21: Superposition of EHMT1 and Suv39h1 to determine the binding pocket of H3K9 pocket

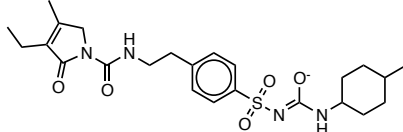
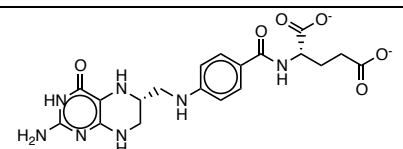
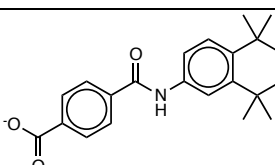
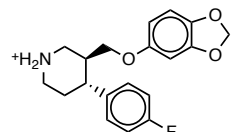
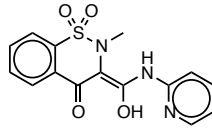
The Suv39h1 structure is in pink and EHMT1 is in green. The mono-methylated H3K9 peptide is coloured in orange and the mono methylated K9 residue is coloured in red. The residues of EHMT1 interacting with the peptide are coloured in yellow.

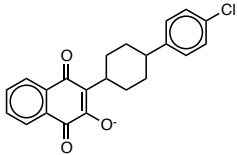
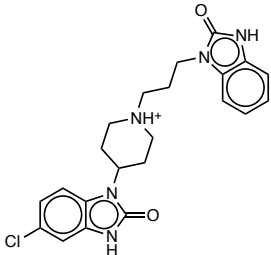
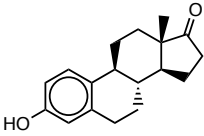
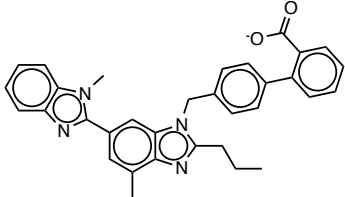
3.6. Small molecule inhibitors obtained from virtual screening

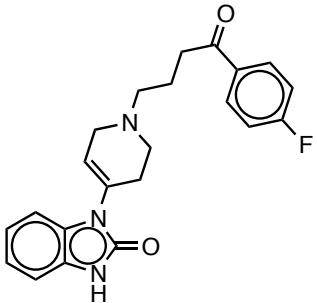
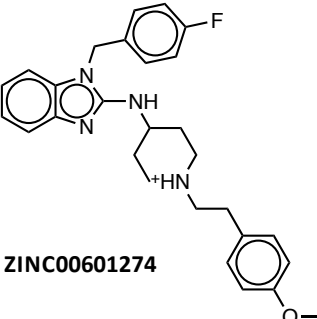
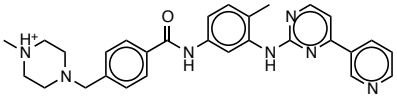
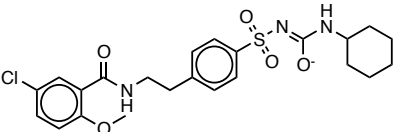
The binding sites of SAM, chaetocin, and N-terminal peptides of histone H3 were targeted and virtual screening was performed in order to find small molecule inhibitors that bound to human Suv39h1 with high affinity. The small molecules predicted to have high binding affinity are listed in Table 3.2.

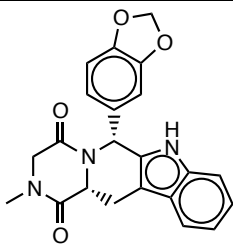
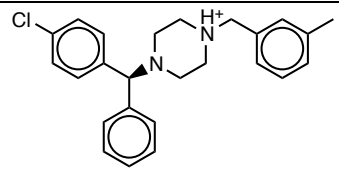
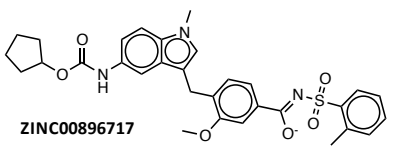
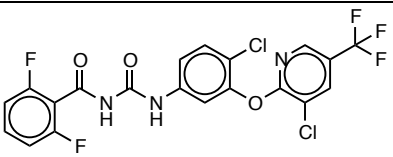
The table consists of the 2-D structure of the small molecules, molecular weight, predicted binding energy (kCal/mol) from AutoDock Vina, partition coefficient (logP), ADMET risk prediction value (Absorption, Distribution, Metabolism, Elimination and Toxicity), and the database to which each molecule belongs. The ADMET risk prediction value was determined by ADMET predictor (Simulations*plus* inc.), a software program that rapidly estimates a number of ADMET properties of small molecules based on their molecular structure. A lower number means lower risk of the molecule failing as a drug and subsequently a higher number corresponds to high risk of the molecule failing as a drug.

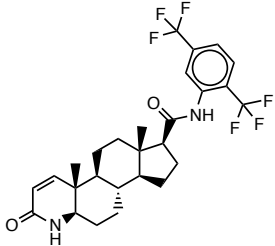
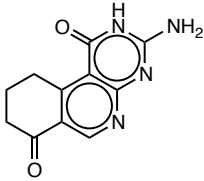
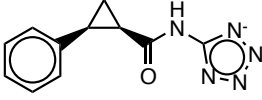
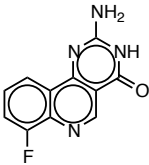
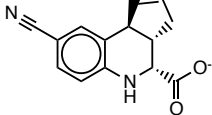
Serial number	Structure	Molecular Weight (kDa)	Binding energy (kCal/mole)	LogP	ADMET risk prediction	Database	Binding site
1	 <p>ZINC01482077</p>	527.64	-10.7	3.28	5	DrugBank	SAM
2	 <p>ZINC15668997</p>	405.44	-10	3.1	2	DrugBank	SAM
3	 <p>ZINC01493878</p>	464.83	-9.8	4.76	6	DrugBank	SAM
4	 <p>ZINC00538312</p>	410.49	-9.7	2.96	4	DrugBank	SAM
5	 <p>ZINC13540266</p>	398.4	-9.6	4.08	4	DrugBank	SAM

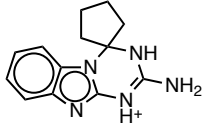
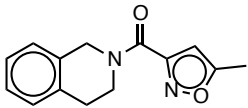
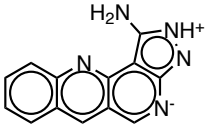
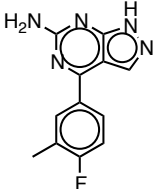
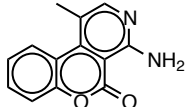
6	 <p>ZINC00537791</p>	490.63	-9.6	3.44	4	DrugBank	SAM
7	 <p>ZINC04228235</p>	445.44	-9.4	2	6	DrugBank	SAM
8	 <p>ZINC00538415</p>	351.45	-9.4	5.37	2	DrugBank	SAM
9	 <p>ZINC00527386</p>	329.37	-9.3	4.05	1	DrugBank	SAM
10	 <p>ZINC12466469</p>	331.35	-9.3	0.88	1	DrugBank	SAM

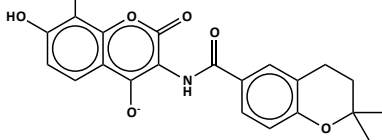
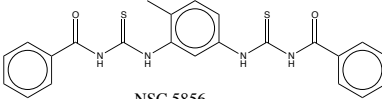
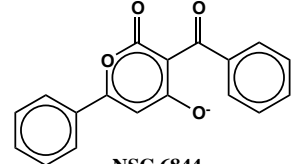
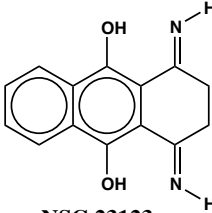
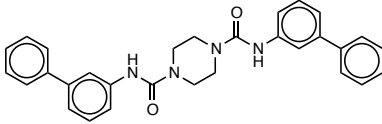
Serial number	Structure	Molecular Weight (kDa)	Binding energy (kCal/mole)	LogP	ADMET risk prediction	Database	Binding site
11	 <p>ZINC12504271</p>	366.85	-9.3	4.96	6	DrugBank	SAM
12	 <p>ZINC19632603</p>	425.92	-9.3	3.31	2	DrugBank	SAM
13	 <p>ZINC13509425</p>	270.37	-9.2	3.24	1	DrugBank	SAM
14	 <p>ZINC01530886</p>	514.63	-7.7	7.6	8	DrugBank	Chaetocin

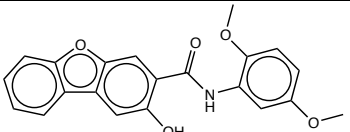
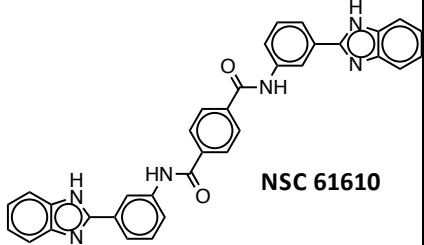
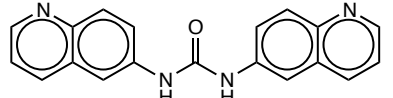
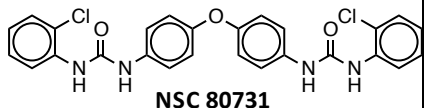
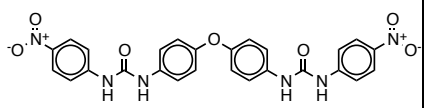
15	 <p>ZINC19796080</p>	379.44	-7.5	3.4	2	DrugBank	Chaetocin
16	 <p>ZINC00601274</p>	458.58	-7.8	5.71	4	DrugBank	H3K9
17	 <p>ZINC19632618</p>	493.62	-7.7	3.89	8	DrugBank	H3K9
18	 <p>ZINC00537805</p>	494.01	-7.7	4.4	3	DrugBank	H3K9

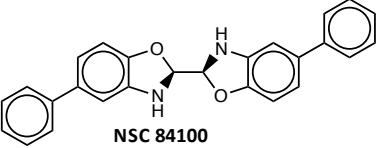
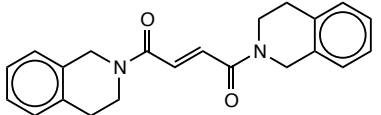
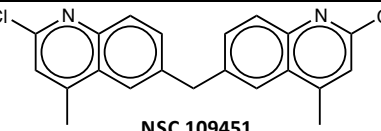
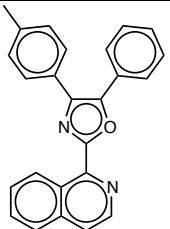
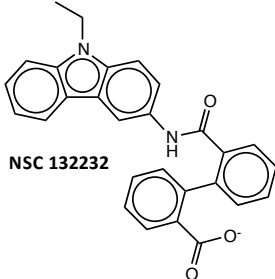
Serial number	Structure	Molecular Weight (kDa)	Binding energy (kCal/mole)	LogP	ADMET risk prediction	Database	Binding site
19	 <p>ZINC03993855</p>	389.41	-7.7	2.36	2	DrugBank	H3K9
20	 <p>ZINC19594557</p>	390.96	-7.5	5.9	5	ZDD	Chaetocin
21	 <p>ZINC00896717</p>	575.69	-8.8	5.75	7	ZDD	H3K9
22	 <p>ZINC02570819</p>	506.22	-8.5	5.68	6	ZDD	H3K9

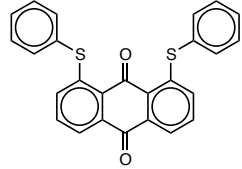
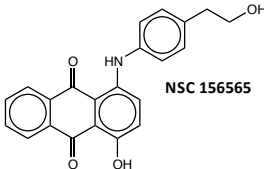
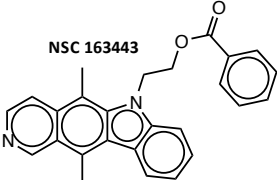
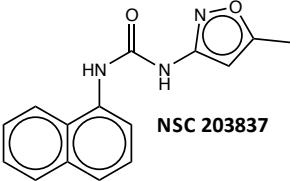
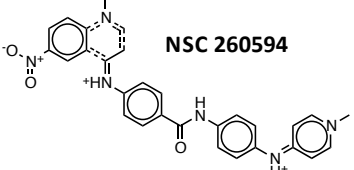
23	 <p>ZINC03932831</p>	528.54	-8.3	6.23	8	ZDD	H3K9
24	 <p>ZINC20673684</p>	230.227	-9.2	0.15	4	ZINC	SAM
25	 <p>ZINC16971127</p>	228.235	-8.9	0.96	3	ZINC	SAM
26	 <p>ZINC26507499</p>	230.202	-9.6	1.33	1	ZINC	SAM
27	 <p>ZINC08419668</p>	239.254	-9.0	0.77	3	ZINC	SAM

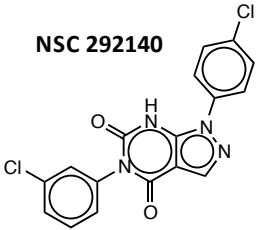
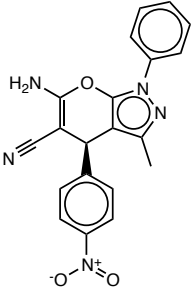
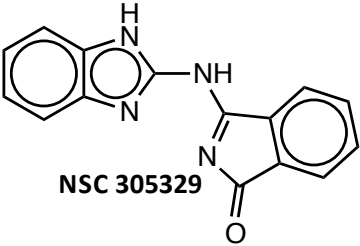
Serial number	Structure	Molecular Weight (kDa)	Binding energy (kCal/mole)	LogP	ADMET risk prediction	Database	Binding site
28	 ZINC13132400	242.306	-9.1	2.31	4	ZINC	SAM
29	 ZINC04750560	242.278	-8.9	1.71	1	ZINC	SAM
30	 ZINC06494057	235.25	-9.2	1.99	0	ZINC	SAM
31	 ZINC20572499	243.245	-9.0	2.04	3	ZINC	SAM
32	 ZINC19015026	226.235	-9.0	2.07	3	ZINC	SAM

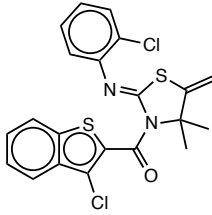
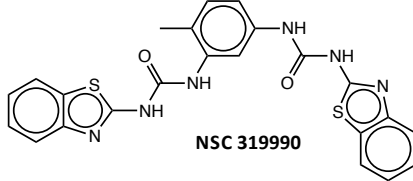
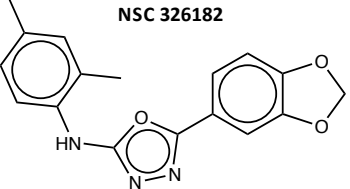
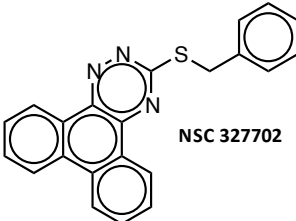
33	 <p>NSC 5157</p>	395.411	-9.1	6.58	1	NCI D2	SAM
34	 <p>NSC 5856</p>	448.57	-10.8	3.2	6	NCI D2	SAM
35	 <p>NSC 6844</p>	292	-9.1	3.72	1	NCI D2	SAM
36	 <p>NSC 23123</p>	240.26	-9.2	2.19	2	NCI D2	SAM
37	 <p>NSC 37553</p>	301.32	-9.4	6.47	9	NCI D2	SAM

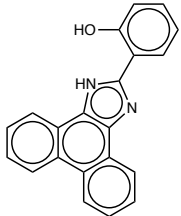
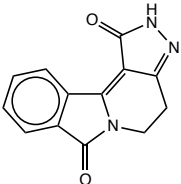
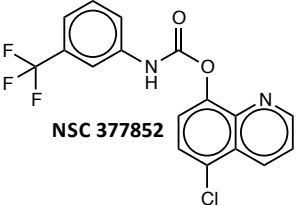
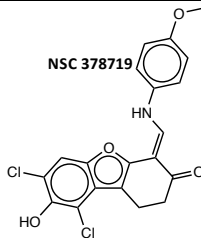
Serial number	Structure	Molecular Weight (kDa)	Binding energy (kCal/mole)	LogP	ADMET risk prediction	Database	Binding site
38	 <p>NSC 50650</p>	363.36	-9.0	4.94	4	NCI D2	SAM
39	 <p>NSC 61610</p>	548.6	-9.0	6.89	10	NCI D2	H3K9
40	 <p>NSC 71881</p>	392.458	-9.4	3.11	4	NCI D2	SAM
41	 <p>NSC 80731</p>	507.4	-8.1	7.36	8	NCI D2	H3K9
42	 <p>NSC 80735</p>	528.5	-8.1	6.02	8	NCI D2	H3K9

43	 <p>NSC 84100</p>	392.45	-9.0	6.5	6	NCI D2	SAM
44	 <p>NSC 87838</p>	346.42	-7.9	2.32	1	NCI D2	Chaetocin
45	 <p>NSC 109451</p>	367.279	-9.1	7.26	5	NCI D2	SAM
46	 <p>NSC 116702</p>	362.4	-10.2	6.18	6	NCI D2	SAM
47	 <p>NSC 132232</p>	433.5	-7.9	6.05	5	NCI D2	Chaetocin

Serial number	Structure	Molecular Weight (kDa)	Binding energy (kcal/mole)	LogP	ADMET risk prediction	Database	Binding site
48	 <p>NSC 156516</p>	424.5	-11.1	7.5	7	NCI D2	SAM
49	 <p>NSC 156565</p>	359.3	-10.6	4.89	3	NCI D2	SAM
50	 <p>NSC 163443</p>	394.4	-11.5	6.15	6	NCI D2	SAM
51	 <p>NSC 203837</p>	367.3	-9.2	3.13	6	NCI D2	SAM
53	 <p>NSC 260594</p>	506.6	-8.5	5.0	5	NCI D2	H3K9

54	<p>NSC 292140</p> 	373.2	-9.7	3.34	1	NCI D2	SAM
55	 <p>NSC 298892</p>	373.4	10.0	2.75	6	NCI D2	SAM
56	 <p>NSC 305329</p>	262.27	-9.5	3.62	2	NCI D2	SAM

Serial number	Structure	Molecular Weight (kDa)	Binding energy (kCal/mole)	LogP	ADMET risk prediction	Database	Binding site
57	 <p>NSC 310324</p>	447.41	-10.4	7	9	NCI D2	SAM
58	 <p>NSC 319990</p>	474.5	-10.4	5.84	7	NCI D2	SAM
59	 <p>NSC 326182</p>	309.32	-9.4	4.64	6	NCI D2	SAM
60	 <p>NSC 327702</p>	353.4	-10.1	5.98	8	NCI D2	SAM

61	 <p>NSC 332670</p>	310.35	-8.0	5.5	4	NCI D2	Chaetocin
62	 <p>NSC 335048</p>	239.23	-9.7	1.72	5	NCI D2	SAM
63	 <p>NSC 377852</p>	366.72	-9.4	4.78	5	NCI D2	SAM
64	 <p>NSC 378719</p>	404.2	-10.4	4.55	4	NCI D2	SAM

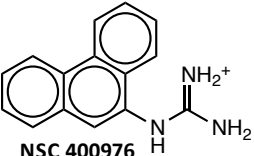
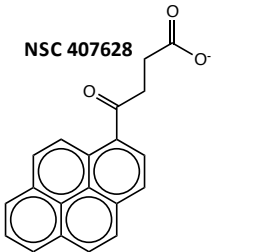
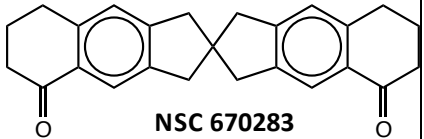
Serial number	Structure	Molecular Weight (kDa)	Binding energy (kCal/mole)	LogP	ADMET risk prediction	Database	Binding site
65	 NSC 400976	236.3	-9.2	3.17	3	NCI D2	SAM
66	 NSC 407628	301.32	-9.3	4.14	5	NCI D2	SAM
67	 NSC 670283	292	-10.3	4.49	6	NCI D2	SAM

Table 3.2: List of small molecules obtained from virtual screening that bind to Suv39h1 with high affinity.

The table lists the structures and id's of the compounds, molecular weight, predicted binding energy from virtual screening, LogP values, ADMET risk prediction, the database that the compound belongs to and the binding site for which it has the corresponding binding energy.

3.7 . Testing of the small molecule inhibitors by *in vitro* assay

An *in vitro* methyltransferase assay was optimized to test the small molecule inhibitors as described in section 2.14.

3.7.1 Fluorescent histone methyltransferase assay

The fluorescent histone methyltransferase assay was performed several times in order to optimize the controls. The samples used were: untransfected HEK293T cell lysate, pCEMM-NTAP-GS(GW)-h*SUV39h1*-855/856 transfected HEK293T cell lysate, HeLa cell lysate, lysates of D5 MEF (*suv39h1* and *suv39h2* double knockout MEF) and W8 MEF (wild type). The positive control was provided with the assay kit and the negative control was prepared by not adding any enzyme. Several issues were encountered while performing the assay. One of the major issues was the consistently high value obtained from the blank sample compared to the positive control. Numerous trials yielded inconsistent results. On one occasion only, the blank sample showed lower absorbance than the positive control; the result of which is shown in Figure 3.22. Even in this case, D5 (*SUV39h1/h2* double knock out MEF) shows methyltransferase activity similar to that of the control W8 MEF, untransfected HEK293T and HeLa cell lysates. These issues necessitated the use of a different strategy to test the small molecules.

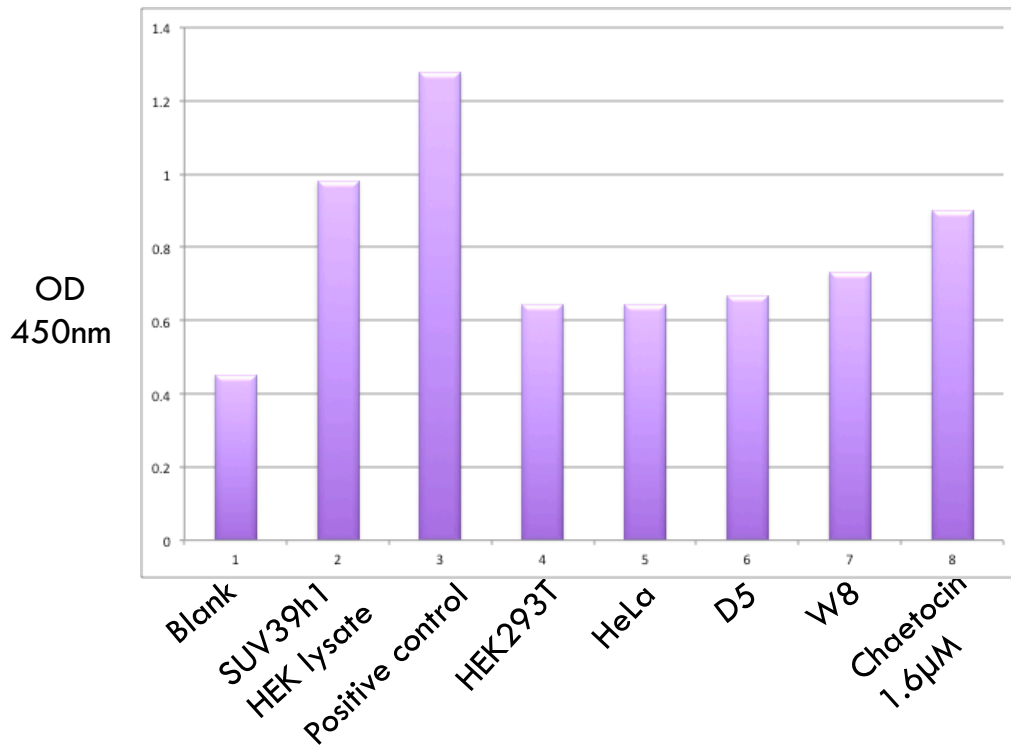


Figure 3.22: Results of fluorescent methyltransferase assay

The chart shows the methyltransferase activity present in each sample in the form of absorbance measured at 450nm. The samples used were: Blank - contains only the assay buffer and no other components, Suv39h1 HEK lysate - lysate from HEK293T cells transfected with pCEMM-NTAP-GS(GW)-h*SUV39h1*-855/856, positive control - supplied in the kit, HEK293T - untransfected HEK293T cell lysate, HeLa - untransfected HeLa cell lysate, D5 - cell lysate of the *SUV39h1* and *SUV39h2* double knock out MEF, W8 - cell lysate of the control (normal) MEF, Chaetocin - 1.6µM of chaetocin tested for inhibition of methyltransferase activity against *SUV39h1* HEK lysate.

3.7.2 Validation of the radioactive *in vitro* histone methyltransferase assay

This assay makes use of radioactively labeled methyl group (from SAM) as a methyl donor, which is incorporated directly onto the H3K9 residue upon methyltransferase reaction. The assay was validated with two positive controls and several negative controls as shown in Table 3.3. The unmethylated and mono-methylated histone H3 peptides were used as substrates for positive controls as they can be methylated by Suv39h1. Histone H3 peptide acetylated at the K9 residue was used as a substrate for negative control as it cannot be methylated. The amount of methyltransferase activity of the Suv39h1 was measured by the counts per minute of radioactivity obtained by scintillation counting. The counts per minute reflect the amount of tritiated methyl groups incorporated onto the H3K9 residue. The positive controls (i.e. H3K9 and H3K9me) show at least 40-fold increase in counts per minute (CPM) (Figure 3.23) over the negative control (i.e. H3K9Ac) thereby validating the assay.

Description	H ³ -SAM (0.55μCi)	Suv39h1 (0.2μg)	H3K9 (5μg)	H3K9me (5μg)	H3K9Ac (5μg)	Chaetocin	DMSO	CPM (Trial1)	CPM (Trial2)
H3K9	✓	✓	✓	✗	✗	✗	✗	17763	16912
H3K9me	✓	✓	✗	✓	✗	✗	✗	11613	10227
H3K9Ac	✓	✓	✗	✗	✓	✗	✗	240	273
No Suv39h1	✓	✗	✓	✗	✗	✗	✗	402	416
No H3- SAM	✗	✓	✓	✗	✗	✗	✗	31	15
No peptide	✓	✓	✗	✗	✗	✗	✗	185	148
Only H3- SAM	✓	✗	✗	✗	✗	✗	✗	405	308
Chaetocin	✓	✓	✓	✗	✗	✓	✗	---	9426
DMSO	✓	✓	✓	✗	✗	✗	✓	---	6167
Blank	✗	✗	✗	✗	✗	✗	✗	29	37

Table 3.3 Experimental plan and results of validation of the *in vitro* histone methyltransferase assay with necessary control

H3K9 and H3K9me were used as substrates in the positive control reaction and they show higher counts per minute upon scintillation counting. The CPM reflects the H³ labeled methyl group being incorporated onto the K9 residue of the histone H3 peptide. H3K9Ac peptide was used as a negative control because the acetylated K9 residue cannot be methylated and therefore shows much lower counts per minute as expected. All other negative controls tested also show very low counts per minute as expected. Chaetocin and DMSO (Dimethyl Sulfoxide) were not tested in trial 1.

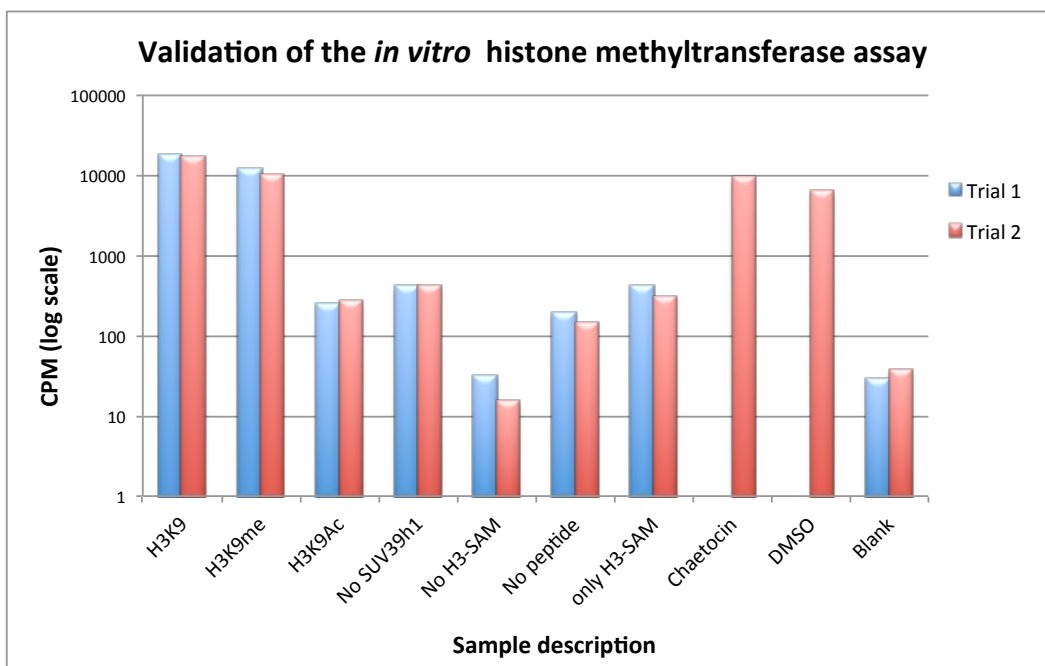


Figure 3.23: Results of the validation of *in vitro* histone methyltransferase assay

Two trials were performed with the positive and negative controls as listed in Table 3.3. Blue bars represent trial 1 and red bars show trial 2. Trial 1 and was done without chaetocin and DMSO samples. Noted that the Y-axis is plotted in logarithmic scale because of the vast difference in CPM readings of different samples. The positive controls (i.e. H3K9 and H3K9me) show at least 40-fold increased activity compared to the negative control (i.e. H3K9Ac).

3.7.3 Testing of the small molecule inhibitors by radioactive *in vitro* assay

The small molecules obtained from virtual screening (Table 3.2) were divided into different batches and tested for their methyltransferase activity based in the order in which they were received. When possible, the compounds were categorized into different batches based on the database to which they belonged. All compounds were tested in 1 μM and 5 μM concentrations. Two trials at 5 μM and 1 μM of each batch of compounds were performed. The assay results are tabulated and followed by a corresponding chart showing the percentage inhibition of methyltransferase activity of each sample. In each batch, unmethylated H3K9 residue was used as substrate for the positive control, acetylated H3K9 residue was used as a substrate for the negative control and DMSO is used as a control for the inhibitors as they are all dissolved in DMSO. The methyltransferase activity of each sample is shown in counts per minute (CPM) and the percentage inhibition of methyltransferase activity is calculated by comparing it to the CPM of unmethylated H3K9 residue (positive control), which represents maximum methyltransferase activity (i.e. zero inhibition).

3.7.3.1 . *In vitro* methyltransferase assay results for the NCI diversity set 2 compounds

The compounds from the NCI diversity set 2 were tested at 1 μ M and 5 μ M concentrations. The compounds were divided into two batches and two trials were done with each batch of compounds at 5 μ M concentration. (Table 3.4 and Table 3.5, Figure 3.24 and Figure 3.25). All the compounds were tested at 1 μ M concentration in a single batch in order to reduce the variability in the assay conditions (Table 3.6, Figure 3.26). The membranes were not allowed to dry before washing them during the second trial at 1 μ M concentration.

NCI id	Sample description	Trial 1 CPM	Trial 2 CPM
	H3K9	8178.4	6770.8
	H3K9Ac	1030.5	336.4
	DMSO	6623.23	5894.32
Chaetocin	1	5088	7488.43
NSC 5157	2	5508.65	7078.64
NSC 5856	3	7353.57	8410.85
NSC 6844	4	6915.17	7955.06
NSC 23123	5	6221.21	7828.07
NSC 37553	6	5342.63	8923.58
NSC 50650	7	4944	7693.38
NSC 61610	8	5376.32	9288.11
NSC 71881	9	4503.5	8206.11
NSC 80731	10	6000	6718.89
NSC 80735	11	4482.5	6175.49
NSC 84100	12	5016.5	8582.54
NSC 87838	13	5223.08	9082.96
NSC 109451	14	5738.29	8510.66
NSC 116702	15	6225.45	9964.49
NSC 132232	16	7403.5	6517.24
NSC 156516	17	5056.5	8761.79

Table 3.4: Results of the inhibition assay for NCI Diversity set 2 compounds (Batch 1) at 5 μ M inhibitor concentration

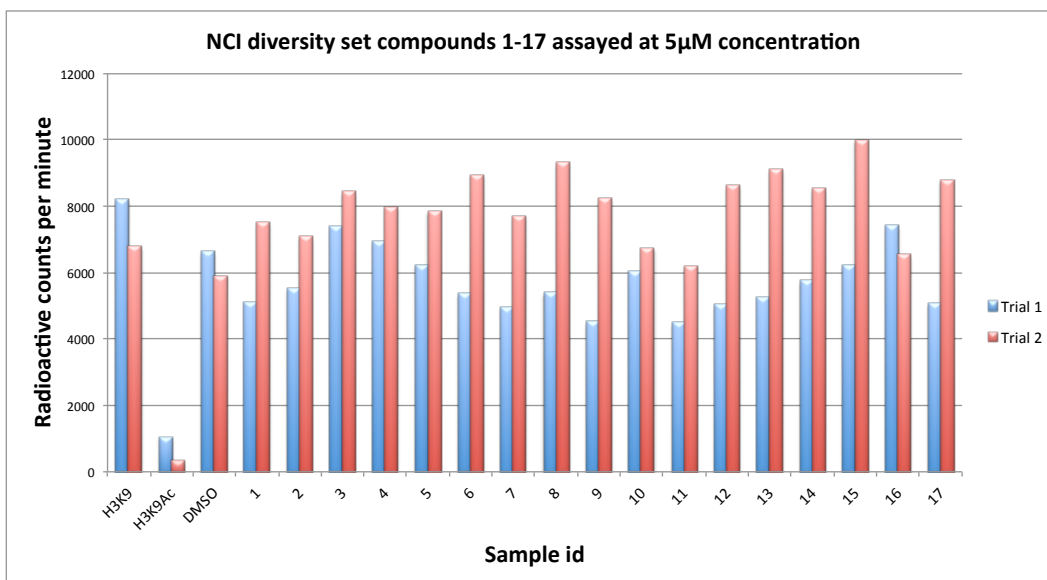


Figure 3.24: Radioactive counts per minute (CPM) of NCI Diversity set compounds (Batch 1) assayed at 5 μ M concentration

The X-axis corresponds to the sample id (Table 3.4) and the Y-axis corresponds to the CPM of radiation from each sample. H3K9 is the positive control and the sample labeled H3K9Ac (acetylated H3K9) is the negative control. All samples show varied levels of inhibition of methyltransferase activity in the first trial (blue bars), where as there is no inhibition by any of the compounds in the second trial (red bars).

NCI id	Sample description	Trial 1 CPM	Trial 2 CPM
	H3K9	5130.2	6734.6
	H3K9Ac	1150.9	396.1
	DMSO	4583.9	8324.12
Chaetocin	18	8061.73	6485.83
NSC 156565	19	7621.74	6205
NSC 203837	20	6259.04	5670.04
NSC 260594	21	6732.05	6630.35
NSC 292140	22	5871.55	6656.86
NSC 298892	23	4231.54	4356.04
NSC 305329	24	5639.96	5449.76
NSC 310324	25	4937.56	3665.24
NSC 319990	26	5031.77	5367.77
NSC 326182	27	6086.29	4909.57
NSC 332670	28	6536.5	3881.36
NSC 335048	29	4912.18	4777.48
NSC 377852	30	5423.29	5439.59
NSC 378719	31	5603.5	5377.3
NSC 400976	32	5870.2	6404.73
NSC 407628	33	5017.3	5278.21
NSC 670283	34	2328.55	5607.72

Table 3.5: Results of the inhibition assay for NCI Diversity set 2 compounds (Batch 2) at 5 μ M inhibitor concentration

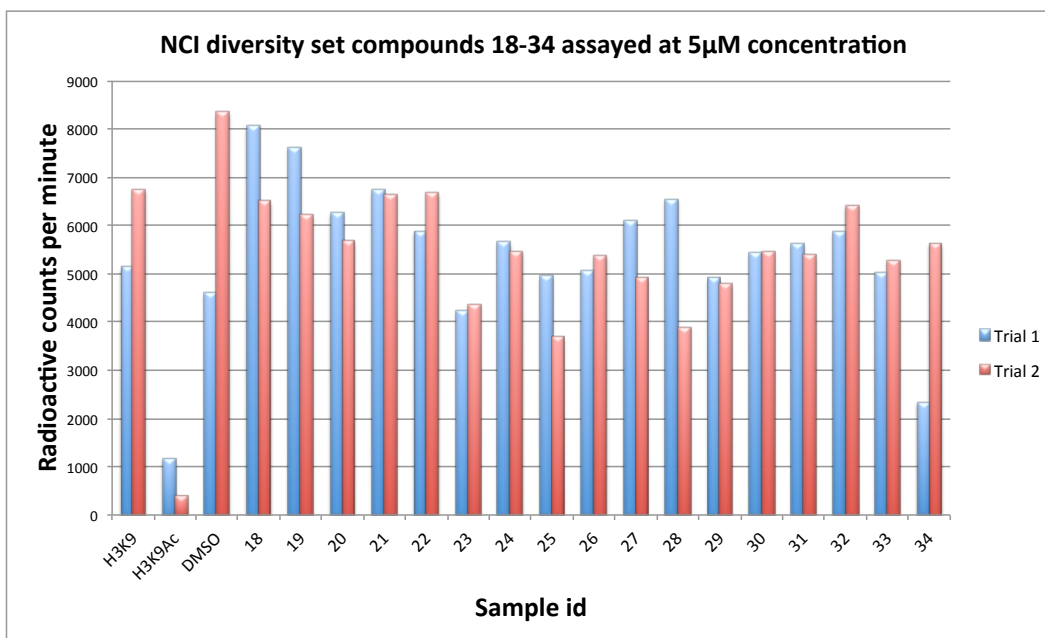


Figure 3.25: Radioactive counts per minute (CPM) of NCI Diversity set compounds (Batch 2) assayed at 5µM concentration

The X-axis corresponds to the sample id (Table 3.5) and the Y-axis corresponds to the CPM of radiation from each sample. H3K9 is the positive control and the sample labeled H3K9Ac (acetylated H3K9) is the negative control. Blue bars represent the results from trial 1 and the red bars represent the results from trial 2.

NCI id	Sample description	Trial 1 CPM	Trial 2 CPM
	Positive control	9117.4	21443.41
	H3K9Ac	482.8	412.80
	Heat Killed Suv39h1	528.9	978.20
	DMSO	9313.14	5260.61
Chaetocin	1	6901.13	9423.02
NSC 5157	2	8412.25	6318.42
NSC 5856	3	8455.06	8650.83
NSC 6844	4	8072.17	5451.82
NSC 23123	5	8101.58	4034.42
NSC 37553	6	10713.91	5958.83
NSC 50650	7	5114.46	5837.64
NSC 61610	8	7210.29	6250.44
NSC 71881	9	3390.45	4258.63
NSC 80731	10	4705.47	5801.25
NSC 80735	11	3560.86	4884.04
NSC 84100	12	7489.63	4299.04
NSC 87838	13	7389.44	4857.85
NSC 109451	14	5759.91	5248.85
NSC 116702	15	9578.1	3755.04
NSC 132232	16	9408.61	5074.86
NSC 156516	17	10467.74	4790.46
NSC 156565	18	13072.31	5364.07
NSC 203837	19	12697.72	4846.46
NSC 260594	20	11948.61	6038.68
NSC 292140	21	10697.99	4082.46
NSC 298892	22	8873.95	3683.85
NSC 305329	23	8699.66	2688.24
NSC 310324	24	9029.98	8010.53
NSC 319990	25	7878.15	7816.13

NSC 326182	26	7956.56	7213.92
NSC 332670	27	16815.77	7247.53
NSC 335048	28	15316.94	8841.36
NSC 377852	29	5298.49	6236.52
NSC 378719	30	7262.47	5584.11
NSC 400976	31	7428.99	7902.96
NSC 407628	32	6874.7	5704.12
NSC 670283	33	6346.4	6401.14

Table 3.6: Results of the inhibition assay for all NCI Diversity set 2 compounds tested at 1 μ M concentration

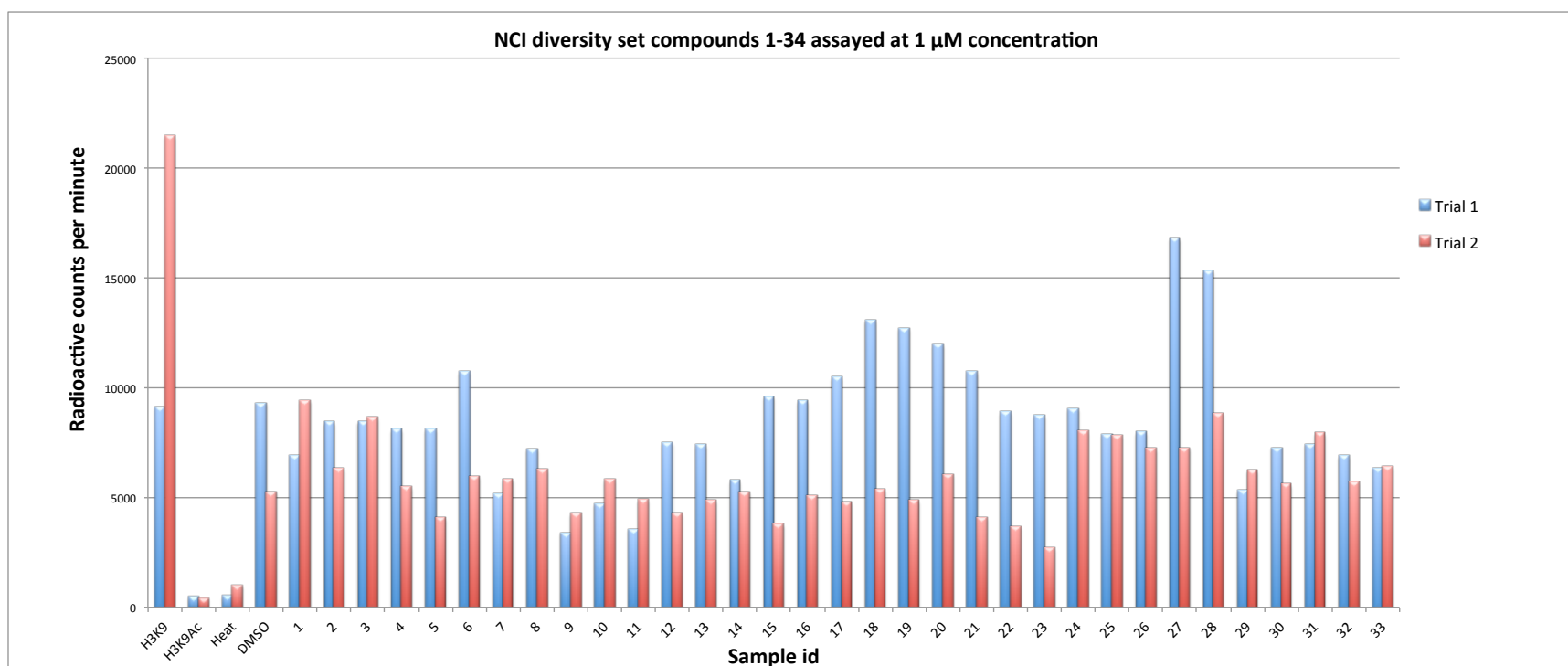


Figure 3.26: Radioactive counts per minute (CPM) of NCI Diversity set compounds assayed at 1 μ M concentration

The X-axis corresponds to the sample id (Table 3.6) and the Y-axis corresponds to the CPM of radiation from each sample. H3K9 is the positive control and the sample labeled H3K9Ac (acetylated H3K9) is the negative control. Blue bars represent the results from trial 1 and the red bars represent the results from trial 2.

3.7.3.2 . *In vitro* methyltransferase assay results for the DrugBank, ZINC and ZDD compounds

The DrugBank, ZINC and ZDD compounds were tested at 1 μM and 5 μM concentrations. The compounds were divided into two batches in the order in which they were received and tested at 5 μM concentration, the results of are listed in Table 3.7 and Table 3.8 and their corresponding plots showing percentage inhibition of activity are shown in Figure 3.27 and Figure 3.28. The assay was performed at 1 μM in a single batch in order to reduce the variability in assay conditions; the results of which are listed in Table 3.9 and the corresponding plot is shown in

Figure 3.29. The membranes were not allowed to dry before washing them during the second trial at 1 μM concentration. Samples 18 (Astemizole) and 20 (Dutasteride) have shown consistent inhibition of activity in both the trials at 5 μm and at 1 μM .

Compound name	Sample description	Trial 1 CPM	Trial 2 CPM
	H3K9	4957.8	6433.6
	H3K9Ac	621.2	714.1
	DMSO	2638.5	6548.4
Chaetocin	1	10427.84	5286.8
Glyburide	2	8035	5617.7
Sorafenib	3	8095	6486.1
Resperidone	4	4386	6051.8
Sulfasalazine	5	8550	5242.2
Glimepiride	6	6735	5089.3
Tetrahydrofolic acid	7	7810.8	5587.06
Piroxicam	8	7330.8	4659.3
Domperidone	9	9245.6	5418.3
Estrone	10	7339	5662.6
Imatinib	11	7816	6608.5

Table 3.7: Results of the inhibition assay for DrugBank, ZINC, ZDD compounds at 5 μ M inhibitor concentration (Batch 1)

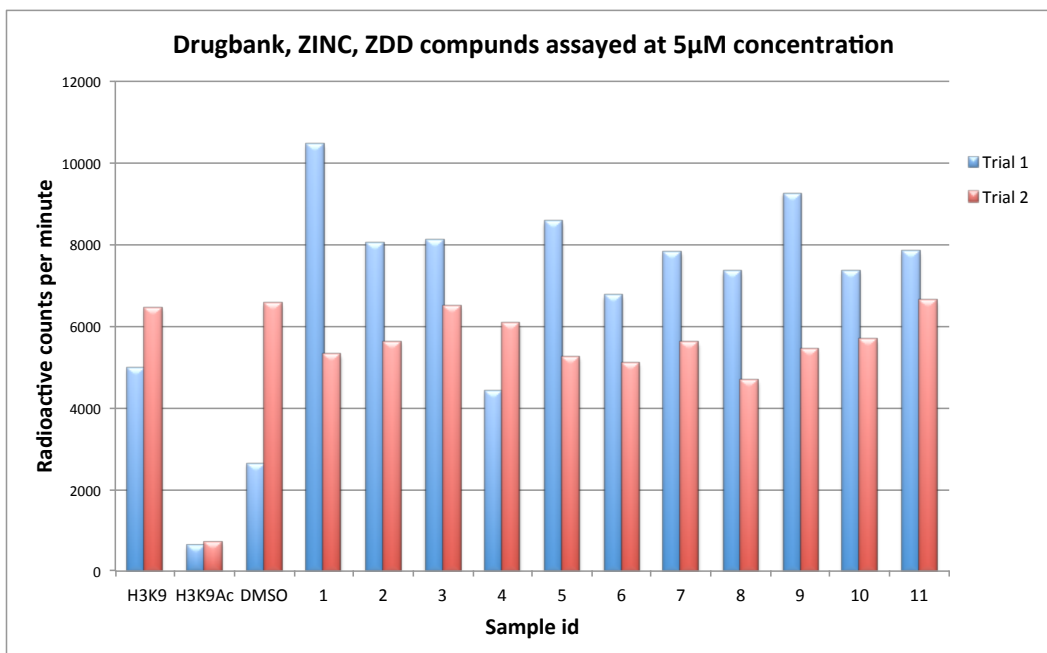


Figure 3.27: Radioactive counts per minute (CPM) of DrugBank, ZINC, ZDD compounds assayed at 5 µM inhibitor concentration (Batch 1).

The X-axis corresponds to the sample id (Table 3.7) and the Y-axis corresponds to the CPM of radiation from each sample. H3K9 is the positive control and the sample labeled H3K9Ac (acetylated H3K9) is the negative control. Blue bars represent the results from trial 1 and the red bars represent the results from trial 2.

Compound name	Sample description	Trial 1 CPM	Trial 2 CPM
	H3K9	10746.4	6504.5
	H3K9Ac	1131	990.7
	DMSO	8362	4914.1
Nebivolol HCl	12	8840.4	7118.4
Tamibarotene	13	8072.8	6560.3
Paroxetine malate	14	8703.9	6314.7
Atovaquone	15	7752.2	8387.8
Tadalafil	16	8631.6	8409.7
Telmisartan	17	8376.4	7084.8
Astemizole	18	7596.4	6005.5
Zafirlukast	19	11652.8	5429.8
Dutasteride	20	5598.6	5811.5

Table 3.8: Results of the inhibition assay for DrugBank, ZINC, ZDD compounds at 5 μ M inhibitor concentration (Batch 2)

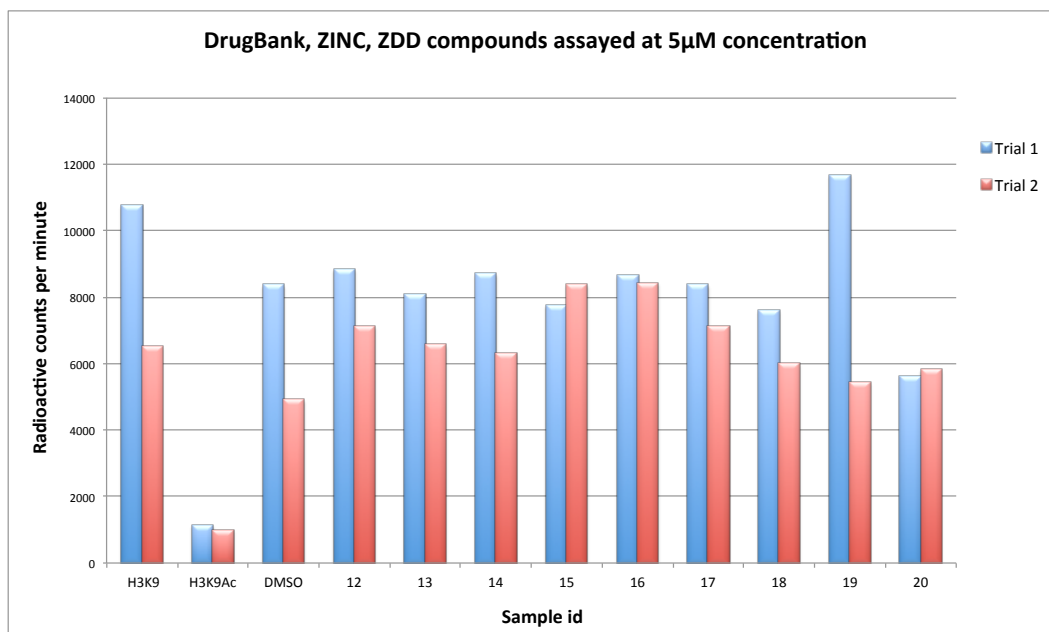


Figure 3.28: Radioactive counts per minute (CPM) of DrugBank, ZINC, ZDD compounds assayed at 5 µM inhibitor concentration (Batch 2).

The X-axis corresponds to the sample id (Table 3.8) and the Y-axis corresponds to the CPM of radiation from each sample. H3K9 is the positive control and the sample labeled H3K9Ac (acetylated H3K9) is the negative control. Blue bars represent the results from trial 1 and the red bars represent the results from trial 2.

Sample number	Sample description	CPM (Trial 1)	CPM (Trial 2)
	H3K9	2244.4	866.20
	H3K9Ac	584.8	51.40
	DMSO	712.8	794.20
1	Chaetocin	1976.6	4779.01
2	Glyburide	1459.2	1197.80
3	Sorafenib	2849.01	1318.20
4	Resperidone	1843.21	1191.41
5	Sulfasalazine	1890.81	3566.82
6	Glimepiride	1819.61	1503.61
7	Tetrahydrofolic acid	2124.21	1580.41
8	Piroxicam	1776.21	1343.61
9	Domperidone	1641.01	50.60
10	Estrone	713.61	1271.81
11	Imatinib	2541.82	835.21
12	Nebivolol HCl	1807.22	2074.42
13	Tamibarotene	824.21	1085.21
14	Paroxetine malate	1713.02	1213.41
15	atovaquone	1682.02	1452.62
16	Tadalafil	1330.21	1027.01
17	Telmisartan	1293.41	1390.02
18	Astemizole	1452.02	1758.42
19	Zafirlukast	1476.82	1275.82
20	Dutasteride	1246.42	1630.22

Table 3.9: Results of the inhibition assay for DrugBank, ZINC, ZDD compounds at 1 μ M inhibitor concentration.

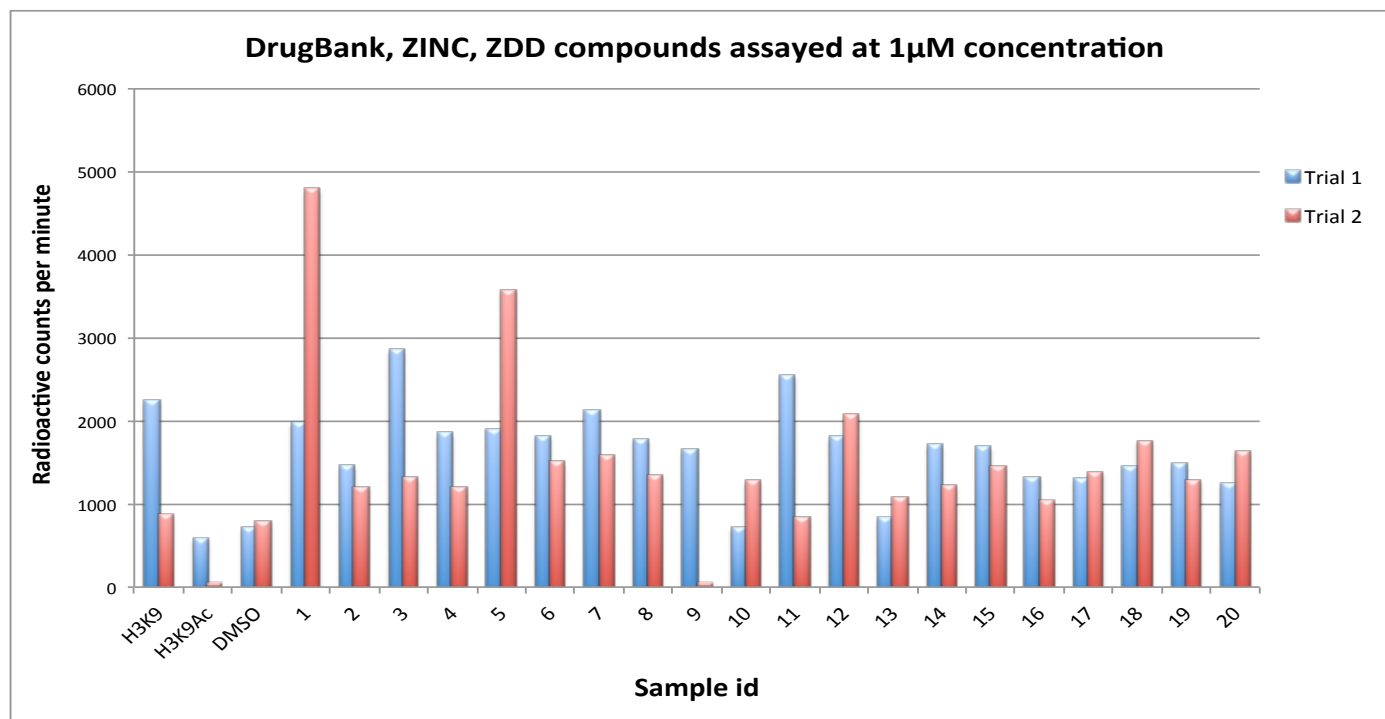


Figure 3.29: Radioactive counts per minute (CPM) of DrugBank, ZINC, ZDD compounds assayed at 1 μ M inhibitor concentration.

The X-axis corresponds to the sample id (Table 3.9) and the Y-axis corresponds to the CPM of radiation from each sample. H3K9 is the positive control and the sample labeled H3K9Ac (acetylated H3K9) is the negative control. Blue bars represent the results from trial 1 and the red bars represent the results from trial 2

Chapter IV – Discussion and Future Directions

4. Chapter 4 – Discussion and Future Directions

Even though several histone methyltransferases have been implicated in different types of cancer, and given the immense potential for this group of enzymes to be used as targets for therapy, this field has remained relatively unexplored. In this thesis, work was done to discover novel small molecule inhibitors of the H3K9 trimethylase, Suv39h1 by homology modeling and virtual screening.

4.1. Expression of human Suv39h1

The task of human Suv39h1 protein expression was performed in order to synthesize active functional human Suv39h1, which could be used for testing against the predicted small molecule inhibitors. As described previously, bacterial and mammalian expression strategies were used. Despite employing several expression vectors and different parameters for the bacterial expression system, human Suv39h1 was always expressed in the insoluble fraction. Protein expression in the insoluble fraction is mainly attributed to high levels of protein expression and formation of inclusion bodies that leads to misfolded proteins. Misfolded proteins are usually functionally inactive and hence it was crucial to express the protein in the soluble fraction. In order to counter the issue of high levels of protein expression, very low amounts of IPTG (0.5 mM and even 0.1 mM in some cases) were used. Reducing the IPTG did result in low protein expression, however it was still expressed in the insoluble fraction. Different fusion tags were used (MBP, GST, N110-His tag) which apart from helping downstream in the purification of the expressed protein have also been shown to

increase the solubility of expressed proteins (Dyson, Shadbolt et al. 2004). The lactose induction system, which is a much milder expression system compared to IPTG based protein induction was also employed, which also resulted in insoluble proteins (Studier 2005).

Denaturation and refolding of the misfolded protein by dialysis in order to obtain functional human Suv39h1 was also tried. However, all the attempts were unsuccessful as the protein precipitated and degraded during dialysis despite the addition of protease inhibitors.

A mammalian protein expression system was employed in order to overcome the issues encountered in the bacterial expression system by using pCEMM-NTAP-h*SUV39h1*-855/856 and HEK293T cells as described previously. Western blot revealed expression of the human Suv39h1 protein in the soluble fraction as shown in Figure 3.1. Later on, the hyperactive and catalytically dead human Suv39h1 were also expressed in a mammalian expression system. However, the protein had to be purified and before being used for the methyltransferase assay, because the cell lysate that had overexpressed TAP-tagged human Suv39h1 also had other histone methyltransferases that could potentially interfere in the methyltransferase reaction. TAP tagged human Suv39h1 was isolated by incubation of the cell lysate with IgG beads. Upon analyzing the purified fraction by western blot no bands were seen, revealing that the protein had not been purified from the lysate. This could have been due to insufficient protein in the lysate or the protein could have degraded during the purification process.

Despite all these efforts into synthesizing active functional human Suv39h1, in the end it was not feasible in terms of time. Therefore commercially available purified human Suv39h1 was obtained from Sigma and was used in all the *in vitro* assays.

4.2. Homology modeling of human Suv39h1 and virtual screening

In an ideal situation, virtual screening is performed with a high-resolution structure of a ligand-bound protein. However, the absence of a three-dimensional structure of human Suv39h1 prompted the construction of its homology model. The homology model was built on the principle that homologous proteins with similar sequence have similar structures. Since the crystal structure of Suv39h2, which is homologous to Suv39h1 is available, it was used as the template as described earlier. The homology model was optimized by MD simulation and validated by a set of tools available at SWISS-MODEL that are routinely used in the field of homology modeling thus giving us a model that is closest to the native structure of human Suv39h1 as possible.

Over the last few years virtual screening has emerged as a powerful and viable tool for the identification of novel and diverse lead compounds that can be further derivatized to form effective small molecule inhibitors or potential drug molecules. With the rapid improvements in homology modeling and docking algorithms, several successful virtual screening studies based on homology models of the proteins have been reported in recent years (Kellenberger, Springael et al. 2007; Noeske, Jirgensons et al. 2007; Schlegel, Laggner et al. 2007; Heinke, Spannhoff et al. 2009).

The amino acids of the human Suv39h1 model were not allowed to move during the docking of the small molecules i.e. rigid docking was done in this experiment. Another way of doing this would be to allow some of the residues to be flexible, especially the ones in the binding site being used for virtual screening. Protein flexibility is of relevance if there is a drastic change in the conformation upon ligand binding and is usually an issue when the homology model is of very low quality, or the protein of interest is an intrinsically unstable protein where the structure is not defined until ligand binding (Cozzini, Kellogg et al. 2008). Since the homology model constructed for human Suv39h1 was optimized by MD simulations, and the final structure chosen is a representative of the largest cluster of structures, it can be believed that all the residues have attained their global energy minimum and the model is highly reliable. This exempts the use of flexible residues, which is more time consuming and requires more computational power. Also, most successful virtual screening studies to date have been accomplished without considering protein flexibility. It should be noted that in these cases there is a very slight change in the conformation upon ligand binding (Bolstad and Anderson 2008).

One of the major problems with discovering new small molecules that target histone methyltransferases is selectivity. Since almost all histone methyltransferases use SAM as the methyl donor, using the binding site of SAM for screening new molecules has an inherent problem, in that it leads to the discovery of non-specific inhibitors that target all histone methyltransferases that use SAM. This can be overcome by using other potential binding pockets on the

surface of the protein. In order to do this, any knowledge of the structure and function of the protein (i.e. protein-protein, protein-DNA interaction sites) is of immense help. In this project, the binding sites of chaetocin and the N-terminal peptide of histone H3 were determined and used for virtual screening. The binding site of chaetocin is a novel site that was determined by blind docking, and the small molecules shown to bind to this site with high affinity should be specific to human Suv39h1. The binding site of the N-terminal peptide of histone H3 was determined by comparison with other histone methyltransferase with known binding sites. Therefore, the caveat for this binding site is that the small molecules shown to have high affinity to this site might also inhibit other H3K9 binding histone methyltransferases like G9a, GLP and EHMT1; although, this should still be better than the ones targeting SAM binding site.

4.3. Development of an *in vitro* histone methyltransferase assay

Development of an *in vitro* histone methyltransferase assay was a significant part of this study. The initial ELISA based fluorescent histone methyltransferase assay was chosen because of its ease of use and safety compared to the radioactive assay. The radioactive assay has inherent disadvantages compared to the fluorescent assay, such as exposure of the personnel to radioactivity, radioactive contamination of the equipment and disposal of the radioactive waste; however, it is also highly specific, sensitive and bypasses the use of antibodies to detect change in methylation levels, which could be one of the major reasons for high background noise in the fluorescent assay. The fact that the negative and positive

controls were indiscernible and the trials were inconsistent made us look for an alternative method and go ahead with the radioactive assay.

The radioactive assay was first carried out with peptides that were not biotinylated under similar reaction conditions as mentioned for the biotinylated peptides. The major difference was during the blotting and washing steps. After the methyltransferase reaction was carried out the reaction mixture containing the non-biotinylated peptides were blotted onto P81 phosphocellulose membranes. This method has been reported several times in the literature although with scant details (Strahl, Ohba et al. 1999; Greiner, Bonaldi et al. 2005). The P81 membrane carries a net negative charge and the peptides are supposed to bind onto the surface of the membrane by virtue of their positive charge. However, after several trials, no consistent results were obtained during the optimization of the assay with different controls. This could be due to non-specific binding of the peptides onto the membrane or improper washing of the blotted membranes. The membranes were washed by a vacuum assisted procedure as well as free floating in a large volume of wash buffer (>30ml/blot) and washed at least 4 times. The blots were then dried and counted by scintillation counting as described earlier.

Radioactive assays produce a lot of background noise and proper washing of the membrane is crucial in order to get rid of the unreacted SAM bound non-specifically to the membrane. In order to see if the washing step was leading to the random variation in the values obtained upon scintillation counting, that step was bypassed and the reaction mixture was run on an 18% polyacrylamide gel (27:1 acrylamide:bis), fixed with 5% glutaraldehyde for 30 minutes and stained

with commassie blue in order to visualize the peptides. The plan was to cut the peptide bands out of the gel and dissolve the bands in the scintillation fluid and count for radioactivity. However, this plan failed because it was extremely difficult to visualize the peptides due to their small size (~2.5 kDa) and therefore the peptide bands could not be isolated from the gel. Alternatively, the gel was exposed to the PHOSPHORIMAGER storage screen for 30 hours. When the screen was imaged after 30 hours, no bands were seen on the screen. The lack of signal could be due to no methyltransferase reaction or the absence of peptide from the gel or longer exposure times are required in order to register a better signal. Therefore this method too seemed unviable because a reasonably high throughput mechanism was required in order to test the compounds.

Finally, the biotinylated peptides and streptavidin membrane that are highly specific to each other were chosen and the assays were performed which resulted in very consistent results during the two optimization trials that were carried out using several controls (Table 3.3).

4.4. Validation of the small molecule inhibitors by the *in vitro* histone methyltransferase assay

Small molecule inhibitors are used at very low concentrations (usually in the range of nM to low μ M) in cell based assays in order to avoid any non-specific binding of the compounds to targets other than the intended target. Using lower concentration also helps in preventing or limiting any potential cytotoxic effects of the small molecules. In accordance with this principle, the small molecules were tested for their inhibitory efficiency at 1 μ M and 5 μ M concentrations in this

project. Two trials were done at 5 μ M concentration for the all compounds, where as only one trial was done at 1 μ M concentration because of limited resources and time. Although the assay was optimized and all the controls were confirmed twice, there is a large degree of variation from one trial to the next when the compounds were tested with this assay. This variation can be seen in the trials done with 5 μ M inhibitor concentrations for all the batches of compounds. This variation can be attributed to the error in the positive control as the percentage activity inhibition is calculated relative to it. Therefore, if the radioactive CPM of the positive control is lower than the samples, as is the case in trials done at 5 μ M concentrations, it seems as if the compounds are increasing the activity of the enzyme rather than inhibiting it. This discrepancy can be resolved by increasing the number of trials per batch. It is important to note that it is quite possible that some of the compounds might indeed be increasing the activity, as the AutoDock only predicts the binding affinity of the compounds and not their inhibition potential.

4.5.Future directions

The next step is to repeat the experiments with the modified protocol i.e. by keeping the membranes hydrated before washing them in order to get more consistent and reliable results. Then dose response assays can be performed to find the IC_{50} values of the top compounds that show inhibition at 1 μ M and 5 μ M. Cell based assays can be done in order to check for the inhibition of H3K9 trimethylation by human Suv39h1 *in vivo* by using antibodies that specifically

target trimethylated H3K9. Studies to analyze the cytotoxicity of these compounds can be done using cell death assays. It will be exciting to check if these compounds can reverse the aberrant methylation of promoters of tumor suppressor genes targeted by Suv39h1.

Histone methylation, which was until recently thought to be an irreversible process, was shown to be dynamic process that is regulated by the interplay between histone methyltransferases and demethylases. The association of these enzymes in different kinds of cancer and other diseases establishes them as therapeutic targets. In recent years more and more information about the epigenetic processes and their association with diseases is being discovered and most importantly there is still a lot of which is yet to be discovered. Currently, the field of epigenetic therapy is still in its infancy. At this point only a few drugs have been approved for therapy and effort is being put into discovering new drugs targeting other epigenetic targets. A dynamic development of this field is imminent and the next generation of cancer therapeutic drugs will definitely routinely include epigenetic drugs either by themselves or in combination with conventional cancer therapy.

Bibliography

- Aagaard, L., G. Laible, et al. (1999). "Functional mammalian homologues of the *Drosophila* PEV-modifier Su(var)3-9 encode centromere-associated proteins which complex with the heterochromatin component M31." EMBO J **18**(7): 1923-1938.
- Aagaard, L., M. Schmid, et al. (2000). "Mitotic phosphorylation of SUV39H1, a novel component of active centromeres, coincides with transient accumulation at mammalian centromeres." J Cell Sci **113** (Pt 5): 817-829.
- Al-Lazikani, B., J. Jung, et al. (2001). "Protein structure prediction." Current opinion in chemical biology **5**(1): 51-56.
- Albert, M. and K. Helin (2010). "Histone methyltransferases in cancer." Semin Cell Dev Biol **21**(2): 209-220.
- Altschul, S. F., W. Gish, et al. (1990). "Basic local alignment search tool." J Mol Biol **215**(3): 403-410.
- Altschul, S. F., T. L. Madden, et al. (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res **25**(17): 3389-3402.
- Arnold, K., L. Bordoli, et al. (2006). "The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling." Bioinformatics **22**(2): 195-201.
- Backman, T. W., Y. Cao, et al. (2011). "ChemMine tools: an online service for analyzing and clustering small molecules." Nucleic Acids Res **39**(Web Server issue): W486-491.
- Baker, D. and A. Sali (2001). "Protein structure prediction and structural genomics." Science **294**(5540): 93-96.
- Bannister, A. J., P. Zegerman, et al. (2001). "Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain." Nature **410**(6824): 120-124.
- Bedford, M. T. and S. Richard (2005). "Arginine methylation an emerging regulator of protein function." Mol Cell **18**(3): 263-272.
- Benkert, P., M. Biasini, et al. (2011). "Toward the estimation of the absolute quality of individual protein structure models." Bioinformatics **27**(3): 343-350.
- Benkert, P., S. C. Tosatto, et al. (2008). "QMEAN: A comprehensive scoring function for model quality assessment." Proteins **71**(1): 261-277.
- Bernard, P. and R. Allshire (2002). "Centromeres become unstuck without heterochromatin." Trends Cell Biol **12**(9): 419-424.

- Bolstad, E. S. and A. C. Anderson (2008). "In pursuit of virtual lead optimization: the role of the receptor structure and ensembles in accurate docking." Proteins **73**(3): 566-580.
- Bonneau, R. and D. Baker (2001). "Ab initio protein structure prediction: progress and prospects." Annual review of biophysics and biomolecular structure **30**: 173-189.
- Bordoli, L., F. Kiefer, et al. (2009). "Protein structure homology modeling using SWISS-MODEL workspace." Nature protocols **4**(1): 1-13.
- Burckstummer, T., K. L. Bennett, et al. (2006). "An efficient tandem affinity purification procedure for interaction proteomics in mammalian cells." Nat Methods **3**(12): 1013-1019.
- Byvoet, P., G. R. Shepherd, et al. (1972). "The distribution and turnover of labeled methyl groups in histone fractions of cultured mammalian cells." Arch Biochem Biophys **148**(2): 558-567.
- C, B. R., J. Subramanian, et al. (2009). "Managing protein flexibility in docking and its applications." Drug Discov Today **14**(7-8): 394-400.
- Carew, J. S., F. J. Giles, et al. (2008). "Histone deacetylase inhibitors: mechanisms of cell death and promise in combination cancer therapy." Cancer Lett **269**(1): 7-17.
- Chakraborty, S., K. K. Sinha, et al. (2003). "SUV39H1 interacts with AML1 and abrogates AML1 transactivity. AML1 is methylated in vivo." Oncogene **22**(34): 5229-5237.
- Cheng, D., J. Cote, et al. (2007). "The arginine methyltransferase CARM1 regulates the coupling of transcription and mRNA processing." Mol Cell **25**(1): 71-83.
- Cheng, X., R. E. Collins, et al. (2005). "Structural and sequence motifs of protein (histone) methylation enzymes." Annu Rev Biophys Biomol Struct **34**: 267-294.
- Chenna, R., H. Sugawara, et al. (2003). "Multiple sequence alignment with the Clustal series of programs." Nucleic Acids Res **31**(13): 3497-3500.
- Cherblanc, F. L., K. L. Chapman, et al. (2013). "Chaetocin is a nonspecific inhibitor of histone lysine methyltransferases." Nat Chem Biol **9**(3): 136-137.
- Chothia, C. and A. M. Lesk (1986). "The relation between the divergence of sequence and structure in proteins." EMBO J **5**(4): 823-826.
- Claussen, H., C. Buning, et al. (2001). "FlexE: efficient molecular docking considering protein structure variations." J Mol Biol **308**(2): 377-395.
- Cornell, W. D., P. Cieplak, et al. (1995). "A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules." Journal of the American Chemical Society **117**(19): 5179-5197.

- Cozzini, P., G. E. Kellogg, et al. (2008). "Target flexibility: an emerging consideration in drug discovery and design." J Med Chem **51**(20): 6237-6255.
- Duvic, M. and J. Vu (2007). "Vorinostat: a new oral histone deacetylase inhibitor approved for cutaneous T-cell lymphoma." Expert Opin Investig Drugs **16**(7): 1111-1120.
- Dyson, M. R., S. P. Shadbolt, et al. (2004). "Production of soluble mammalian proteins in Escherichia coli: identification of protein features that correlate with successful expression." BMC Biotechnol **4**: 32.
- Edgar, R. C. and S. Batzoglou (2006). "Multiple sequence alignment." Current opinion in structural biology **16**(3): 368-373.
- Ewing, T. J., S. Makino, et al. (2001). "DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases." J Comput Aided Mol Des **15**(5): 411-428.
- Fenaux, P., G. J. Mufti, et al. (2009). "Efficacy of azacitidine compared with that of conventional care regimens in the treatment of higher-risk myelodysplastic syndromes: a randomised, open-label, phase III study." Lancet Oncol **10**(3): 223-232.
- Fernandez-Fuentes, N., B. Oliva, et al. (2006). "A supersecondary structure library and search algorithm for modeling loops in protein structures." Nucleic acids research **34**(7): 2085-2097.
- Fernandez-Fuentes, N., E. Querol, et al. (2005). "Prediction of the conformation and geometry of loops in globular proteins: testing ArchDB, a structural classification of loops." Proteins **60**(4): 746-757.
- Fiser, A., R. K. Do, et al. (2000). "Modeling of loops in protein structures." Protein science : a publication of the Protein Society **9**(9): 1753-1773.
- Friesner, R. A., J. L. Banks, et al. (2004). "Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy." J Med Chem **47**(7): 1739-1749.
- Fritsch, L., P. Robin, et al. (2010). "A subset of the histone H3 lysine 9 methyltransferases Suv39h1, G9a, GLP, and SETDB1 participate in a multimeric complex." Mol Cell **37**(1): 46-56.
- Gasteiger, J. and M. Marsili (1978). "A new model for calculating atomic charges in molecules." Tetrahedron Letters **19**(34): 3181-3184.
- Good, A. C., S. R. Krystek, et al. (2000). "High-throughput and virtual screening: core lead discovery technologies move towards integration." Drug Discov Today **5**(12 Suppl 1): 61-69.
- Goodsell, D. S., G. M. Morris, et al. (1996). "Automated docking of flexible ligands: applications of AutoDock." J Mol Recognit **9**(1): 1-5.

- Greiner, D., T. Bonaldi, et al. (2005). "Identification of a specific inhibitor of the histone methyltransferase SU(VAR)3-9." Nat Chem Biol **1**(3): 143-145.
- Halgren, T. A., R. B. Murphy, et al. (2004). "Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening." J Med Chem **47**(7): 1750-1759.
- Hartwell, L. (1992). "Defects in a cell cycle checkpoint may be responsible for the genomic instability of cancer cells." Cell **71**(4): 543-546.
- Heard, E., C. Rougeulle, et al. (2001). "Methylation of histone H3 at Lys-9 is an early mark on the X chromosome during X inactivation." Cell **107**(6): 727-738.
- Heinke, R., A. Spannhoff, et al. (2009). "Virtual screening and biological characterization of novel histone arginine methyltransferase PRMT1 inhibitors." ChemMedChem **4**(1): 69-77.
- Heit, R., J. B. Rattner, et al. (2009). "G2 histone methylation is required for the proper segregation of chromosomes." J Cell Sci **122**(Pt 16): 2957-2968.
- Hetyenyi, C. and D. van der Spoel (2002). "Efficient docking of peptides to proteins without prior knowledge of the binding site." Protein science : a publication of the Protein Society **11**(7): 1729-1737.
- Higashimoto, K., P. Kuhn, et al. (2007). "Phosphorylation-mediated inactivation of coactivator-associated arginine methyltransferase 1." Proc Natl Acad Sci U S A **104**(30): 12318-12323.
- Hoofst, R. W., G. Vriend, et al. (1996). "Errors in protein structures." Nature **381**(6580): 272.
- Huang, J. and S. L. Berger (2008). "The emerging field of dynamic lysine methylation of non-histone proteins." Curr Opin Genet Dev **18**(2): 152-158.
- Irwin, J. J., T. Sterling, et al. (2012). "ZINC: A Free Tool to Discover Chemistry for Biology." J Chem Inf Model.
- Jacobson, M. P., D. L. Pincus, et al. (2004). "A hierarchical approach to all-atom protein loop prediction." Proteins **55**(2): 351-367.
- Jansson, M., S. T. Durant, et al. (2008). "Arginine methylation regulates the p53 response." Nat Cell Biol **10**(12): 1431-1439.
- Jauch, R., H. C. Yeo, et al. (2007). "Assessment of CASP7 structure predictions for template free targets." Proteins **69 Suppl 8**: 57-67.
- Jenuwein, T. and C. D. Allis (2001). "Translating the histone code." Science **293**(5532): 1074-1080.
- Jenuwein, T., G. Laible, et al. (1998). "SET domain proteins modulate chromatin domains in eu- and heterochromatin." Cell Mol Life Sci **54**(1): 80-93.

- Kang, M. Y., B. B. Lee, et al. (2007). "Association of the SUV39H1 histone methyltransferase with the DNA methyltransferase 1 at mRNA expression level in primary colorectal cancer." *Int J Cancer* **121**(10): 2192-2197.
- Karagiannis, T. C. and A. El-Osta (2006). "Clinical potential of histone deacetylase inhibitors as stand alone therapeutics and in combination with other chemotherapeutics or radiotherapy for cancer." *Epigenetics* **1**(3): 121-126.
- Kellenberger, E., J. Y. Springael, et al. (2007). "Identification of nonpeptide CCR5 receptor agonists by structure-based virtual screening." *J Med Chem* **50**(6): 1294-1303.
- Kim, K. C. and S. Huang (2003). "Histone methyltransferases in tumor suppression." *Cancer Biol Ther* **2**(5): 491-499.
- Kopp, J. and T. Schwede (2004). "Automated protein structure homology modeling: a progress report." *Pharmacogenomics* **5**(4): 405-416.
- Kornberg, R. D. (1974). "Chromatin structure: a repeating unit of histones and DNA." *Science* **184**(4139): 868-871.
- Kouzarides, T. (2007). "Chromatin modifications and their function." *Cell* **128**(4): 693-705.
- Krouwels, I. M., K. Wiesmeijer, et al. (2005). "A glue for heterochromatin maintenance: stable SUV39H1 binding to heterochromatin is reinforced by the SET domain." *J Cell Biol* **170**(4): 537-549.
- Kubicek, S., R. J. O'Sullivan, et al. (2007). "Reversal of H3K9me2 by a small-molecule inhibitor for the G9a histone methyltransferase." *Mol Cell* **25**(3): 473-481.
- Lachner, M., D. O'Carroll, et al. (2001). "Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins." *Nature* **410**(6824): 116-120.
- Laskowski, R. A., V. V. Chistyakov, et al. (2005). "PDBsum more: new summaries and analyses of the known 3D structures of proteins and nucleic acids." *Nucleic Acids Res* **33**(Database issue): D266-268.
- Laskowski, R. A., MacArthur, M.W., Moss, D.S. and Thornton, J.M. (1993) "PROCHECK: A program to check the stereochemical quality of protein structures." **26**, 283-291.
- Loenen, W. A. (2006). "S-adenosylmethionine: jack of all trades and master of everything?" *Biochem Soc Trans* **34**(Pt 2): 330-333.
- Luger, K., A. W. Mader, et al. (1997). "Crystal structure of the nucleosome core particle at 2.8 Å resolution." *Nature* **389**(6648): 251-260.
- Luger, K. and T. J. Richmond (1998). "The histone tails of the nucleosome." *Curr Opin Genet Dev* **8**(2): 140-146.
- Macarron, R., M. N. Banks, et al. (2011). "Impact of high-throughput screening in biomedical research." *Nat Rev Drug Discov* **10**(3): 188-195.

- Martin, C. and Y. Zhang (2005). "The diverse functions of histone lysine methylation." Nat Rev Mol Cell Biol **6**(11): 838-849.
- McManus, K. J., V. L. Biron, et al. (2006). "Dynamic changes in histone H3 lysine 9 methylations: identification of a mitosis-specific function for dynamic methylation in chromosome congression and segregation." J Biol Chem **281**(13): 8888-8897.
- Melcher, M., M. Schmid, et al. (2000). "Structure-function analysis of SUV39H1 reveals a dominant role in heterochromatin organization, chromosome segregation, and mitotic progression." Mol Cell Biol **20**(10): 3728-3741.
- Melo, F. and E. Feytmans (1998). "Assessing protein structures with a non-local atomic interaction energy." Journal of molecular biology **277**(5): 1141-1152.
- Michalsky, E., A. Goede, et al. (2003). "Loops In Proteins (LIP)--a comprehensive loop database for homology modelling." Protein engineering **16**(12): 979-985.
- Miranda, T. B., C. C. Cortez, et al. (2009). "DZNep is a global histone methylation inhibitor that reactivates developmental genes not silenced by DNA methylation." Mol Cancer Ther **8**(6): 1579-1588.
- Morris, G. M., D. S. Goodsell, et al. (1998). "Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function." J Comput Chem **19**(14): 1639-1662.
- Musacchio, A. and E. D. Salmon (2007). "The spindle-assembly checkpoint in space and time." Nat Rev Mol Cell Biol **8**(5): 379-393.
- Nguyen, C. T., D. J. Weisenberger, et al. (2002). "Histone H3-lysine 9 methylation is associated with aberrant gene silencing in cancer cells and is rapidly reversed by 5-aza-2'-deoxycytidine." Cancer Res **62**(22): 6456-6461.
- Nielsen, S. J., R. Schneider, et al. (2001). "Rb targets histone H3 methylation and HP1 to promoters." Nature **412**(6846): 561-565.
- Noeske, T., A. Jirgensons, et al. (2007). "Virtual screening for selective allosteric mGluR1 antagonists and structure-activity relationship investigations for coumarine derivatives." ChemMedChem **2**(12): 1763-1773.
- Notredame, C., D. G. Higgins, et al. (2000). "T-Coffee: A novel method for fast and accurate multiple sequence alignment." J Mol Biol **302**(1): 205-217.
- O'Carroll, D., H. Scherthan, et al. (2000). "Isolation and characterization of Suv39h2, a second histone H3 methyltransferase gene that displays testis-specific expression." Mol Cell Biol **20**(24): 9423-9433.
- Ohta, S., J. C. Bukowski-Wills, et al. (2010). "The protein composition of mitotic chromosomes determined using multiclassifier combinatorial proteomics." Cell **142**(5): 810-821.

- Pearson, W. R. and D. J. Lipman (1988). "Improved tools for biological sequence comparison." *Proc Natl Acad Sci U S A* **85**(8): 2444-2448.
- Peitsch, M. C. (1996). "ProMod and Swiss-Model: Internet-based tools for automated comparative protein modelling." *Biochem Soc Trans* **24**(1): 274-279.
- Peng, H. P. and A. S. Yang (2007). "Modeling protein loops with knowledge-based prediction of sequence-structure alignment." *Bioinformatics* **23**(21): 2836-2842.
- Peters, A. H., S. Kubicek, et al. (2003). "Partitioning and plasticity of repressive histone methylation states in mammalian chromatin." *Mol Cell* **12**(6): 1577-1589.
- Peters, A. H., D. O'Carroll, et al. (2001). "Loss of the Suv39h histone methyltransferases impairs mammalian heterochromatin and genome stability." *Cell* **107**(3): 323-337.
- Piekarz, R. L., R. Frye, et al. (2009). "Phase II multi-institutional trial of the histone deacetylase inhibitor romidepsin as monotherapy for patients with cutaneous T-cell lymphoma." *J Clin Oncol* **27**(32): 5410-5417.
- Qi, C., J. Chang, et al. (2002). "Identification of protein arginine methyltransferase 2 as a coactivator for estrogen receptor alpha." *J Biol Chem* **277**(32): 28624-28630.
- Qian, C. and M. M. Zhou (2006). "SET domain protein lysine methyltransferases: Structure, specificity and catalysis." *Cell Mol Life Sci* **63**(23): 2755-2763.
- Ramachandran, G. N., C. Ramakrishnan, et al. (1963). "Stereochemistry of polypeptide chain configurations." *J Mol Biol* **7**: 95-99.
- Rando, O. J. (2007). "Global patterns of histone modifications." *Curr Opin Genet Dev* **17**(2): 94-99.
- Rea, S., F. Eisenhaber, et al. (2000). "Regulation of chromatin structure by site-specific histone H3 methyltransferases." *Nature* **406**(6796): 593-599.
- Reuter, G. and P. Spierer (1992). "Position effect variegation and chromatin proteins." *Bioessays* **14**(9): 605-612.
- Rice, J. C. and C. D. Allis (2001). "Histone methylation versus histone acetylation: new insights into epigenetic regulation." *Curr Opin Cell Biol* **13**(3): 263-273.
- Rice, J. C., S. D. Briggs, et al. (2003). "Histone methyltransferases direct different degrees of methylation to define distinct chromatin domains." *Mol Cell* **12**(6): 1591-1598.
- Rost, B. (1999). "Twilight zone of protein sequence alignments." *Protein Eng* **12**(2): 85-94.
- Rufino, S. D., L. E. Donate, et al. (1996). "Analysis, clustering and prediction of the conformation of short and medium size loops connecting regular

- secondary structures." Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing: 570-589.
- Santos-Rosa, H., R. Schneider, et al. (2002). "Active genes are tri-methylated at K4 of histone H3." Nature **419**(6905): 407-411.
- Schlegel, B., C. Laggner, et al. (2007). "Generation of a homology model of the human histamine H(3) receptor for ligand docking and pharmacophore-based screening." J Comput Aided Mol Des **21**(8): 437-453.
- Schotta, G., A. Ebert, et al. (2002). "Central role of Drosophila SU(VAR)3-9 in histone H3-K9 methylation and heterochromatic gene silencing." EMBO J **21**(5): 1121-1131.
- Schotta, G., M. Lachner, et al. (2004). "A silencing pathway to induce H3-K9 and H4-K20 trimethylation at constitutive heterochromatin." Genes Dev **18**(11): 1251-1262.
- Schwede, T., J. Kopp, et al. (2003). "SWISS-MODEL: An automated protein homology-modeling server." Nucleic acids research **31**(13): 3381-3385.
- Shepherd, G. R., J. M. Hardin, et al. (1971). "Methylation of lysine residues of histone fractions in synchronized mammalian cells." Arch Biochem Biophys **143**(1): 1-5.
- Shi, Y., F. Lan, et al. (2004). "Histone demethylation mediated by the nuclear amine oxidase homolog LSD1." Cell **119**(7): 941-953.
- Soding, J., A. Biegert, et al. (2005). "The HHpred interactive server for protein homology detection and structure prediction." Nucleic acids research **33**(Web Server issue): W244-248.
- Strahl, B. D. and C. D. Allis (2000). "The language of covalent histone modifications." Nature **403**(6765): 41-45.
- Strahl, B. D., R. Ohba, et al. (1999). "Methylation of histone H3 at lysine 4 is highly conserved and correlates with transcriptionally active nuclei in Tetrahymena." Proc Natl Acad Sci U S A **96**(26): 14967-14972.
- Studier, F. W. (2005). "Protein production by auto-induction in high density shaking cultures." Protein Expr Purif **41**(1): 207-234.
- Tachibana, M., K. Sugimoto, et al. (2001). "Set domain-containing protein, G9a, is a novel lysine-preferring mammalian histone methyltransferase with hyperactivity and specific selectivity to lysines 9 and 27 of histone H3." J Biol Chem **276**(27): 25309-25317.
- Tachibana, M., J. Ueda, et al. (2005). "Histone methyltransferases G9a and GLP form heteromeric complexes and are both crucial for methylation of euchromatin at H3-K9." Genes Dev **19**(7): 815-826.
- Tan, J., X. Yang, et al. (2007). "Pharmacologic disruption of Polycomb-repressive complex 2-mediated gene repression selectively induces apoptosis in cancer cells." Genes Dev **21**(9): 1050-1063.

- Tian, H., L. Tang, et al. (2011). "Lactose Induction Increases Production of Recombinant Keratinocyte Growth Factor-2 in Escherichia coli." International Journal of Peptide Research and Therapeutics **17**(2): 123-129.
- Totrov, M. and R. Abagyan (1997). "Flexible protein-ligand docking by global energy optimization in internal coordinates." Proteins Suppl **1**: 215-220.
- Triebel, R. C., B. M. Beach, et al. (2002). "Structure and catalytic mechanism of a SET domain protein methyltransferase." Cell **111**(1): 91-103.
- Trott, O. and A. J. Olson (2010). "AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading." J Comput Chem **31**(2): 455-461.
- Tsukada, Y., J. Fang, et al. (2006). "Histone demethylation by a family of JmjC domain-containing proteins." Nature **439**(7078): 811-816.
- Vandel, L., E. Nicolas, et al. (2001). "Transcriptional repression by the retinoblastoma protein through the recruitment of a histone methyltransferase." Mol Cell Biol **21**(19): 6484-6494.
- Vasquez, M. (1996). "Modeling side-chain conformation." Curr Opin Struct Biol **6**(2): 217-221.
- Verdonk, M. L., J. C. Cole, et al. (2003). "Improved protein-ligand docking using GOLD." Proteins **52**(4): 609-623.
- Vos, L. J., J. K. Famulski, et al. (2006). "How to build a centromere: from centromeric and pericentromeric chromatin to kinetochore assembly." Biochem Cell Biol **84**(4): 619-639.
- Wallner, B. and A. Elofsson (2005). "All are not equal: a benchmark of different homology modeling programs." Protein Sci **14**(5): 1315-1327.
- Wang, H., W. An, et al. (2003). "mAM facilitates conversion by ESET of dimethyl to trimethyl lysine 9 of histone H3 to cause transcriptional repression." Mol Cell **12**(2): 475-487.
- Wijermans, P. W., M. Lubbert, et al. (2005). "An epigenetic approach to the treatment of advanced MDS; the experience with the DNA demethylating agent 5-aza-2'-deoxycytidine (decitabine) in 177 patients." Ann Hematol **84 Suppl 1**: 9-17.
- Wishart, D. S., C. Knox, et al. (2006). "DrugBank: a comprehensive resource for in silico drug discovery and exploration." Nucleic acids research **34**(Database issue): D668-672.
- Wissmann, M., N. Yin, et al. (2007). "Cooperative demethylation by JMJD2C and LSD1 promotes androgen receptor-dependent gene expression." Nat Cell Biol **9**(3): 347-353.

- Wolf, S. S. (2009). "The protein arginine methyltransferase family: an update about function, new perspectives and the physiological role in humans." Cell Mol Life Sci **66**(13): 2109-2121.
- Wu, H., J. Min, et al. (2010). "Structural biology of human H3K9 methyltransferases." PLoS One **5**(1): e8570.
- Wysocka, J., C. D. Allis, et al. (2006). "Histone arginine methylation and its dynamic regulation." Front Biosci **11**: 344-355.
- Xiang, Z., P. J. Steinbach, et al. (2007). "Prediction of side-chain conformations on protein surfaces." Proteins **66**(4): 814-823.
- Yang, A. S., K. D. Doshi, et al. (2006). "DNA methylation changes after 5-aza-2'-deoxycytidine therapy in patients with leukemia." Cancer Res **66**(10): 5495-5503.
- Zhang, X., H. Tamaru, et al. (2002). "Structure of the Neurospora SET domain protein DIM-5, a histone H3 lysine methyltransferase." Cell **111**(1): 117-127.
- Zhang, Y. and D. Reinberg (2001). "Transcription regulation by histone methylation: interplay between different covalent modifications of the core histone tails." Genes Dev **15**(18): 2343-2360.
- Zhou, H. and Y. Zhou (2002). "Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction." Protein science : a publication of the Protein Society **11**(11): 2714-2726.

Appendix

Structure function analysis of CENP-35/PHF2 during mitosis

ABSTRACT

CENP-35/PHF2 is a novel kinetochore associated protein that was found recently during a study to determine the proteins associated with mitotic chromosomes. CENP-35 is also a H3K9 demethylase, which makes it the first histone demethylase that has been shown to localize to the kinetochore. The localization of full-length CENP-35 and several truncation mutants that were prepared was studied by fluorescence microscopy. Since, there were very few mitotic cells expressing CENP-35 upon over-expression of CENP-35 by transient transfection, the effect of over-expression of CENP-35 on the entry of cells into mitosis was studied by live cell imaging.

List of Figures

Figure 1.1: Localization of CENP-35 during mitosis as shown by Ohta et al....	194
Figure 2.1: Vectors used for cloning CENP-35	196
Figure 2.2: List of primers used to clone CENP-35.....	199
Figure 3.1: Schematic representation of the truncation mutants of CENP-35 constructed	205
Figure 3.2: Fluorescence images of TMR direct ligand labeled HeLa cells transfected with HaloTag-CENP-35.	209
Figure 3.3: Fluorescence images of HeLa cells transfected with EGFP-tagged full- length CENP-35 (pEGFP-CENP35-1094/1095).....	210
Figure 3.4: Fluorescence images of HeLa cells transfected with pEGFP-CENP- 35-Ch1056/1055 (N-terminal fragment from 50-369aa)	211
Figure 3.5: Fluorescence images of HeLa cells transfected with pEGFP-CENP- 35- 1056/1061 (N-terminal fragment).....	212
Figure 3.6: Fluorescence images of HeLa cells transfected with pEGFP-CENP- 35- 1064/1053 (C-terminal fragment).....	213
Figure 3.7: Quantification of the cells in mitosis based on CENP-F staining.	215
Figure 3.8: Live cell imaging of HeLa cells transiently transfected with HaloTag fused CENP-35	217
Figure 3.9: Expression of CENP-35 fragments for antibody production	219
Figure 4.1: Variation in CENP-35 sequence.....	221

List of Tables

Table 3.1: List of the full length and truncation mutant constructs of CENP-35 prepared.....	206
---	-----

1. Chapter 1 – Introduction

1.1. CENP-35

In a study conducted to determine and characterize the proteins associated with mitotic chromosomes, CENP-35/PHF2, a novel histone demethylase was discovered to localize at the kinetochore during mitosis (Ohta, Bukowski-Wills et al. 2010) (Figure 1.1). It has a N-terminal PHD zinc finger domain, a JmjC domain at the C-terminal of the PHD zinc finger domain, a Nuclear Localization Signal (NLS) domain and a PEST sequence domain, which is a protein degradation signal peptide and is usually associated with proteins with a short half-life (Figure 3.1). JmjC domain containing proteins are usually associated with histone demethylation. Accordingly, in another independent study, CENP-35 was found to demethylate mono-methylated H3K9 in interphase cells (Wen, Li et al. 2010). In addition, the PHD finger domain was found to bind to tri-methylated H3K4 by using peptide pull down assays (Wen, Li et al. 2010). Since this is the first histone demethylase that has been reported to localize to the kinetochore during mitosis, we were interested in understanding its function at the kinetochore and the relationship between its structure and function during mitosis.

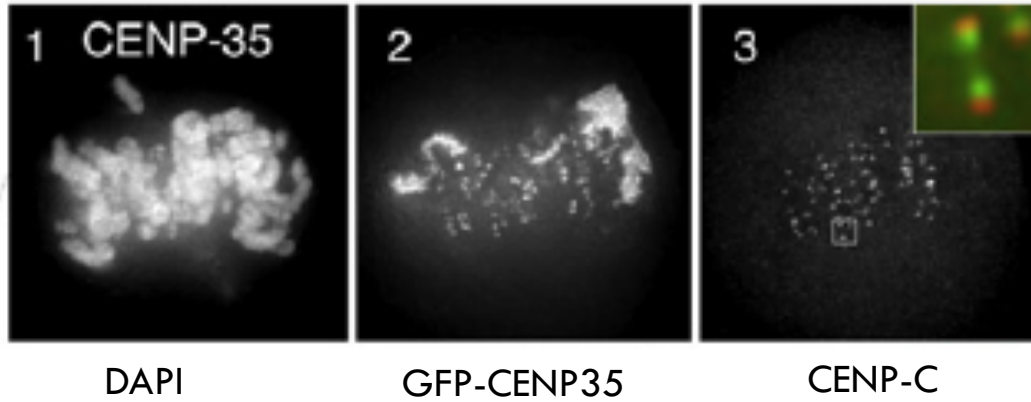


Figure 1.1: Localization of CENP-35 during mitosis as shown by Ohta et al.

The DNA is stained with DAPI (panel 1) and CENP-C (panel 3) is used a kinetochore marker. GFP-CENP-35 (panel 2) is shown to localize to the kinetochores and along the chromosomal arms at some places.

2. Chapter 2 - Materials and methods

2.1. Cloning and mutagenesis of CENP-35

To study the structural and functional role of CENP-35 at the kinetochore, several truncation mutant constructs and a full-length CENP-35 construct tagged with EGFP were prepared. Several of the truncation mutants were also tagged with MBP and GST fusion proteins in order to synthesize protein that could be used to generate antibodies against CENP-35. A list of the gateway vectors used for cloning CENP-35 mutant constructs is shown in Figure 2.1 and the primers used to clone CENP-35 are shown in Figure 2.2. A schematic diagram showing the position and size of the protein fragments is shown in Figure 3.1 and the list of constructs prepared with the vectors used is shown in Table 3.1.

Vector	Fusion tag	Gateway	Antibiotic Resistance	Source of the vector
pDONR221	None	Yes	Kanamycin	Invitrogen
pEGFP	Emerald green fluorescent protein	Yes	Ampicillin	Invitrogen
pDEST-N112-MBP	Maltose Binding protein and 6-His	Yes	Ampicillin	(Dyson, Shadbolt et al. 2004)
pDEST-N112-GST	Glutathione S-transferase and 6-His	Yes	Ampicillin	

Figure 2.1: Vectors used for cloning CENP-35

Ch1052 CENP-35 N-terminal *attB1* linker primer 27mer
Tm=66.8°C
5'-AA AAA GCA GGC TCC ATG GCG ACG GTG C-3'
K A G C M A T V

Ch1053 CENP-35 N-terminal *attB1* linker primer 35mer
Tm=66.1°C
5'-A GAA AGC TGG GTCA AAG GAG TAG TTT CCC GTT CCG-3'
* L L L K G N R

Ch1054 CENP-35 5' *attB1* linker primer (2567-2584) **aa182**
29mer Tm=63.6°C
5'-AA AAA GCA GGC TAC AGC ACC AAC CGC AAG -3'
K A G Y S T N R K

Ch1055 CENP-35 3' *attB2* linker primer (3132-3114) **aa369**
32mer Tm=65.4°C
5'-A GAA AGC TGG GT-CA CGC AGT TTC AAA GTT GGG -3'
* A T E F N P

Ch1056 CENP-35 5' *attB1* linker primer (2170-2192) **aa50**
31mer Tm=63.5°C
5'- 5'-AA AAA GCA GGC TGC CCA AAC TGT GAG AAA AC -3'
K A G C P N C E K

Ch1057 CENP-35 3' *attB2* linker primer (3396-3380) **aa458**
32mer Tm=60.3°C
5'-A GAA AGC TGG GTC-A CGA GGC GAC AGT ATT CAC -3'
* S A V T N V

Ch1058 CENP-35 3' *attB2* linker primer (3544-3526) **aa507**
33mer Tm=66.3°C
5'-A GAA AGC TGG GTCA CTT GGG GAT TTT GGA TGG C -3'
* K P I K S P

Ch1059 CENP-35 3' *attB2* linker primer (3655-3639) **aa544**
30mer Tm=65.0°C
5'-A GAA AGC TGG GTCA GTT GGG GAT GGT GGG TG -3'
* N P I T P

Ch1060 CENP-35 3' *attB2* linker primer (3814-3797) **aa597**
32mer Tm=61.3°C
5'-A GAA AGC TGG GTCA GCT CTT GGT TTG CTC TCG -3'
* S K T Q E R

Ch1061 CENP-35 3' *attB2* linker primer (4078-4061) **aa685**
32mer Tm=61.3°C
5'-A GAA AGC TGG GTCA CTC ACC GTC ATC CGA CAC -3'
* E G D D S V

Ch1062 CENP-35 3' *attB2* linker primer (4570-4553) **aa849**
32mer Tm=63.4°C
5'-A GAA AGC TGG GTCA AGC CCT CTT CAG CAG TCG -3'
* A R K L L R

Ch1063 CENP-35 3' *attB2* linker primer (4744-4727) **aa907**
32mer Tm=63.9°C
5'-A GAA AGC TGG GTCA TGT TGG GCT GTA GGG AGC -3'
* T P S Y P A

Ch1064 CENP-35 5' *attB1* linker primer (3402-3419) **aa461**
30mer Tm=64.8°C
5'-AA AAA GCA GGC TAT GAG GTG TGT GAC GGG G -3'
K A G Y E V C D G

Ch1065 CENP-35 5' *attB1* linker primer (3520-3538) **aa500**
32mer Tm=66.4°C
5'-AA AAA GCA GGC TTG CCC AAG CCA TCC AAA ATC -3'
K A G L P K P S K I

Ch1066 CENP-35 5' *attB1* linker primer (3632-3649) **aa537**
29mer Tm=65.1°C
5'-AA AAA GCA GGC TCA GCC TCA CCC ACC ATC -3'
K A G S A S P T I

Ch1067 CENP-35 5' *attB1* linker primer (4714-4731) **aa898**
31mer Tm=65.9°C
5'-AA AAA GCA GGC TCA GGC TCA GAC GAC GCT CC -3'
K A G S G S D D A

Ch1068 CENP-35 C-terminal with ORF *attB2* linker primer
34mer Tm=66.1°C
5'-A GAA AGC TGG GTA AAG GAG TAG TTT CCC GTT CCG-3'
F A P Y L L L K G N R

Ch1094 CENP-35 N-terminal full-length *attB1* linker primer
47mer, Tm=63.6°C, GC%=55.5
5'-G GGG ACA AGT TTG TAC AAA AAA GCA GGC TCG ATC GCT TTC GAA GGA G-3'
G T S L Y K K A G S I A F E G

Ch1095 CENP-35 C-terminal full length *attB2* linker primer
52mer, Tm=66.1°C,
5'-G GGG AC CAC TTT GTA CAA GAA AGC TGG GTCA AAG GAG TAG TTT CCC GTT CCG-3'
* L L L K G N R

Ch1098 CENP-35 N-terminal *attB1* linker primer (1998-2015)
Tm=63.6°C
5'-AA AAA GCA GGC T-CG ATC GCT TTC GAA GGA G-3'
K A G S I A F E G

Ch1099 CENP-35 3' *attB2* linker primer (2874-2854) **aa283**
Tm=63.4°C
5'-A GAA AGC TGG GT-CA GCG CTC ATA CAG GGA GAT G-3'
* R E Y L S I

Ch1100 CENP-35 3' *attB2* linker primer (2185-2167) **aa54**
Tm=64.5°C
5'-A GAA AGC TGG GTTA-CTC ACA GTT TGG GCA GTG G-3'
* E C N P C H

Ch1101 CENP-35 3' *attB2* linker primer (2263-2246) **aa80**
Tm=66.0°C
5'-A GAA AGC TGG GTTA-CTT GAC GTC AGG CGC TTG-3'
* K V D P A Q

Ch1102 CENP-35 3' *attB2* linker primer (2359-2342) **aa112**
Tm=63.3°C
5'-A GAA AGC TGG GTTA-CTG ACT TCC TGG CAC ACG-3'
* Q S G P V R

Ch1103 CENP-35 5' *attB1* linker primer (2037-2054) **aa6**
Tm=65.7°C
5'-AA AAA GCA GGC T-CC GTG TAC TGC GTC TGC C-3'
K A G S V Y C V C

Ch1104 CENP-35 5' *attB1* linker primer (2055-2073) **aa12**
Tm=63.4°C
5'-AA AAA GCA GGC T-GG CTC CCC TAC GAC GTT AC-3'
K A G W L P Y D V

Figure 2.2: List of primers used to clone CENP-35

2.2. Cell Culture

HeLa 9 (CCL-2) were grown as a monolayer in DMEM low glucose media supplemented with 2mM L-glutamine and 10% FBS in a humidified incubator at 37°C with 5% CO₂.

Cells were seeded onto 22X22 mm coverslips (No. 1.5) in 35mm dishes at 1x10⁵ cells/dish and transfected 24 hours later. Cultures were enriched for mitotic cells by arresting them in metaphase with 10µM MG-132 (a 26S proteasome inhibitor which prevents the separation of sister chromatids) approximately 24 hours after transfection for 2 hours. This results in a high proportion of cells stuck in metaphase. The cells were then subsequently fixed and analyzed by fluorescent microscopy.

2.3. PEI transfection

For each transfection, 2µg of plasmid DNA was diluted in 150µl of Opti-MEM[®] and vortexed. 10µl PEI (1mg/ml) was combined with 100µl of Opti-MEM[®] per sample in separate tube and vortexed and incubated at room temperature for 5 minutes. The diluted PEI solution was added to each plasmid sample, vortexed and incubated for 15 minutes at room temperature to allow plasmid:PEI complex formation. This solution was further diluted with 2ml of media and then added to each dish containing HeLa cells and the dishes were gently shaken to mix. The cells were incubated at 37°C in a humidified 5% CO₂ incubator for 8 hours before the media was changed.

2.4. HaloTag technology

The full-length CENP-35 construct was purchased from Kazusa DNA Research Institute (Clone name pFN21ASDA0662). This clone has a N-terminal fusion of the HaloTag protein (Promega) with CENP-35. The HaloTag protein (33kDa) is an engineered, catalytically inactive derivative of a hydrolase and has very high specificity and binds covalently to a specific set of ligands. The HaloTag technology can be used for cellular imaging (live cell and fixed cell), protein purification from bacterial and mammalian cells, protein interaction studies and protein-DNA interaction studies (Randall Learish 2005; Melissa McCornack 2008; Méndez 2010). In this case HaloTag was used for cellular imaging and there are several HaloTag ligands that are tagged with different kinds fluorescent dyes, which can be excited at different wavelengths.

The TMR Direct (Promega) ligand, which is a membrane permeable ligand and has an excitation and emission wavelength of 555nm and 585nm respectively was used for live cell and fixed cell imaging of CENP-35 in HeLa cells.

2.5. Fluorescence microscopy

For fluorescence microscopy of GFP-tagged constructs, HeLa cells were seeded onto 22mm² coverslips at a density of 5×10^4 cells/ml (2ml) in a 35mm dish for 24 to 48 hours prior to transfection. The cells were transiently transfected using PEI as described above. The cells were fixed with 3.5% paraformaldehyde in PBS with 10mM PIPES (pH 6.8) for 7 minutes at RT and permeabilized in KB (50mM

Tris-HCl, pH7.4, 150mM NaCl, 0.1% BSA) with 0.2% Triton X-100 for 5 minutes at room temperature and then washed with KB buffer for 5 min. DNA was stained with 0.1µg/ml 4',6'-diaminophenylindole (DAPI). Rabbit anti-CENP-F (1:1500) and Alexa 488 goat anti-rabbit (1:1000, Molecular probes) antibodies were used to detect CENP-F. Coverslips were mounted with 1mg/ml Mowiol4-88 (Calbiochem) in phosphate buffer (pH 7.4)

For imaging cells transfected with HaloTag-CENP-35, the cells were grown as described and 24 hours later, the cells were labeled with TMR Direct ligand (Promega) by incubating with the cells for 15 minutes. The ligand was removed by washing with PBS and the cells were fixed and permeabilized as described.

Cells were visualized with 100X Plan-Apochromatic objective on a Zeiss AxioPlan 2 microscope. Images were captured with a Photometrics CoolSNAP HQCCD camera (Roper scientific Inc., Trenton NJ) controlled by Metamorph 6.0 software (Universal Imaging Corporation, Downingtown, PA). Image processing was performed using Adobe Photoshop CS2.

2.6. Live cell imaging

Cells expressing CENP-35 protein fused with HaloTag were plated on glass bottom imaging dishes with glass bottom. The cells were labeled with TMR direct ligand for 10 hours and DNA was stained with Hoechst for 20 minutes and washed off. The cells were imaged using the spinning disk confocal microscope with a stage warmer and a Zeiss 710 LSM 63x Plan-Apochromatic 1.4 Oil DIC

M27 objective. The cells were imaged at 15 minutes time interval. The images were then analyzed with the Imarisx64 7.3.0 Surface rendering algorithm.

3. Chapter 3 – Results

3.1. Cloning and Mutagenesis of CENP-35

Initially, because of its large size it was difficult to amplify the full-length CENP-35. However, by varying the PCR parameters and designing new primers incorporating the *attB* linker sequences (Ch1094 and Ch1095) the full-length CENP-35 was amplified successfully. The amplified fragment was then recombined in frame with EGFP using the Gateway system and the pEGFP construct.

Several truncation mutants of CENP-35 tagged with different fusion protein were prepared. Figure 3.1 is a schematic representation of the size and position of the truncation mutants prepared. A list of the CENP-35 constructs prepared is provided in Table 3.1. The table lists the primers used to make the truncation, size of the truncation mutant and the vectors used to clone the fragment. Some of the fragments were cloned into MBP and GST constructs in order to synthesize protein for antibody production.

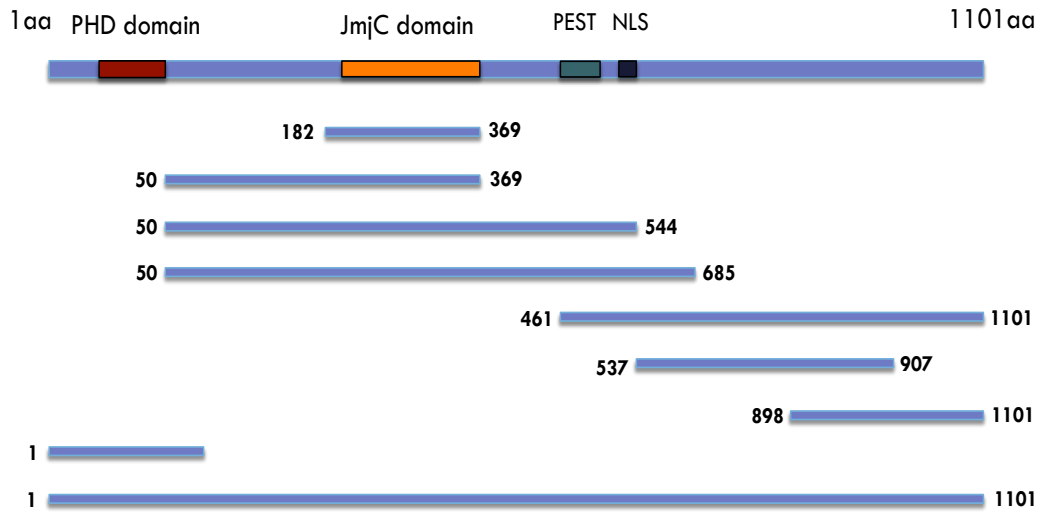


Figure 3.1: Schematic representation of the truncation mutants of CENP-35 constructed

CENP-35 is a 1101 amino acid protein and consisting of an N-terminal PHD domain, a JmjC domain, PEST and NLS domains. Truncation mutants are represented by the blue lines in their relative positions to the full length CENP35.

Primer numbers	Size of fragment (bp)	Amino acids encoded	Vectors used
1054/1055	2567-3132, 565	182-369	pEGFP pDEST-N112-GST
1056/1053	2170-5327, 3157	50-1101	pDONR221
1056/1055	2170-3132, 962	50-369	pEGFP pDEST-N112-MBP pDEST-N112-GST
1056/1060	2170-3814, 1644 (contains JmjC, PEST and NLS domains)	50-597	pEGFP pDEST-N112-MBP pDEST-N112-GST
1056/1061	2170-4078, 1908 (contains JmjC, PEST and NLS domains)	50-685	pEGFP, pDEST-N112-GST
1064/1061	3402-4078, 676	461-685	pDEST-N112-MBP
1064/1053	3402-5327, 1925	461-1101	pEGFP
1066/1053	3632-5327, 1695 (C-terminal without any of the domains)	537-1101	pDEST-N112-MBP pDEST-N112-GST
1066/1063	3632-4744, 1112	537-907	pEGFP pDEST-N112-MBP pDEST-N112-GST
1067/1053	4714-5327, 613	898-1101	pEGFP
1103/1101	2037-2263, 74	6-80	pEGFP pDEST-N112-MBP pDEST-N112-GST
1103/1102	2037-2359, 106	6-112	pEGFP pDEST-N112-MBP pDEST-N112-GST
1094/1095	1998-5327	Full-length	pEGFP

Table 3.1: List of the full length and truncation mutant constructs of CENP-35 prepared

3.2. CENP-35 localization during the cell cycle

Studies of the localization of full-length CENP-35 with HaloTag fusion protein show that the protein localizes at the nucleolus during interphase and disappears during mitosis as seen in Figure 3.2. CENP-F was used as a mitotic marker (Liao, Winkfein et al. 1995). CENP-F localizes to the kinetochores in late G2 phase and remains there through early anaphase (Rattner, Rao et al. 1993).

The full-length and truncation mutant constructs tagged with EGFP fusion protein were also used to study the localization pattern of CENP-35 during the cell cycle. The full-length CENP-35 fused with EGFP construct (pEGFP-CENP-35-Ch1094/1095) was used to transfect HeLa cells and the localization of GFP-CENP-35 was studied by fluorescence microscopy. Even though the GFP-CENP-35 localized to nucleoli during interphase, surprisingly, unlike the HaloTag fused CENP-35, the EGFP fused CENP-35 localized to the chromosome arms during mitosis as seen in Figure 3.3.

The N-terminal fragment (50-369aa) containing the JmjC domain was tagged with EGFP (pEGFP-CENP-35-Ch1056/1055) and transfected into HeLa cells and the localization was observed. This fragment localizes to the nucleoli during interphase, at the chromosomal arms during prometaphase and at the spindle poles and chromosomal arms during metaphase as seen in Figure 3.4.

Another N-terminal fragment (50-685aa) containing the JmjC domain, PEST and NLS was tagged with EGFP (pEGFP-CENP-35-Ch1056/1061) and transfected into HeLa cells and the localization was observed. This fragment also localizes to the nucleoli during interphase, at the chromosomal arms during

prometaphase and at the spindle poles and chromosomal arms during metaphase as seen in Figure 3.5.

Another fragment containing the PEST and NLS upto the C-terminal of CENP-35 (461-1101) was cloned into pEGFP (pEGFP-CENP-35-Ch1064/1053) and transfected into HeLa cells and the localization was observed. This fragment also localizes to the nucleoli during interphase, at the chromosomal arms during prometaphase and at the spindle poles and chromosomal arms during metaphase as seen in Figure 3.6.

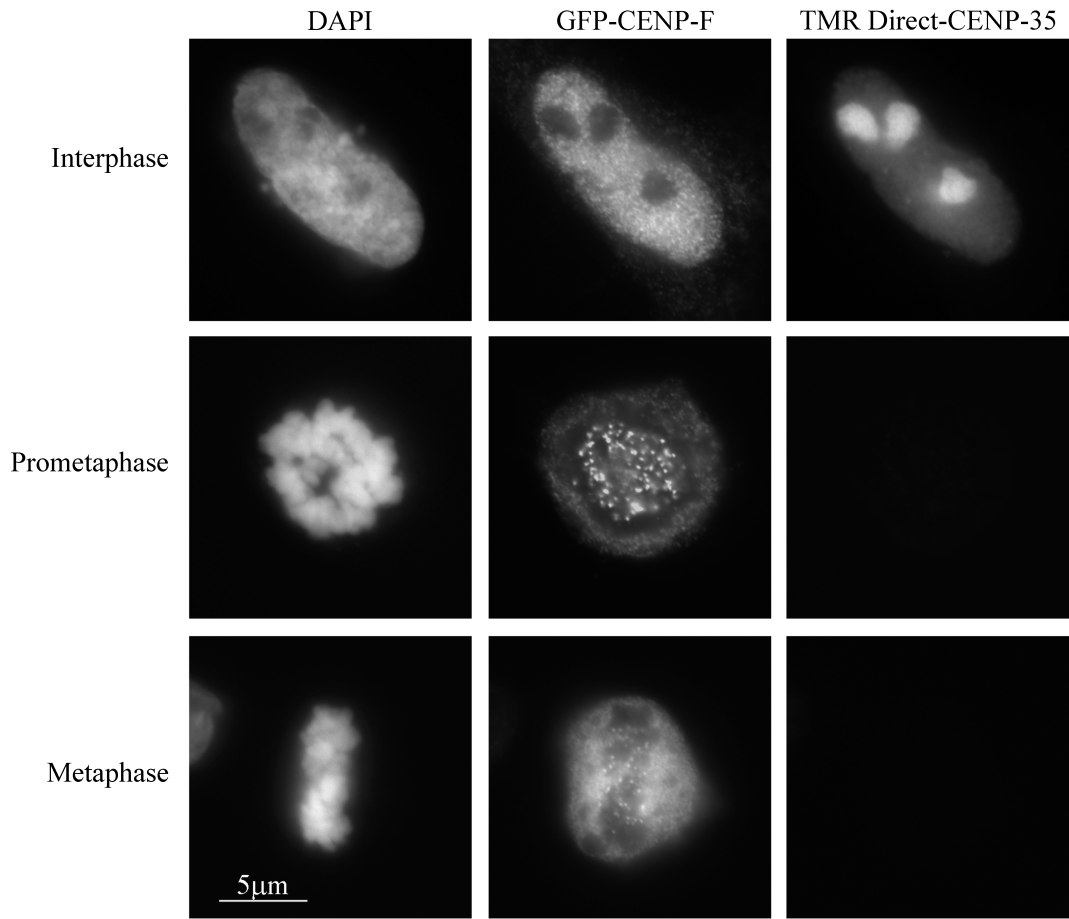


Figure 3.2: Fluorescence images of TMR direct ligand labeled HeLa cells transfected with HaloTag-CENP-35.

HeLa cells were transfected with the HaloTag-CENP-35 construct and the cells were labeled with TMR direct ligand. The top panels are interphase cells, middle are pro-metaphase and the bottom are metaphase cells. CENP-F was used as a centromere marker and was stained by using anti-CENP-F antibodies (middle panels). DNA was stained with DAPI (left panels) TMR direct ligand labeling shows that CENP-35 localizes to the nucleoli in interphase and is not found in mitotic cells. Scale bar = 5µm

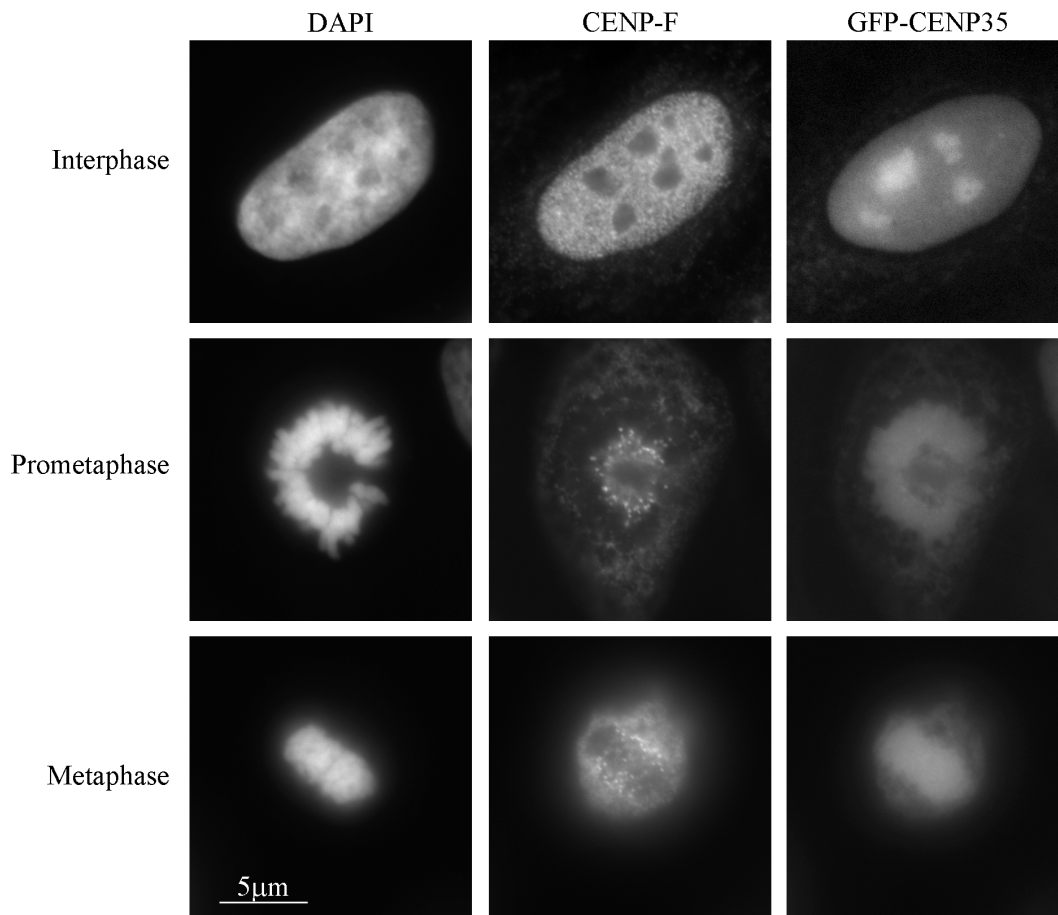


Figure 3.3: Fluorescence images of HeLa cells transfected with EGFP-tagged full-length CENP-35

HeLa cells were transfected with pEGFP-CENP35-1094/1095 EGFP construct. The top panels are interphase cells, middle are pro-metaphase and the bottom are metaphase cells. CENP-F was used as a centromere marker and was stained by using anti-CENP-F antibodies (middle panels). DNA was stained with DAPI (left panels). EGFP-CENP-35 localizes to the nucleoli in interphase and to the chromosomal arms during mitosis. Scale bar = 5µm.

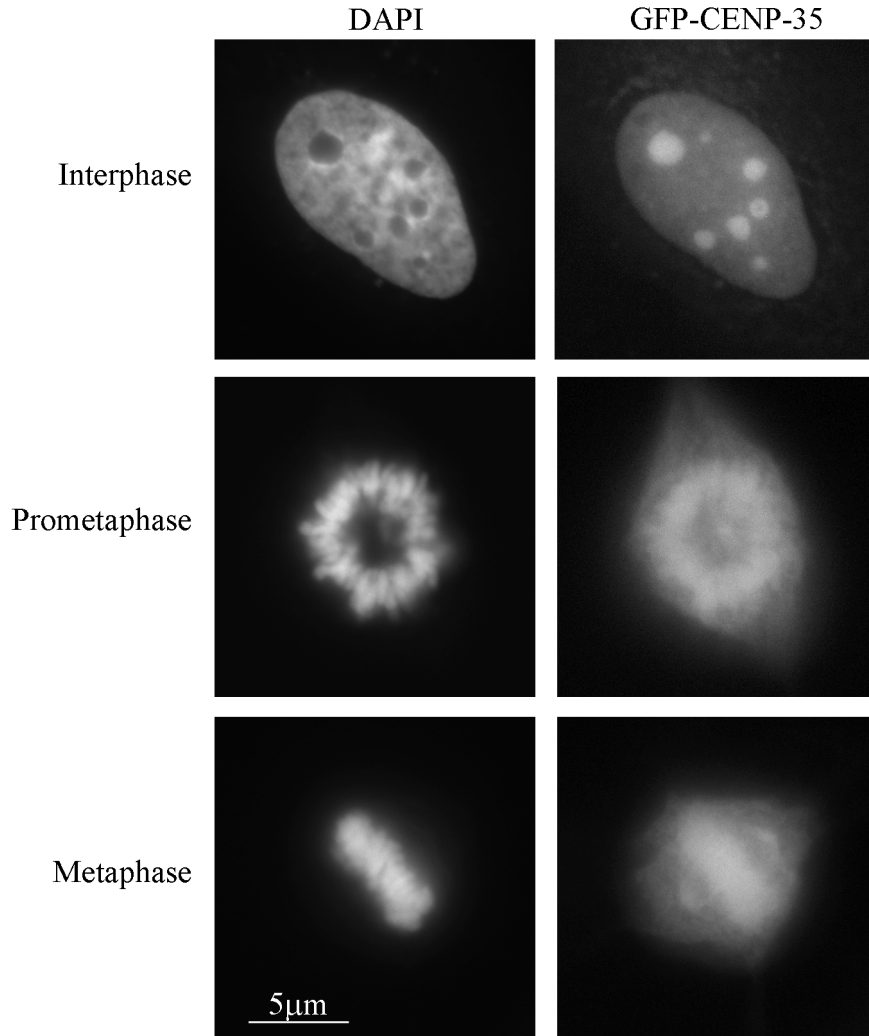


Figure 3.4: Fluorescence images of HeLa cells transfected with pEGFP-CENP-35-Ch1056/1055 (N-terminal fragment from 50-369aa)

HeLa cells were transfected with pEGFP-CENP-35-Ch1056/1055. CENP-F was used as a centromere marker and was stained by using anti-CENP-F antibodies. However no immunofluorescence of CENP-F was seen due to miscellaneous reasons and therefore the pictures have not been shown. DNA was stained with DAPI. This fragment localizes to the nucleoli in interphase, chromosomal arms during prometaphase and at the spindle poles and chromosomal arm during metaphase. Scale bar = 5µm.

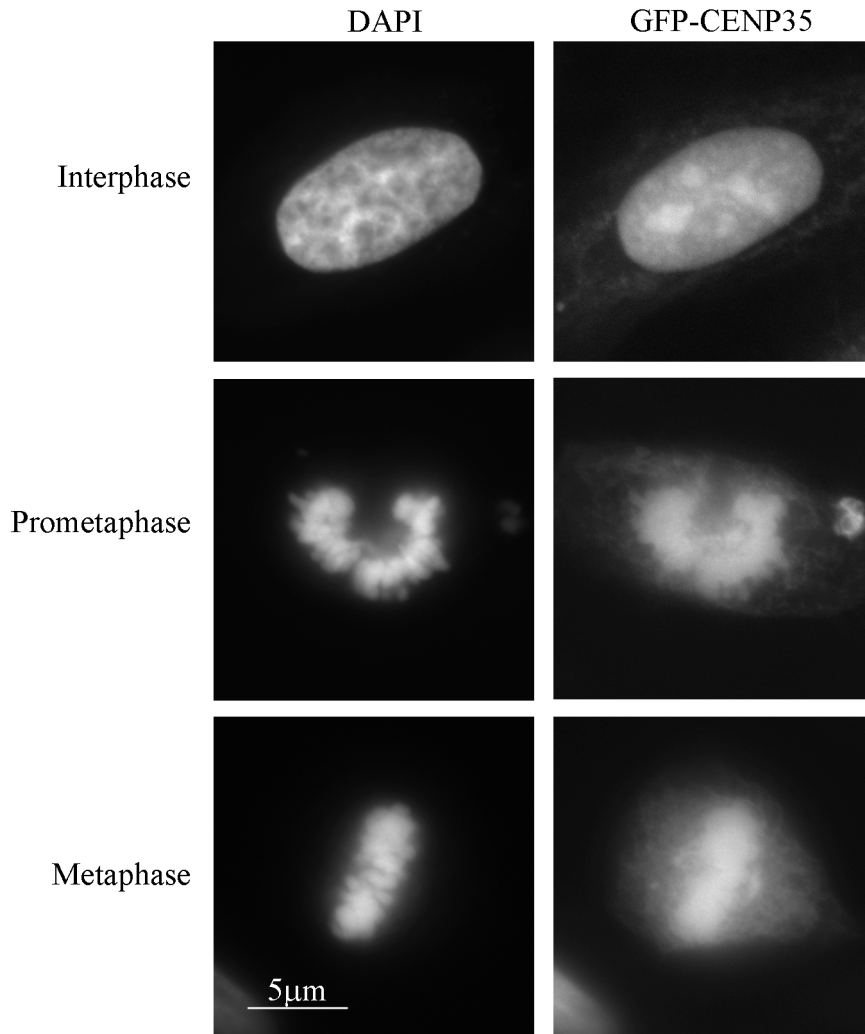


Figure 3.5: Fluorescence images of HeLa cells transfected with pEGFP-CENP-35- 1056/1061 (N-terminal fragment)

HeLa cells were transfected with pEGFP-CENP-35-Ch1056/1061. CENP-F was used as a centromere marker and was stained by using anti-CENP-F antibodies. DNA was stained with DAPI. However no immunofluorescence of CENP-F was seen due to miscellaneous reasons and therefore the pictures have not been shown. This fragment localizes to the nucleoli in interphase, chromosomal arms during prometaphase and at the spindle poles and chromosomal arm during metaphase. Scale bar = 5µm.

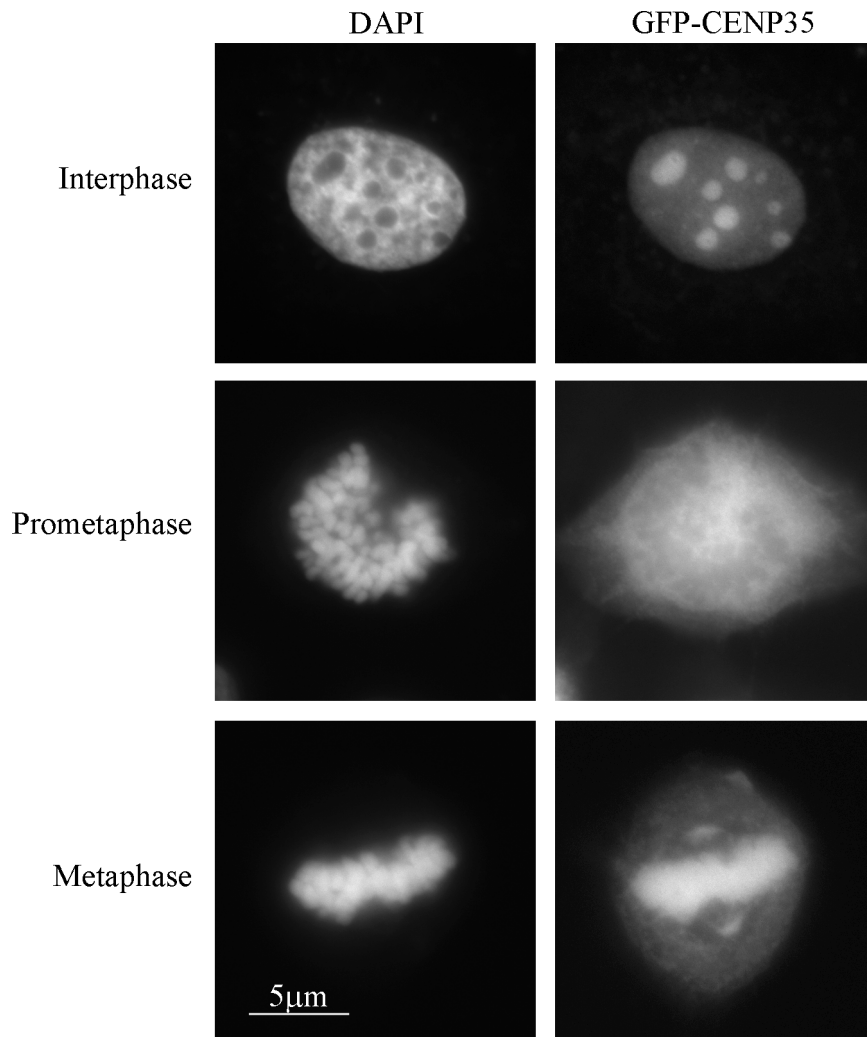


Figure 3.6: Fluorescence images of HeLa cells transfected with pEGFP-CENP-35- 1064/1053 (C-terminal fragment)

HeLa cells were transfected with pEGFP-CENP-35-Ch1064/1053. CENP-F was used as a centromere marker and was stained by using anti-CENP-F antibodies. DNA was stained with DAPI. However no immunofluorescence of CENP-F was seen due to miscellaneous reasons and therefore the pictures have not been shown. This fragment localizes to the nucleoli in interphase, chromosomal arms during prometaphase and at the spindle poles and chromosomal arm during metaphase. Scale bar = 5µm.

3.3. Cell cycle stage quantification of CENP-35 transfected cells

Despite the treatment with MG-132 in order to enrich for mitotic cells, very few mitotic cells were seen upon overexpression of CENP-35. This lead us to hypothesize that overexpression of CENP-35 prevents the cells from entering into G2/M phase. In order to test this hypothesis HeLa cells were transfected with full-length GFP-CENP-35 and were fixed 24, 36 and 48 hours post transfection. The number of cells in mitotisis was determined by CENP-F staining. Untransfected HeLa cells were used as controls. 100 cells were counted per slide. The results show that there is no correlation between CENP-35 overexpression and the ability of cells to enter mitosis as seen in Figure 3.7.

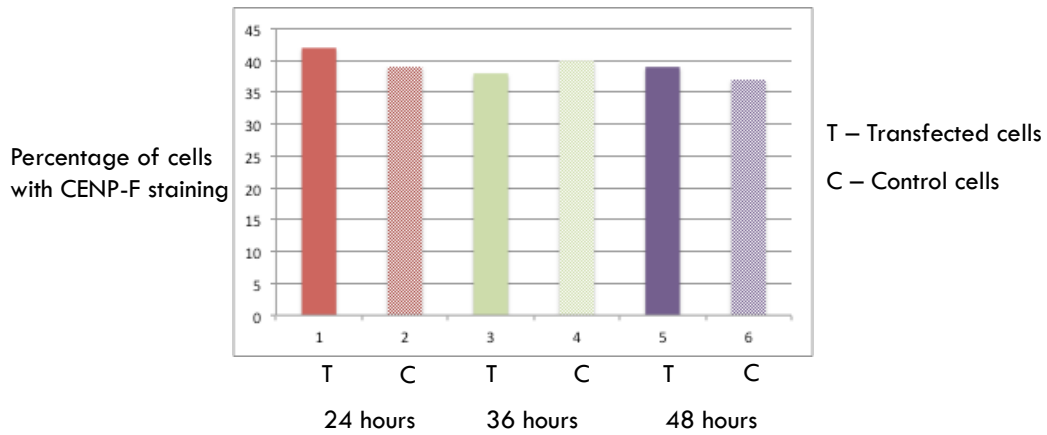


Figure 3.7: Quantification of the cells in mitosis based on CENP-F staining.

HeLa cells were transfected with full-length GFP-CENP-35 and the cells were fixed after 24, 36 and 48 hours. Untransfected HeLa cells were used as controls. 100 cells were counted per slide and the number of cells in mitosis was determined based on CENP-F staining. The percentage of cells in mitosis fixed at different time points shows there is very little difference between CENP-35 transfected and untransfected cells in their ability to enter mitosis.

3.4. Live cell imaging

Live cell imaging was done to study the effect of overexpression of CENP-35 during mitosis. The cells were transfected with CENP-35 fused with HaloTag and the labeled with TMR direct ligand and the DNA was stained with Hoechst. The cells in interphase were chosen and imaged for 10 hours with images taken at 15 minute time interval. Even after 10 hours, it was seen that the cells did not enter mitosis.

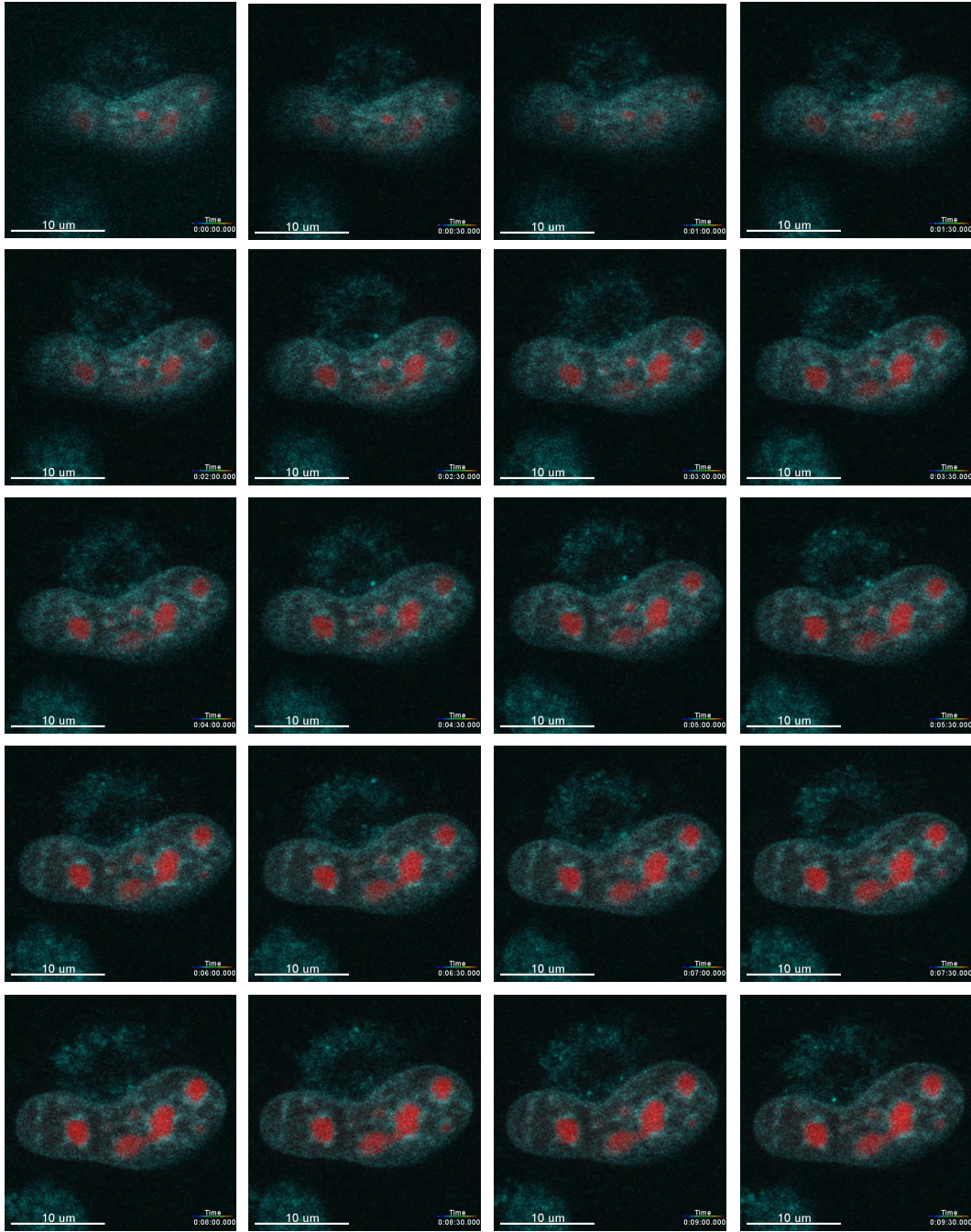


Figure 3.8: Live cell imaging of HeLa cells transiently transfected with HaloTag fused CENP-35

HeLa cells overexpressing Halo-tag-CENP-35 do not enter mitosis even after 10 hours. The HaloTag fused CENP-35 is labeled with TMR direct ligand and is shown in red. DNA is stained with Hoechst and is shown in cyan.

3.5. Expression of CENP-35 truncation fragment proteins

Several truncation mutant constructs fused with MBP and GST proteins were prepared as listed in Table 3.1. The truncation mutant constructs were transformed into *E.coli* and protein expression was induced by the addition of 0.5mM and 1.0mM IPTG. Upon analyzing the soluble and insoluble fractions, it was found that none of the protein fragments were expressed in the soluble fraction.

Even though they were expressed in the insoluble fraction CENP-35-MBP-Ch1064/1061 and CENP-35-GST-1056/1061 were expressed in large quantities by large induction (400ml of LB media and 1.0mM IPTG). The insoluble fraction was resolved on a 12.5% polyacrylamide gel by SDS-PAGE stained with aqueous coomassie blue and the band at the right size was cut out and frozen. The antibodies are being generated by in collaboration with Dr. Tim J. Yen at Fox Chase Cancer Institute, USA.

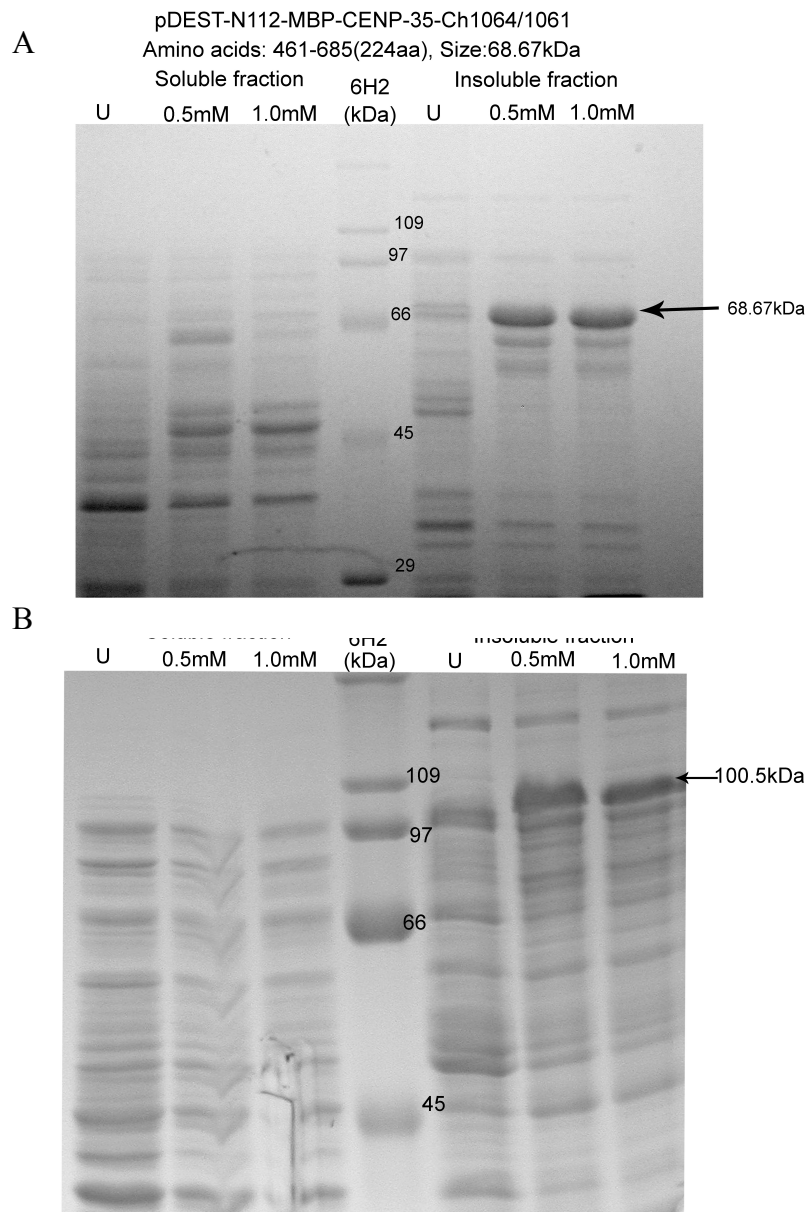


Figure 3.9: Expression of CENP-35 fragments for antibody production

Expression of CENP-35-MBP-1064/1061 (461-685aa) and CENP-35-GST-1056/1061 (50-685aa) fragments were induced in *E.coli* by the addition of 0.5mM and 1.0mM IPTG. The soluble and insoluble fractions were separated by SDS-PAGE. Both the proteins were expressed in the insoluble fraction as shown by the arrow marking 100.5 kDa corresponding to the correct MW for the fusion proteins. SDS-6H2 (Sigma-Aldrich) was used as protein marker. The sizes of the molecular marker are shown in the center lane.

4. Chapter 4 –Discussions and Future Directions

The results obtained from the localization of full-length CENP-35 fused with HaloTag and EGFP-CENP-35 are different during prometaphase and metaphase. The results of these experiments show that HaloTag fused CENP-35 does not localize at the chromosomal arms during prometaphase and metaphase, where as the EGFP-CENP-35 does localize. However these results are still different from the reported localization pattern as seen in Figure 1.1. These discrepancies might be due to the variation in the experimental techniques and the level of expression of CENP-35 upon transient transfection.

Another variant of CENP-35 that is different from the one used in this experiment has been reported (NM_005392.3). These sequences are 98% identical to each other with the 2% variation occurring at the N-terminal, which is shown in Figure 4.1. This variation in the sequence might be another reason for the difference in localization pattern reported by Ohta et al. and the results reported in this thesis.

```

PHF2_NM_005392. 1 ATGGCGACGGTGCCCGTGTACTGCGTCTGCCGGCTCCCCTACGACGTTACCCGCTTCATG 60
HaloTag-CENP-35 1 ATGGCGACGGTGCCCGTGTACTGCGTCTGCCGGCTCCCCTACGACGTTACCCGCCCC-CG 59
*****
PHF2_NM_005392. 61 ATCGAGTGCAGCGCCTGCAAGGACTGGTTCCACGGCAGCTGTGTTGGGGTGAAGAGGAG 120
HaloTag-CENP-35 60 CGCGGCCCGGCCCGGCCCGGCCCGGCCCGGCCCGGCCCGGCCCGGCCCGGCCCGGCCCGGCC 119
** * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
PHF2_NM_005392. 121 GA-GGCGCCCGACATCGACATATACCACTGCCCAAACGTGTGAGAAAACCCATGGGAAGTC 179
HaloTag-CENP-35 120 GACGGCGCCCGACATCGACATATACCACTGCCCAAACGTGTGAGAAAACCCATGGGAAGTC 179
** *****

```

Figure 4.1: Variation in CENP-35 sequence

Although the CENP-35 sequence used in this experiment and another reported CENP-35/PHF2 sequence (NM_005392.3) are 98% identical, a 2% difference in their sequences is concentrated at the N-terminal region as shown in the figure.

Although, HaloTag ligands are claimed to be extremely specific to the HaloTag protein, the amount of protein required for the binding of the ligand has to be optimized. Further experiments, will have to be done in order to optimize the over expression of HaloTag fused CENP-35 and proper labeling by the TMR ligand.

The HaloTag technology was used primarily because of the difficulties associated with the amplification of the full-length CENP-35 due to its large size. Since, this issue was resolved by designing better primers and optimizing PCR parameters and the full-length CENP-35 was cloned successfully into pEGFP vector, it might be unnecessary to do the optimization experiments with the HaloTag fused CENP-35.

The truncation mutants were used to study the localization patterns and their effect on the different stages of mitosis. Despite several efforts, it was not possible to amplify fragments starting at the N-terminal (starting from amino acid 1). The N-terminal region has very low GC content, therefore the primers designed were unable to bind properly and amplify the N-terminal fragment. This was however overcome later by increasing the length of the primer and including the *attB* linker sequence (Ch1094). This primer can now be used to design the extreme N-terminal truncation mutants. Since, this was done later on, the primer nearest to the N-terminal (aa50) was used for preparing N-terminal truncation mutants. The localization patterns of the N-terminal truncation mutants and the C-terminal mutant were found to be similar. The truncation mutants localize to the nucleolus during interphase, at the chromosomal arms during prometaphase and the spindle poles and at the chromosomal arms during metaphase.

Although the cell cycle quantification studies revealed that there is no correlation between CENP-35 overexpression and entry into mitosis, the live cell imaging studies contradicted this. However, it needs to be kept in mind that the live cell imaging was done with the HaloTag fused CENP-35 and this needs to be repeated with the pEGFP-CENP-35 to prove the effect of overexpression of CENP-35 on mitotic entry. Apart from the full-length CENP-35, live cell imaging with the truncation mutants must also be done in order to study their effect during mitosis. Further experiments are necessary in order to decipher the structure and functional relationship of CENP-35 during mitosis.

Bibliography:

- Dyson, M. R., S. P. Shadbolt, et al. (2004). "Production of soluble mammalian proteins in *Escherichia coli*: identification of protein features that correlate with successful expression." BMC Biotechnol **4**: 32.
- Liao, H., R. J. Winkfein, et al. (1995). "CENP-F is a protein of the nuclear matrix that assembles onto kinetochores at late G2 and is rapidly degraded after mitosis." J Cell Biol **130**(3): 507-518.
- Melissa McCornack, T. S., and Michael Slater (2008) "Expression of Fusion Proteins: How to Get Started with the HaloTag® Technology." 13-15.
- Méndez, J., Murphy, N., Benink, H., Daniels, D. and Urh, M. (2010). Efficient Isolation, Identification and Labeling of Intracellular Mammalian Protein Complexes.
- Ohta, S., J. C. Bukowski-Wills, et al. (2010). "The protein composition of mitotic chromosomes determined using multiclassifier combinatorial proteomics." Cell **142**(5): 810-821.
- Randall Learish, C. Z., Christine Andrews and Georgyi Los (2005). "Perform Multicolor Live- and Fixed-Cell Imaging Applications with HaloTag® Interchangeable Labeling Technology." cell notes: 4-8.
- Rattner, J. B., A. Rao, et al. (1993). "CENP-F is a .ca 400 kDa kinetochore protein that exhibits a cell-cycle dependent localization." Cell Motil Cytoskeleton **26**(3): 214-226.
- Wen, H., J. Li, et al. (2010). "Recognition of histone H3K4 trimethylation by the plant homeodomain of PHF2 modulates histone demethylation." J Biol Chem **285**(13): 9322-9326.