In silico ligand fitting/docking, computational analysis and biochemical/biophysical validation for protein-RNA recognition and for rational drug design in diseases

In silico Ligand Fitting/Docking, Computational Analysis and Biochemical/Biophysical Validation for Protein-RNA Recognition and for Rational Drug Design in Diseases

Anathe Olivia Maria
Patschull Lafitte-Laplace

Thesis submitted for the degree of Doctor of Philosophy

Institute of Structural and Molecular Biology

Department of Biological Sciences
Birkbeck College
University of London

October 2013

Declaration

I, Anathe Olivia Maria Patschull Lafitte-Laplace, hereby declare that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated.

Anathe Olivia Maria Patschull Lafitte-Laplace
October 2013

Abstract

Kaposi's sarcoma-associated herpesvirus, is a double-stranded DNA γ - herpesvirus and the main causative agent of Kaposi's sarcoma (KS). γ - herpesviruses undergo both lytic and latent replication cycles; and encode proteins that modulate host transcription at the RNA level, by inducing decay of certain mRNAs. Here we describe a mechanism that allows the viral endo-/exonuclease SOX to recognise mRNA targets on the basis of an RNA motif and fold. To induce rapid RNA degradation by subverting the main host mRNA degradation pathway SOX was shown to directly bind Xrn1. This may shed light as to how some viruses evade the host antiviral response and how mRNA degradation processes in the eukaryotic cell are involves in this.

Contents

Acknowledgements

"Doubt grows with knowledge."
- Johann Wolfgang von Goethe -

“You can only be afraid of what you think you know.”
- Jiddu Krishnamurti -

“Creativity requires the courage to let go of certainties.“
- Erich Fromm -

I am grateful to my supervisors, Dr. Tracey Barrett and Dr. Irilenia Nobeli, for their patience and the opportunity to conduct research on RNA for the past 3 years. I am thankful to the Wellcome Trust for the Studentship and Funding. I would like to thank the Biophysics, X-ray and Rosalind Franklin Laboratory managers for the facilities Dr. Tina Daviter, Dr. Ambrose Cole/Dr. Nora Cronin and Dr. Robert Sarra /Dr. Stella Geddes/Dr. Renos Savva. For his help with the Microscale Thermophoresis experiments I would like to thank Dr. James Wilkinson. “Professor Schrödinger” would like to thank Dr. David Holdershaw and Richard Westlake for their IT support and endless sarcasm. I am indebted to Dr. Mun Peak Nyon, Dr. Louise Briggs and Dr. Claire Bagnéris for their knowledge and patience teaching me the ropes, 谢谢. I would like to thank Prof. Richard Goldstein, my Thesis Committee Chair, for very insightful questions, Dr. Russell Hamilton for inspiring conversation about RNA and various RNA prediction methods and Prof. David Moss for advice of statistical methods. And thank you to Dr. Adrian Shepherd for many encouraging chats. I would also like to thank Dr. Irilenia Nobeli and our three Bioinformatics MSc. Students for great work on project in the wider RNA field. I would like to thank all the Rayne-Wolfson, Rosalind Franklin and 3rd floor Computational Lab members and my fellow Wellcome Trust and BBSRC students for a great atmosphere and many friendships.

To my friends and colleagues who lived through this with me, Alex Lim, Jessie Yeung, Chia-Ying Chou, Altin Sula, Denis O’Leary, Peg, Natalie Dawson, Paul Moody, Anna Adams, Katie Griffiths, Paul Ashford, Daven Vasishtan, Oliver Willhoft, Julia Wenger, Gilles Phan, Yuriy Chaban, Guillaume Gouget, James Smith, Myriam De Gregorio, Moncef Nafir, Wei Nee Lim, Katia Gicquel, Sandra Baumgartner and Dr. Marcia Ben-Shoshan thank you.

To my teachers and mentors, who believed in me and encouraged me along the life-long learning path: Dr. Irilenia Nobeli, Dr. Bibek Gooptu, Dr. Michael Titheradge (University of Sussex), Dr. Mohammed Meah (University of East London), Mr. Florent Genatio (Lycée Français de Hambourg) and the late Ms. Chantal Briffod (Maternelle and Lycée Français de Hambourg).

And last but not least I would like to thanks my family, Susanne, Olivier, Idgie, Oma, Opa and Patrick, who have been there for me in the dark hours and have been a source of strength.

Publications

These publications are linked to the work presented in the Addition Appendix Chapter 7.

Patschull, AOM, Gooptu, B, Ashford, P, Daviter, T, Nobeli, I (2012) "*In Silico* Assessment of Potential Druggable Pockets on the Surface of alpha(1)-Antitrypsin". *Plos One*, 7(5). DOI**:** 10.1371/journal.pone.0036612

Nyon, MP, Segu, L,Cabrita, LD, Levy, GR, Kirkpatrick, J, Roussel, BD, **Patschull, AOM,** Barrett, TE, Ekeowa, UI, Kerr, R, Waudby, CA, Kalsheker, N, Hil, M, Thalassinos, K, Lomas, DA, Christodoulou, J, Gooptu, B (2012) "Structural Dynamics Associated with Intermediate Formation in an Archetypal Conformational Disease". *Structure*, 20(3), pp. 504-512. DOI: 10.1016/j.str.2012.01.012

Patschull, AOM, Segu, L, Nyon, MP, Lomas, DA, Nobeli, I, Barrett, TE, Gooptu, B, (2011) " Therapeutic target-site variability in a1-antitrypsin characterized at high resolution". *Acta Crystallographica Section F-Structural Biology And Crystallization Communications,* 67, pp. 1492-1497. DOI: 10.1107/S1744309111040267

Chang, YP, Mahadeva, R, **Patschull, AOM,** Nobeli, I, Ekeowa, UI, McKay, AR, Thalassinos, K, Irving, JA, Haq, I, Nyon, MP, Christodoulou, J, Ordonez, A, Miranda, E, Gooptu, B, (2011) "Targeting Serpins In High-Throughput And Structure-Based Drug Design" *Methods In Enzymology, Vol 501: Serpin Structure And Evolution*, 501, pp. 139-175. DOI: 10.1016/B978-0-12-385950-1.00008-0

Abbreviations

#		
	2D	**Two-Dimensional**
	3D	**Three-Dimensional**
	3'UTR	**Three Prime Untranslated Region**
	5'UTR	**Five Prime Untranslated Region**
	40S	**Eukaryotic Small Ribosomal Subunit**
	6-FAM	***6*-Carboxyfluorescein**
	80S	**Eukaryotic Ribosome**
A		
	Å	**Ångström**
	Aly	**Aly/REF Export Factor**
	APS	**Ammonium Persulfate**
	ARE	**Adenylate-Uridylate-Rich Element**
	AUF1	**AU-Rich Element RNA-Binding Protein 1**
B		
	Blast	**Basic Local Alignment Search Tool**
	BGLF5	**EBV Shutoff Alkaline Exonuclease (SOX)**
	bp	**Base Pair**
	BPS	**Branch Point Site**
	BSA	**Bovine Serum Albumin**
C		
	°C	**Degree Celsius**
	CBC	**Cap-Binding Complex**
	CBP80	**Cap Binding Protein 80**
	CCD	**Charge-Coupled Device Detectors**
	CCP4	**Collaborative Computational Project, Number 4**
	CCR4-NOT	**C-C Chemokine Receptor Type 4 NOT**
	Compseq	**Composition Sequence**
	Coot	**Crystallographic Object-Oriented Toolkit**
	Cu	**Copper**
	CV	**Column Volume**
D		
	Dcp1-Dcp2	**mRNA-Decapping Enzyme 1- mRNA-Decapping Enzyme 2**
	DcpS	**Scavenger mRNA-Decapping Enzyme**
	DEAD-Box	**Asp-Glu-Ala-Asp Box Helicase**
	DNA	**Deoxyribonucleic Acid**
	DNase	**Deoxyribonuclease**
	ds	**Double Stranded**
	dsDNA	**Double Stranded Deoxyribonucleic Acid**
	DsRed2	**Red Fluorescent Protein**
	dsRNA	**Double Stranded Ribonucleic Acid**
	DTT	**Dithiothreitol**
E		
	EBV	**Epstein–Barr Virus (HHV-4)**

	EDTA	**Ethylenediaminetetraacetic Acid**
	e.g.	**exempli gratia**
	eIF3j	**Eukaryotic Translation Initiation Factor 3 Subunit J**
	eIF2a	**Eukaryotic Translation Initiation Factor 2A**
	eIF4E	**Eukaryotic Translation Initiation Factor 4E**
	EJC	**Exon Junction Complex**
	EMBOSS	**European Molecular Biology Open Software Suite**
	EMSA	**Electron Mobility Shift Assay**
	eRF1	**Eukaryotic Translation Termination Factor 1**
	eRF3	**Eukaryotic Translation Termination Factor 3**
	ESE	**Exonic Splicing Enhancer**
	ESS	**Exonic Splicing Silencer**
F		
	FA	**Fluorescence Anisotropy also Fluorescence Polarization Anisotropy**
	FARFAR	**Fragment Assembly of RNA with Full-Atom Refinement**
G		
	GAPDH	**Glyceraldehyde 3-Phosphate Dehydrogenase**
	GFP	**Green Fluorescent Protein**
H		
	h	**Hour**
	HBB	**β-Globin**
	HCl	**Hydrochloric Acid**
	HCMV	**Human Cytomegalovirus (HHV-5)**
	HEK293T	**Human Embryonic Kidney 293 T Cells**
	HHV	**Human Herpesvirus Virus**
	HHV-1	**Human Herpesvirus Virus-1 (HSV-1)**
	HHV-2	**Human Herpesvirus Virus-2 (HSV-2)**
	HHV-3	**Human Herpesvirus Virus-3 (VZV)**
	HHV-4	**Human Herpesvirus Virus-4 (EBV)**
	HHV-5	**Human Herpesvirus Virus-5 (HCMV)**
	HHV-6	**Human Herpesvirus Virus-6 (Roseolovirus)**
	HHV-7	**Human Herpesvirus Virus-7 (Roseolovirus)**
	HHV-8	**Human Herpesvirus Virus-8 (KSHV)**
	HIV	**Human Immunodeficiency Virus**
	HPLC	**High-Performance Liquid Chromatography**
	HPLC-IEX	**High-Performance Liquid Chromatography Ion Exchange**
	HSO	**Host Shutoff**
	HSV-1	**Herpes Simplex Virus 1 (HHV-1)**
	HSV-2	**Herpes Simplex Virus 2 (HHV-2)**
	HuR	**Human Antigen R (ELAV-like Protein 1)**
I		
	i.e.	**id est**
	IEX	**Ion Exchange**
	IL-6	**Interleukin 6**
	IPTG	**Isopropyl β-D-1-Thiogalactopyranoside**
	IR	**Infrared**

	IRE1	**Serine/Threonine-Protein Kinase/Endoribonuclease IRE1**
	ISE	**Intronic Splicing Enhancer**
	ISS	**Intronic Splicing Silencer**
K		
	K	**Kelvin**
	kb	**Kilo Base**
	K_d	**Dissociation Constant**
	kDa	**Kilo Dalton**
	KS	**Kaposi's Sarcoma**
	KSHV	**Kaposi's Sarcoma-Associated Herpesvirus**
L		
	LB	**Lysogeny Broth**
	LLG	**Log-Likelihood Gradient**
	Lsm1-7	**Like Sm 1-7**
M		
	M	**Meter**
	MAD	**Multi-Wavelength Anomalous Dispersion**
	MES	**2-(N-Morpholino) Ethanesulfonic Acid**
	MFE	**Minimum Free Energy**
	MHV68	**Murine Herpesvirus 68**
	MR	**Molecular Replacement**
	MIR	**Multiple Isomorphous Replacement**
	mRNA	**Messenger Ribonucleic Acid**
	miRNA	**Micro Ribonucleic Acid**
	MST	**Microscale Thermophoresis**
	MWCO	**Molecular Weight Cut Off**
N		
	NCS	**Non-Crystallographic Symmetry**
	NGD	**No-Go Decay**
	NLS	**Nuclear Localization Signal**
	NMD	**Nonsense-Mediated mRNA Decay**
	NMR	**Nuclear Magnetic Resonance**
	NPC	**Nuclear Pore Complexes**
	NSD	**Non-Stop Decay**
	NusA	**N-Utilization Substance Protein A**
O		
	OD_{600}	**Optical Density 600 nm**
	ORF37	**Open Reading Frame 37 (Shutoff Alkaline Exonuclease (SOX))**
	ORF57	**Open Reading Frame 57 (mRNA Export Factor ICP27 Homologue)**
P		
	PABP	**Poly(A)-Binding Protein**
	PAGE	**Polyacrylamide Gel Electrophoresis**
	PAN2-PAN3	**PAB-Dependent Poly(A)-Specific Ribonuclease Subunit 2** **PAB-Dependent Poly(A)-Specific Ribonuclease Subunit 3**
	PARN	**Poly(A)-Specific Ribonuclease**

	PDB	**Protein Data Bank**
	PHENIX	**Python-based Hierarchical ENvironment for Integrated Xtallography**
	PMR1	**Polysomal Ribonuclease 1**
	Pol I	**RNA Polymerase I**
	Pol II	**RNA Polymerase II**
	Pol III	**RNA Polymerase III**
	Poly(A)tail	**Polyadenylation Tail**
	PPT	**Polypyrimidine Tract**
	pre-mRNA	**Precursor Messenger Ribonucleic Acid**
	PTC	**Premature Termination Codon**
R		
	REC102	**Meiotic Recombination Protein REC102**
	RefSeq	**Reference Sequence**
	RNA	**Ribonucleic Acid**
	RNase	**Ribonuclease**
	RNase PH	**Ribonuclease PH**
	RNP	**Ribonucleoprotein**
	RPS3	**40S Ribosomal Protein S3**
S		
	SAD	**Single-Wavelength Anomalous Diffraction**
	SDS	**Sodium Dodecyl Sulfate**
	siRNA	**Small Interfering Ribonucleic Acid**
	SIRAS	**Single Isomorphous Replacement with Anomalous Scattering**
	SMG1	**Morphogenetic Effect on Genitalia 1**
	SMG5	**Morphogenetic Effect on Genitalia 5**
	SMG6	**Morphogenetic Effect on Genitalia 6**
	SMG7	**Morphogenetic Effect on Genitalia 7**
	snRNA	**Small Nuclear Ribonucleic Acid**
	snRNP	**Small Nuclear Ribonucleoproteins**
	SOX	**Shutoff and Exonuclease**
	SR	**Serine-Arginine–Rich Proteins**
	SRE1	**SOX-Resistant Element**
	SRSF1	**Serine/Arginine-Rich Splicing Factor 1**
	ssRNA	**Single Stranded Ribonucleic Acid**
	SURF	**SMG1 -UPF1-eRF1-eRF3 Complex**
T		
	TAP	**Tandem Affinity Purification Protein**
	TBE	**Tris/Borate/EDTA**
	TBE-Urea	**Tris/Borate/EDTA-Urea**
	TEMED	**Tetramethylethylenediamine**
	TEV	**Tobacco Etch Virus Protein**
	THO	**Subcomplex of the TREX Complex**
	TREX	**Transcription and Export Complex**
U		
	UAP56	**DEAD-Box RNA Helicase**
	UPF1	**Up-Frameshift Factor 1**
	UPF2	**Up-Frameshift Factor 2**
	UPF3	**Up-Frameshift Factor 3**

	U1 snRNP	**U1 Small Nuclear Ribonucleoprotein**
	U2 snRNP	**U2 Small Nuclear Ribonucleoprotein**
	U3 snRNP	**U3 Small Nuclear Ribonucleoprotein**
	U4 snRNP	**U4 Small Nuclear Ribonucleoprotein**
	U5 snRNP	**U5 Small Nuclear Ribonucleoprotein**
V		
	vhs	**Viral Host Shutoff**
	vIL-6	**Viral Interleukin 6**
	VZV	**Varicella-Zoster Virus**
X		
	XBP1	**X-Box-Binding Protein 1**
	Xrn1	**Exoribonuclease 1**

List of Figures

List of Tables

Chapter 1: Introduction

1.1. Kaposi's Sarcoma-associated Herpesvirus

The *herpesviridae* comprise over 130 viruses, which target a wide range of host's covering invertebrates and vertebrates, from amphibians to humans and livestock (Brown and Newcomb, 2011). *Herpesviridae* are well adapted to evade the immune response and as such, following the primary infection, establish lifelong latent infections in the host, which can reactivate to lead to recurrent infections and chronic diseases (Malik and Schirmer, 2006). Each individual herpesvirus is well adapted to its specific host and target tissues. The primary and/or recurring infections are frequently asymptomatic, but in immunosuppressed individuals herpesviruses can be life-threatening, where infections can lead to different types of cancer or autoimmune diseases (Desailloud and Hober, 2009, Caselli et al., 2012). Herpesviruses have a double stranded linear deoxyribonucleic acid (DNA) genome, which is contained in an icosahedral capsid wrapped in viral proteins and viral messenger ribonucleic acids (mRNA) (tegument), which then is enveloped by a lipid bilayer membrane to form the virion.

Eight human herpesviruses (HHVs 1-8) have been identified to date. There are three subfamilies α-, β- and γ-*herpesviridae*. The α-*herpesviridae* contains herpes simplex virus types 1 and 2, (HSV-1 and 2 or HHV-1 and 2), varicella zoster virus (VZV or HHV-3), whose primary target is the mucoepithelial tissue and latency is established in neurons. The β-*herpesviridae* includes human cytomegalovirus (HCMV or HHV-5), which targets monocytes and lymphocytes (in which it establishes latency) as well as epithelial cells. This subfamily also includes roseoloviruses (HHV-6 and HHV-7), which target and establish latency in T cells (Levy, 1997, Dockrell, 2003, Salahuddin et al., 1986). Finally the γ-*herpesviridae* that encompasses the Epstein–Barr virus (EBV or HHV-4) and Kaposi's sarcoma-associated herpesvirus (KSHV or HHV-8), both of which establish latency in B cells and target lymphocytes and epithelial cells. The animal model for KSHV is the murine herpesvirus 68 (MHV68) in mouse.

KSHV is the most recently identified human herpesvirus, which, due to its lymphotropism and its intense immune evasion strategies, establishes lifelong

infections (Boshoff and Chang, 2001). Hence, the viral proteins and mechanisms that are involved in establishing latency and immune evasion are critical for KSHV pathogenesis (Dourmishev et al., 2003, Gray et al., 2012). Through its infection of endothelial cells, KSHV is the main causative agent of Kaposi's sarcoma (KS), which is associated with immunosuppressed patients and the most common human immunodeficiency virus (HIV) associated cancer. KSHV infection has also been linked to other lymphoproliferative disorders that include primary effusion lymphoma and multicentric castleman's disease (Boshoff and Weiss, 1998, Arvanitakis et al., 1996, Cesarman and Knowles, 1999).

KSHV, EBV and MHV68 are closely related γ-herpesviruses that share many proteins and molecular mechanisms and are thus intensively studied. They have a biphasic life cycle, hence undergo both a lytic and latent phase. During latency, the viral genome is circularised and found as a nuclear episome, which is tethered via the histone to the chromatin during mitosis (Rezaee et al., 2006). In the latency phase viral gene expression is repressed by the limited expression of genes in a small subjection of the episome. The viral genome is approximately 165 to 170 kilobases (kb) in length (Renne et al., 1996) and contains 86 mainly intronless genes (Rezaee et al., 2006). The lytic phase is characterized by increased viral gene expression and rapid and global host mRNA decay prior to viral replication, which climaxes with cells lysis and release of the viral progeny (Glaunsinger and Ganem, 2004). The rapid and global decay of host mRNA transcripts is a conserved mechanism amongst the *herpesviridae* and has been termed host shutoff (HSO) (Glaunsinger and Ganem, 2006). Viral genome maturation, a process required for encapsidation following replication of the viral genome, and HSO during the lytic cycle, are both dependent on a single nuclear and cytoplasmic protein called shutoff and exonuclease (SOX).

1.1.1. SOX: Role in the Lytic Phase

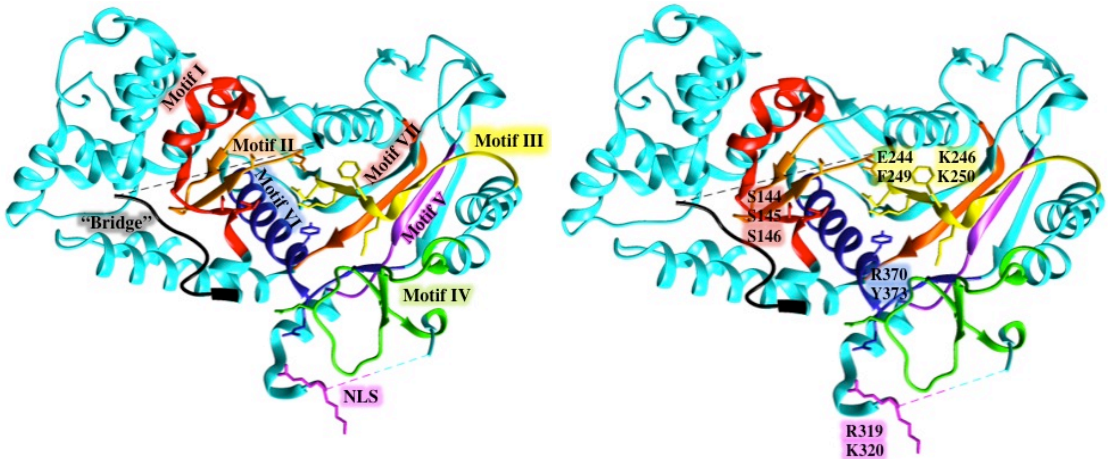
HSO, the rapid and global degradation of host mRNA transcripts is thought to serve two main purposes. These are to promote the evasion of host anti-viral response whilst enabling re-deployment of the transcriptional and translational

machinery for the overproduction of viral mRNA transcripts and proteins (Buisson et al., 2009). Although SOX, an alkaline 5' to 3' exonuclease, has been implicated in this process, it is likely to work in conjunction with other cellular factors (Clyde and Glaunsinger, 2011). Homologues of SOX have been identified in MHV68 (called muSOX) and in EBV (called BGLF5), whose mechanisms of action closely resemble those identified in SOX. It was shown in mice that viral pathogenesis, establishment of latency and reactivation was influenced by the HSO activity of muSOX and vhs (the homologue of SOX in HSV-1) (Richner et al., 2011, Strelow and Leib, 1995, Strelow and Leib, 1996, Strelow et al., 1997). The direct involvement of SOX in mRNA decay was demonstrated when the half life of a reporter green fluorescent protein (GFP) mRNA was reduced in the presence of SOX, compared with experiments where SOX was knocked out in (human embryonic kidney 293 T) HEK293T transfected cells (Glaunsinger et al., 2005).

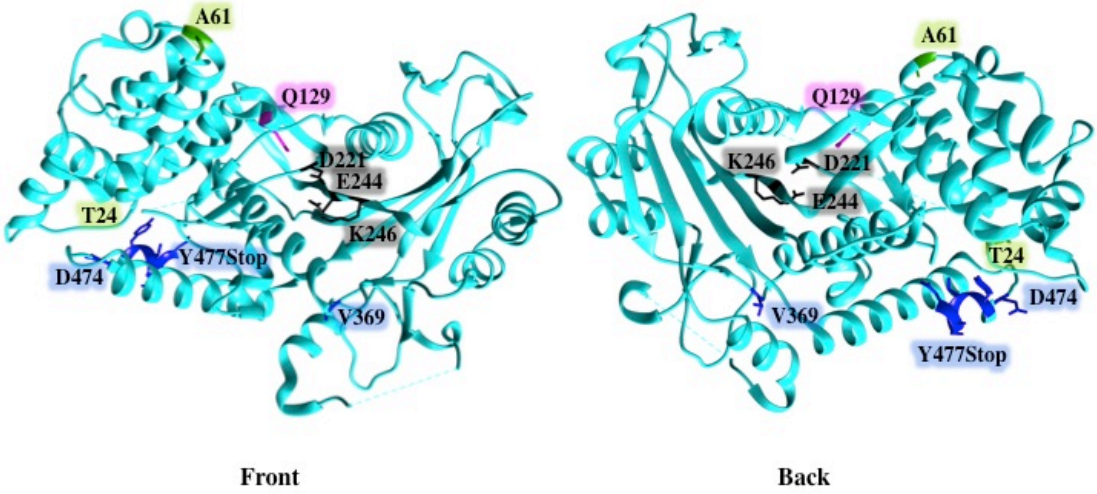
1.1.2. SOX: Structure and Conserved Motifs

SOX and its γ-herpesvirus homologues have 67% identity within a set of seven motifs that are highly conserved. These seven motifs are important for their common nuclease function and have been confirmed by both biochemical and structural studies involving wild-type proteins and a complex involving duplex DNA (Goldstein and Weller, 2004, Glaunsinger and Ganem, 2004, Bagneris et al., 2011). Motif I contains the residues S144, S145 and S146, which together with S219 of motif II form a "serine cluster". The residues within the conserved PD(D/E)XK sequence that has an essential catalytic role are located within motifs II and III. Where the D (D221) and the D/E (E244) and K (K246) residues are found respectively in motifs II and III. Motif III has been implicated in DNA binding along with motif VI (K246, F249, K250 and Y373). Motifs III, IV (E300) and VI (R370) are located in a cleft at the centre of the SOX molecule termed "the canyon" that harbours the active site and DNA binding interface. This canyon effectively subdivides SOX into two lobes (one formed by the N-terminus of the molecule, the other by the C-terminus) that are effectively spanned by a polypeptide "bridge loop" (P164-G180) that resides directly above the active site and is thus thought to be involved in nuclease activity. At the base of the active site canyon, between motifs IV and V, SOX has a second nuclear localization signal (NLS)

A



B



Front **Back**

C

Motif Name	Corresponding Residues in SOX
Motif I	V122 - F148
Motifs II	G206 - D221
Motif III	Y243 - E257
Motif IV	F281 - W311
Motif V	N335 - L345
Motif VI	V365 - I388
Motif VII	I433 - F444
Bridge Loop	Q154 - G180
NLS	P315 - K320

Figure 1-1: SOX Motifs and Residues Involved in Nucleotide Recognition and Nuclease Activity

A) The seven motifs in SOX, NLS and "bridge loop". Motif I in red, motifs II) in orange, motif III in yellow, motif IV in green, motif V in pink, motif VI in blue, motif VII in dark orange, "bridge loop in black and NLS in magenta. B)

The N-terminal (in green T24 and A61) and C-terminal (in blue D474 and Y477Stop) mutants involved in HSO, the catalytic residues involve in deoxyribonuclease (DNase) and HSO activity (in black D221, E244 and K246) and the catalytic DNase mutant (in pink Q129). C) Table clarifying the residues involved each SOX motif.

(315-PRKKRK-320). This NLS is also found in BGLF5 and is highly conserved among the γ -herpesvirus homologues (Glaunsinger et al., 2005, Buisson et al., 2009, Bagneris et al., 2011) (Figure 1-1A).

As SOX was originally shown to function in the resection of newly replicated viral genomes (Buisson et al., 2009) the presence of the PD-(D/E)XK motif resulted in its original classification as a type II restriction DNA endo/exonuclease (Bujnicki and Rychlewski, 2001). RNA endonucleolytic activity, however, has been found in several viral ribonucleases (RNase) that contain this motif (Yuan et al., 2009, Morin et al., 2010). They can harbour within their active site both RNA endonuclease and exonuclease activities (Covarrubias et al., 2011, Yang et al., 2009, Mathy et al., 2007). The apo and holo double stranded deoxyribonucleic acid (dsDNA) bound structures of SOX were obtained and the residues involved in the catalytic site were confirmed as D221, E244 and K246 (Figure 1-1 B) (Protein Data Bank (PDB) ID: 3fhd and 3pov, respectively) (Bagneris et al., 2011, Dahlroth et al., 2009). Furthermore, SOX was confirmed to have RNase activity in addition to its DNase activity. *In vitro* experiments in which D221 and E244 were mutated to S221 and S244 in SOX confirmed that the same catalytic machinery is utilized for both 5' to 3' exonucleolytic RNase and DNase activities in SOX, muSOX and BGLF5 (Buisson et al., 2009, Glaunsinger et al., 2005, Bagneris et al., 2011). Experiments in which these residues were mutated resulted in inhibition of HSO *in vivo* (Covarrubias et al., 2011, Buisson et al., 2009). In addition, a number of SOX non-catalytic residues have been identified that when mutated are HSO defective *in vivo* (HEK 293T cells), but DNase active. These are referred to as HSO mutants. They map to the N-terminus (T24I and A61T), the C-terminus (V369I, D474N and Y477Stop) and the "Bridge Loop" (P176S) of the protein and suggest that RNA binding involves recognition by

distinct structural motifs to those required for DNA (Q129H) (Figure 1-1 B). These differences in modes of association between RNA/DNA and SOX were consistent with studies on BGLF5 (Buisson et al., 2009) and could indicate the involvement of host/viral cofactors interaction to further mediate HSO (Bagneris et al., 2011, Glaunsinger et al., 2005).

1.1.3. SOX: HSO and the Involvement of SOX and Host Co-Factors

Although knockdown of SOX by small interfering RNA (siRNA) silencing is alone sufficient to eliminate HSO (Covarrubias et al., 2011), the extent of its participation in this mechanism remains to be fully established. The HSO mutants nor the D221S catalytic mutant, while leading to abrogation of HSO and RNase activity, did not lead to mislocalisation of SOX within the cell. Experiments in HEK 293T cells with a GFP reporter mRNA demonstrated that SOX decreased the cytoplasmic presence of GFP mRNA, but not the nuclear fraction, while decreasing the overall cellular GFP reporter mRNA quantity. This is an indication that even though SOX contains an NLS and is found in the nucleus, its RNA degradation activity is likely to be occurring in the cytoplasm, where host mRNA degradation takes place (Glaunsinger et al., 2005).

The degradation of host mRNA is uninhibited by the presence of the mRNA 5' cap (Covarrubias et al., 2011). In addition SOX was shown to specifically degrade RNA transcribed by the RNA polymerase II (pol II), which is responsible for transcription of mRNA and most small nuclear ribonucleoproteins (snRNA) and microRNA (miRNA) (Covarrubias et al., 2011). This was demonstrated by constructing GFP reporters that were pol I or pol III transcribable while expressing SOX, which did not lead to GFP reporter RNA turnover (Covarrubias et al., 2011). Pol I and pol III transcription does not result in the addition of the 5' cap and polyadenylation tail (poly(A)tail), nor does the RNA transcript go through the same mRNA maturation pathways, e.g. nuclear splicing. In addition, it was shown that these mRNA transcripts needed to be translationally competent, as they co-sedimented with the ribonucleoprotein (RNP), the eukaryotic small ribosomal subunit (40S) pre-initiation complex (eIF3j, eIF2a and RPS3) and eukaryotic

ribosome (80S) complex. Further analysis also revealed a decreased polysome population and increased 80S monosome population, in response to SOX over expression, which is consistent with degradation of actively translating mRNAs (Covarrubias et al., 2011). Recently Hendrickson *et al*. (2009) reported that the bulk of cytoplasmic mRNAs are polysome-associated, suggesting that targeting translationally active mRNAs would allow the virus to target the majority of host mRNA for degradation as observed in HSO (Hendrickson et al., 2009). The link between translation and degradation is reminiscent of host mRNA surveillance mechanisms, such as nonsense mediated decay (NMD), which works with the exosome and exoribonuclease 1 (Xrn1).

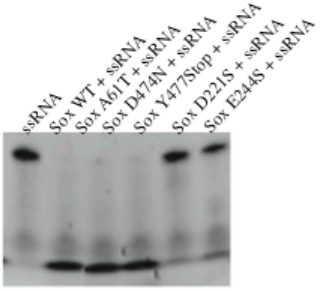
A depletion study of host RNA degradation enzymes revealed that the 5' to 3' exonuclease Xrn1 was necessary for complete mRNAs degradation in SOX-expressing cells. Further, yeast two hybrid assays carried out by collaborators (Ebrahimi, B.; unpublished work) indicated that human Xrn1 and SOX were interacting. In the absence of Xrn1, the mRNAs were only partially degraded and the degradation intermediates of the SOX mRNA targets accumulated (Covarrubias et al., 2011, Kronstad and Glaunsinger, 2012). Xrn1 was also shown to co-sediment in both the RNP and 40S fractions, akin to SOX. It was suggested that SOX's RNase activity could not account for the rapid and global SOX-induced mRNA degradation owing to its poor affinity for single stranded RNA (ssRNA) substrates (Bagneris et al, 2011). Additional experiments with siRNA demonstrated the reliance of the SOX-induced HSO degradation pathway on Xrn1 (Covarrubias et al., 2011). Xrn1, which is highly conserved in eukaryotes, is an 5'-3' exoribonuclease that functions to degrade cytoplasmically localised mRNAs, as part of the host mRNA surveillance pathways (Garneau et al., 2007). Xrn1's 5' to 3' exonuclease activity generally requires initial deadenylation and decapping of mRNAs, but notably, this rate-limiting step appears not to be required for SOX-induced turnover (Garneau et al., 2007). One way to make mRNAs accessible to Xrn1 turnover, prior to or without deadenylation or decapping, is via endonucleolytic cleavage, which is similar to mechanisms in the host mRNA surveillance pathways. Intriguingly, recent evidence suggests that this mechanism of viral nuclease cleavage followed by Xrn1 mediated degradation appears to be

ubiquitous amongst the *herpesviridae* and *coronaviridae* (Gaglia et al., 2012, Kronstad and Glaunsinger, 2012).

1.1.4. SOX: SOX-Mediated mRNA Decay and Target Sequences

Previous published experiments from the Barrett Group that focused on the ability of SOX to turnover single stranded RNA showed that the HSO mutants did not abrogate ssRNA degradation (Figure 1-2) (Bagneris et al., 2011). This was in contradiction to the previously published *in vivo* work, which showed that HSO mutants did abrogate GFP-mRNA turnover (Glaunsinger et al., 2005). This was suggestive of the need for either a specific RNA sequence element, a structured motif or a co-factor.

A



B

dsDNA	5′-GGGGATCCTCCCAGTCGACC-3′
	FAM-3′-CCCCTAGGAGGATCAGCTGG-5′
dsDNA-5′P	P-5′-GGGGATCCTCCCAGTCGACC-3′-FL
	3′-CCCCTAGGAGGATCAGCTGG-5′
dsRNA-5′P	P-5′-UGGUUUACAUGUUCCAAUAUU-3′-FL
	3′-ACCAAAUGUACAAGGUUAUAA-5′
ssRNA-5′P	P-5′-UGGUUUACAUGUUCCAAUAUU-3′-FL

Figure 1-2: HSO SOX Mutants do not Abrogate ssRNA Turnover

A) The TBE-Urea gel from the Bagneris et al., 2011 publication showing that the HSO mutants (A61T, D474N and Y477Stop) do not abrogate the turnover of ssRNA in vitro. B) Oligonucleotides used in assays in Bagneris et al., 2011.

As previously mentioned, in the absence of Xrn1, mRNAs were only partially degraded and shorter RNA degradation intermediates were seen to accumulate (Covarrubias et al., 2011, Kronstad and Glaunsinger, 2012). In experiments with the three reporter mRNAs GFP, red fluorescent protein (DsRed2) and β-globin (HBB), frequently used in these studies, the accumulation of degradation intermediates of

defined length was also observed (See Appendix A). These intermediates are reminiscent of the morphogenetic effect on genitalia 6 (SMG6)-cleavage products seen during NMD (Eberle AB et al., 2009, Kashima et al., 2010). The GFP and DsRed2 reporter mRNAs are of similar length (1.2-1.5 kb), whilst the lengths of their degradation intermediates varied (GFP fragment ~ 1.1 kb and DsRed2 fragment ~ 600 base pairs (bp)). Furthermore, the degradation intermediates for GAPDH (glyceraldehyde 3-phosphate dehydrogenase) and HBB were similarly varied (Covarrubias et al., 2011). This indicates that the generation of the intermediates is not controlled by a positional cue around or by the translation initiation site. The fact that the degradation intermediates for each reporter gene have a consistent defined-length was indicative of cleavage being directed to a specific location or sequence within the mRNA (Covarrubias et al., 2011). The Glaunsinger group sequenced the degradation intermediates for the three reporter mRNAs. All three mRNAs had only a UGAAG sequence in common 2-3 nucleotides upstream of the cleavage site (Covarrubias et al., 2011). Interestingly, transposition of the 201-nucleotide long sequence containing the UGAAG was sufficient to prevent cleavage and subsequent degradation by Xrn1 in cells transfected with SOX and GFP-containing vectors. However, when a 25 nucleotide construct encompassing the UGAAG motif was used, this proved to be insufficient (Covarrubias et al., 2011, Clyde and Glaunsinger, 2011). These results thus suggest a structural element to the SOX mRNA recognition process in conjunction with a requirement for the conserved motif. In addition, it was suggested that HSO is a two-step mechanism requiring initial cleavage of mRNAs by SOX to catalyse substrates for Xrn1. It would thus appear that SOX subverts the normal functioning of Xrn1 since this process occurs in the absence of deadenylation and decapping.

1.2. mRNA and mRNA Decay

Large-scale studies and analyses indicate that as many as half of all changes in the transcriptome and proteome can be attributed to altered rates of mRNA decay (Garneau et al., 2007). From transcription, 5'capping, splicing, polyadenylation, mRNA export to translation, at every point a transcript is subject to elaborate control leading to varied protein expression and altered signals. Thus, many cellular mechanisms and proteins are solely dedicated to tune the rate of mRNA degradation

(Jinek et al., 2011, Schoenberg, 2011). It is becoming increasingly apparent that these finely tuned mechanisms are the targets for pathogenic viruses, which through their subversion are able to promote replication of the viral genome through re-deployment of the host cell translational mechanism. These host mechanisms will therefore be reviewed along with their recently discovered involvement in the pathogenicity of the KSHV virus, the focus of this thesis.

1.2.1. Host pre-mRNA

As previously mentioned, eukaryotic mRNA undergoes maturation, including precursor mRNA (pre-mRNA) splicing within the nuclei. The first maturation step is the addition of a 7-methylguanosine (m7G) via a 5'-5' triphosphate linkage to the first nucleotide of the transcript, which cannot be easily degraded by 5' to 3' exonucleases. This 5' cap is thus a protective element that also functions as a recognition motif for proteins involved in nuclear export, 5' proximal intron excision, translation initiation and translation (Shatkin and Manley, 2000). The poly(A)tail is added to the maturing mRNA at the end of transcription. The poly(A)tail also protects the mRNA from cytoplasmic mRNA degradation and is involved in mRNA export and translation mechanisms (Kapp and Lorsch, 2004, Konarska et al., 1984). mRNA splicing and transport across the nuclear pore complex (NPC), are intrinsically linked and subject to quality control (Houseley et al., 2006, Fasken and Corbett, 2005, Dimaano and Ullman, 2004, Saguez et al., 2005, Vinciguerra and Stutz, 2004). The vast majority of human genes contain introns and express more than one mRNA by a process called alternative splicing, which is highly regulated (Maquat, 2004, Kan et al., 2001). This process allows the translation of functionally diverse protein isoforms. Splicing has to be precise and is a complex process, but has the advantage of diversifying the proteome (Black, 2003, Hastings and Krainer, 2001, Burge et al., 1999). When splicing is inefficient or inaccurate, this can lead to a shift in the translational reading frame, which then introduces a premature termination codon (PTC). mRNA transcripts that contain a PTC are detected by cellular control mechanisms and are targeted for NMD.

1.2.1.1. mRNA Splicing

The average human gene contains 9 exons and 8 introns (Sakharkar et al., 2004). An exon is the nucleotide sequence of the pre-mRNA that remains present within the final mature RNA product, while the intron is removed from the mRNA. The exons are defined by short classical splice-site sequences at the intron/exon borders, which are GU at the 5' splice site and AG at the 3' splice site (Figure 1-3) (Black, 2003, Hastings and Krainer, 2001). The 3' splice site can be further defined by an upstream polypyrimidine tract (PPT), which recruits factors to the 3' splice site and to the A branch site sequence (BPS). snRNA bind the splice-site sequences and promote the assembly of the spliceosome, a large ribonucleoprotein complex. This ribonucleoprotein complex is made up of five small nuclear ribonucleoproteins (snRNP) and another 100 other proteins (Maquat, 2004) .

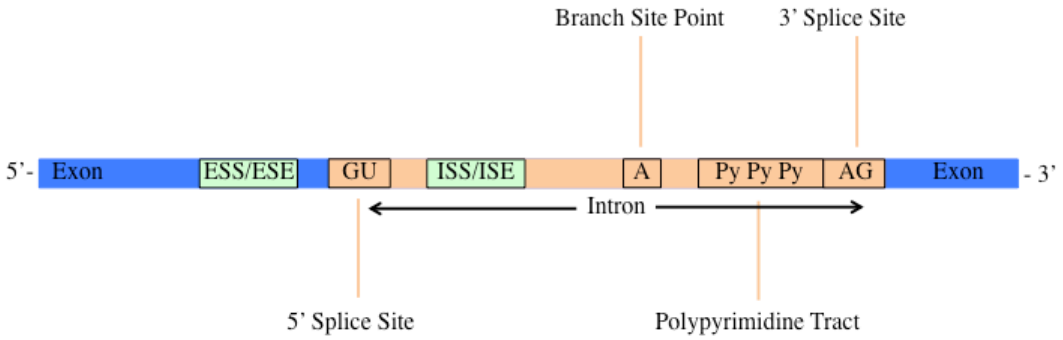


Figure 1-3: pre-mRNA and Splicing Signals

pre-mRNAs contain exons and introns, the latter are excised so that the exons on each side of the intron regions form an mRNA. To facilitate excision of the intron, the splice site contains a 5' GU dinucleotide, an A branch site, a polypyrimidine tract and a 3' splice site. Additional cis elements can be found in the exons, such as exonic splicing enhancer and silencer (ESE and ESS) and intronic splicing enhancer and silencers (ISE and ISS).

The spliceosome firstly recognizes the intron/exon boundaries and secondly catalyses the excision reaction, removing the introns and joining exons. Each snRNP contains a single uridine rich snRNA. The U1 snRNP binds the 5' splice site, while U2 snRNP binds the branch site (Liu, 2002, Du and Rosbash, 2002). Binding occurs via RNA-RNA interaction between the snRNA and the pre-mRNA

(Faustino and Cooper, 2003). Following recognition, pre-mRNA is bent to bind the three other RNPs to form the spliceosome. The final spliceosome complex undergoes a conformational change to cleave the RNA at the 5' GU splice site and forms a lariat at the A branch site. The intron is then cleaved out at the 3' AG splice site, where the two exons ligate together (Burge et al., 1999, Liu, 2002, Lallena et al., 2002, Du and Rosbash, 2002).

The short and degenerate splice sites are not sufficient for splice-site recognition (Lim and Burge, 2001) and must be distinguished from pseudo splice site sequences that resemble classical splice sites but are never used (Black, 2003). Additional *cis* elements, such as exonic and intronic splicing enhancers (ESEs and ISEs) and exonic and intronic splicing silencers (ESSs and ISSs), build a network of interactions across exons as well as across introns to allow the correct exon recognition and hence accurate splicing (Berget, 1995, Reed, 1996).

During the second splicing step the exon junction complex (EJC) is deposited 20-24 nucleotides upstream of each exon-exon junction, when the lariat has formed and the exons are ligated together (Le Hir et al., 2000, Shibuya et al., 2004). The EJC remains bound to the mRNP, during nuclear export and in the cytoplasm. Proteins bind or get released from the EJC during the transport. The EJC has major influences on surveillance and localization of the spliced mRNA and leads to translation enhancement (Tange et al., 2004). Recent sequencing research by various groups have identified purine-rich sequences flanking the EJC binding sites, which showed a high GAAGA content and is thought to be a binding site of EJC associated factors (Long and Caceres, 2009, Singh et al., 2012, Saulière et al., 2012). This motif resembles known binding sites for several serine-arginine–rich proteins (SR proteins), serine/arginine-rich splicing factor 1 (SRSF1) (Sanford et al., 2009). SR proteins bind with very high specificity to purine-rich sequences in RNA and are important for both constitutive and alternative splicing (Long and Caceres, 2009). In other studies, purine-rich GAAG repeats were identified as exonic splicing regulatory elements in plants and animals and as exonic splicing enhancers in other vertebrates (Chasin, 2007, Tacke and Manley, 1995, Thomas et al., 2012, Pertea et al., 2007). SRs also function as exonic splicing regulatory proteins, such as in alterative splicing.

1.2.1.2. Alternative Splicing

Alternative splicing allows one gene to express multiple mRNAs that encode proteins with diverse and even antagonistic functions, by the joining of different 5' and 3' splice sites thereby skipping exons and thus e.g. a potential regulatory domain of a protein. Alternative splicing can also lead to altered translation, stability and/or localization of the mRNA by removing or inserting regulatory *cis* elements. Alternative splicing itself is regulated by ESEs and ISEs and ESSs and ISSs (Grabowski, 1998, Smith and Valcarcel, 2000). It is cell specific and can lead to mis-spliced transcripts with deleterious effects. These deleterious transcripts are detected and processed by the cellular posttranscriptional quality control mechanism, NMD (Danckwardt et al., 2002).

1.2.1.3. mRNA Nuclear Export

All mRNA needs to pass through the NPCs to exit the nucleus to the cytoplasm, where they can be translated. mRNAs pass through the NPC as large RNPs complexes. These become export-competent by associating with the export adapter transcription-coupled export (TREX) complex and the export receptor tandem affinity purification protein (TAP), which targets the mRNA to the NPC (Soller, 2006, Schumann et al., 2013).

As mRNAs vary greatly in sequence, length and structure, their recognition by the TREX complex must target common features of the host mRNAs. In eukaryotes, mRNA export is intrinsically linked to mRNA maturation, i.e. 5'capping and splicing, leading to the stepwise assembly of the export-competent mRNP. The TREX complex gets localized to the 5' end of the mRNA in a splicing dependent manner by Aly/REF export factor (Aly) and THO, two components of the TREX complex, which bind directly to the cap binding protein 80 (CBP80), which in turn is part of the cap-binding complex (CBC) (Cheng et al., 2006, Lejeune et al., 2002, Chi et al., 2013). It has been shown that Aly, together with the Asp-Glu-Ala-Asp box helicase (DEAD-box helicase) UAP56, interact with the EJC, which is known to influence translation, surveillance and localization of the spliced mRNA (Chang et al., 2007, Giorgi and Moore, 2007, Nott et al., 2004). Thus both capping and splicing are crucial for formation of the TREX complex

leading to nuclear export of the host mRNA (Zhou et al., 2000, Masuda et al., 2005). Finally to be considered export competent by the NPC, the TREX containing mRNP needs to be "handed over" from Aly to TAP, the export receptor (Jackson et al., 2012, Stewart, 2007) (Figure 1-4).

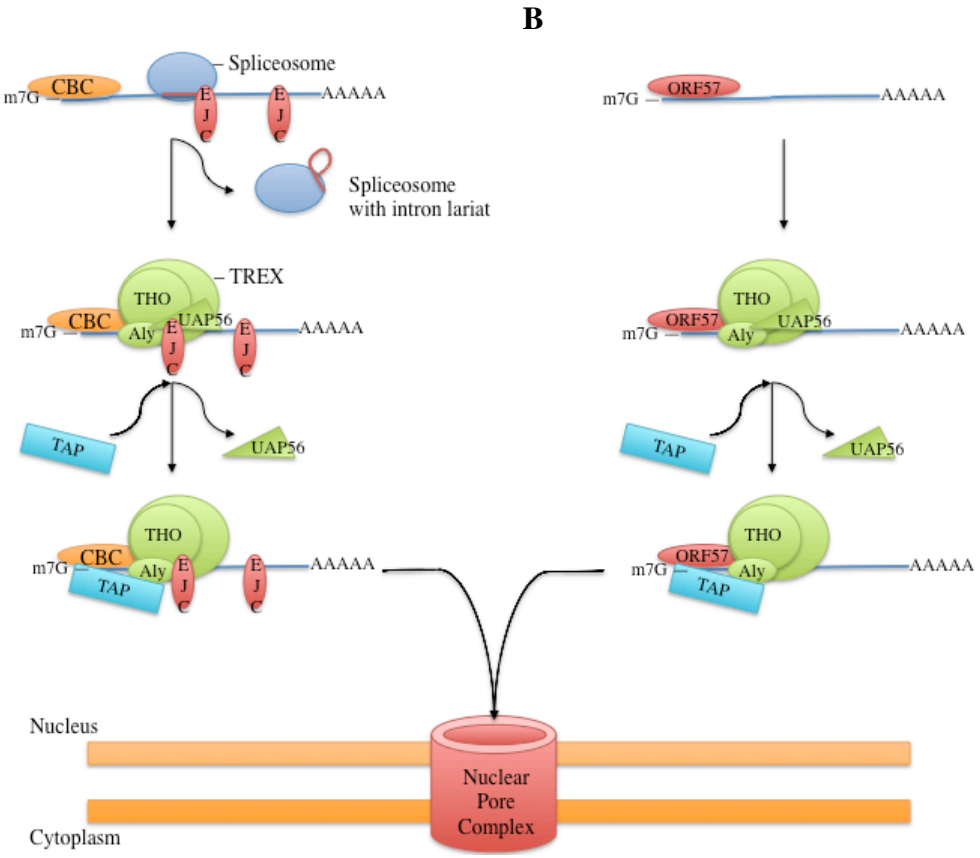


Figure 1-4: RNA Processing and Nuclear Transport for Host and Viral mRNA

A) On the left the host mRNA processing and export is represented. Here EJC is recruited onto the mRNA at the vicinity of the introns, to be spliced by the spliceosome, with whom it interacts. Once the introns are removed the spliceosomes release leaving the EJCs at the exon junctions. The TREX complex then binds EJC and the 5' CBC. It is the ability of TREX to bind EJC that allows it to recognize spliced mRNAs and thus makes these exportable. The binding of TAP to Aly and THO induces a conformational change leading to the release of UAP56 and to the translocation through the NPC. B) On the right KSHV mRNA is intronless and thus is not bound by EJC. ORF57 binds the KSHV mRNA and recruits the TREX complex the mRNA, allowing the viral mRNA to export.

1.2.2. KSHV mRNA

To achieve effective viral proliferation, gene expression and viral replication are intimately linked and tightly regulated. This is achieved by timely viral gene expression with 4 stages of transcription; latent, immediate early lytic, early lytic, and late lytic gene expression. To achieve a high level of replication the virus takes over the host gene expression machinery during lytic infection. The KSHV transcripts are translated in a cap-dependent manner (Conrad and Steitz, 2005) and most KSHV mRNAs are intronless. The host machinery by default tends to express intronless genes at much lower levels than intron-containing mRNAs (Conrad and Steitz, 2005). In host and KSHV and other viruses, intronless genes often contain *cis*-acting elements, which allow enhanced gene expression despite the missing introns (Donello et al., 1998, Huang and S., 1995, Huang and Liang, 1993).

To overcome this, KSHV also has the immediate-early protein open reading frame 57 (mRNA export factor ICP27 homologue) (ORF57) (Boyne et al., 2010, Jackson et al., 2011), which is a *trans*-acting regulatory protein that enhances expression of intronless viral genes and allows the export of viral intronless mRNAs (Malik and Schirmer, 2006). Interestingly the ORF57 gene itself is monocistronic and contains one intron. This allows ORF57 to be efficiently expressed and processed prior to its action to allow viral early and late lytic intronless transcripts to be transcribed and translated (Jackson et al., 2012).

As previously mentioned, host mRNA transcription, post transcription modification, nuclear export, translation, localization, protein stability and the quality control mechanisms are intrinsically linked (Luo and Reed, 1999, Valencia et al., 2008). The recruitment of TREX to EJC, that allows nuclear export of mRNAs for translation, is splice-dependent (Masuda et al., 2005, Schumann et al., 2013), which in turn explains the preference for intron-containing transcripts (Nott et al., 2004). This poses a significant stumbling block for KSHV lytic intronless mRNAs as they do not undergo splicing and therefore cannot recruit TREX via the splicing-dependent mechanism (Schumann et al., 2013). To overcome this, KSHV uses ORF57 (Luo and Reed, 1999, Valencia et al., 2008) allowing the viral intronless transcripts to bind the TREX complex (Tunnicliffe et al., 2010, Boyne et

al., 2008). ORF57 interacts directly with the RNA and with the export adapter protein Aly, recruiting it to the 5′ end of the RNA. Aly is normally recruited in a splicing-dependent manner. ORF57 in conjunction with Aly then recruits the remaining TREX complex allowing the viral mRNAs to translocate through the nuclear pore in a TREX RNP complex (Malik et al., 2004, Boyne and Whitehouse, 2009, Boyne and Whitehouse, 2006, Jackson et al., 2012). KSHV ORF57 mediated mRNA export appears to be highly complex and has yet to be fully understood (Malik et al., 2004, Jackson et al., 2012, Stutz and Izaurralde, 2003). But Jackson et al. were able to determine that ORF57 recruits only TREX and not EJC to the viral intronless mRNAs, and that ORF57 is able to enhance translation of viral transcripts, thus overcoming the lack of translation enhancement by EJC (Jackson et al., 2012).

Finally ORF57 regulation and disruption of viral and host RNA processing, leading to the export of the viral intronless genes (Hardy and Sandri-Goldin, 1994) may also contribute to HSO together with SOX to give the virus a kinetic edge over the host cell in gene expression (Hardy and Sandri-Goldin, 1994, Whitehouse et al., 1998, Ruvolo et al., 1998, Jackson et al., 2012).

1.2.3. mRNA Decay

1.2.3.1. mRNA Decay in Host Homeostasis and Antiviral and Proviral Response

mRNA stability is assured by the 3’ poly(A) tail and the 5’ cap, each associate with protein stability factors, the poly(A)-binding protein (PABP) and eIF4E, respectively (Schoenberg, 2011, Garneau et al., 2007). These two protective systems at the ends of mRNA tend to inhibit exonucleolytic decay. Thus there are two main directions in which mRNA degradation can be initiated: a) a 3’ to 5’ decay pathway where the 3’ end is deadenylated, followed by 3’ to 5’ degradation by the exosome, a large complex of 3’ to 5’ exonuleases and b) 5’- 3’ decay pathway exists in which the 5’ cap is cleaved followed by degradation by the 5’ to 3’ exonuclease Xrn1. After decapping and/or deadenylation, degradation can occur in a 3’ to 5’ or 5’ to 3’ direction. Another alternative to exonucleolytic degradation

of mRNA is endonucleolytic decay in which an endonuclease would need to cleave internally, followed by exonucleolytic degradation (Garneau et al., 2007).

Different mRNAs within the same cell have distinct lifetimes (stabilities). In mammalian cells, mRNA lifetimes range from several minutes to days. The limited lifetime of mRNA enables a cell to alter protein synthesis rapidly in response to its changing needs. There are many mechanisms that lead to the destruction of an mRNA or sequestering of the mRNA to stress granules. Besides interfering with cellular mRNA trafficking, viruses can stimulate mRNA turnover (Walsh and Mohr, 2011) and host mRNA decay pathways, which had been thought to serve primarily in an antiviral manner.

1.2.3.2. 3' to 5' mRNA Decay – Deadenylation and Exosome

Nearly all mammalian mRNA decay is initiated by poly(A)-tail shortening; deadenylation (Yamashita et al., 2005, Chen et al., 2002, Zheng et al., 2008). This includes decay mediated by adenylate-uridylate-rich elements (ARE) in the 3' untranslated region (3'-UTR), by destabilizing elements in the protein coding regions and NMD. Deadenylation predominantly leads to decapping by the nucleases RNA-decapping enzyme 1- mRNA-decapping enzyme 2 (Dcp1-Dcp2) at the 5' end followed by 5' to 3' exonucleolytic degradation by the main cytoplasmic 5' to 3' exonuclease Xrn1 (Parker and Song, 2004, Beelman and Parker, 1995, Coller and Parker, 2004, Cougot et al., 2004). Alternatively, deadenylation can be followed by 3' to 5' exonucleolytic degradation by the exosome in a 3' to 5' fashion (Allmang et al., 1999, Mitchell et al., 1997).

Deadenylation is biphasic (Yamashita et al., 2005). C-C chemokine receptor type 4 NOT (CCR4–NOT), PAB-dependent poly(A)-specific ribonuclease subunit 2 -PAB-dependent poly(A)-specific ribonuclease subunit 3 (PAN2–PAN3) and poly(A)-specific ribonuclease (PARN) are characteristic eukaryotic deadenylases, each with unique properties. PAN2–PAN3 activity is not inhibited by the PABPs, while CCR4–NOT and PARN activity is inhibited by it (Chen and Shyu, 2010). In the first phase PAN2–PAN3 carries out the initial shortening of the poly(A)tail to a length of 100-80 nucleotides (Mangus et al., 2003). In the second phase the

deadenylation of the now shortened and PABP free poly(A)tail is presented to another deadenylase, CCR4–NOT or PARN. Deadenylation by CCR4-NOT is coincident with 5'decapping and 5' to 3' exonucleolytic decay by Xrn1. PARN's deadenylase activity is enhanced by the presence of the 5' cap on the mRNA, but

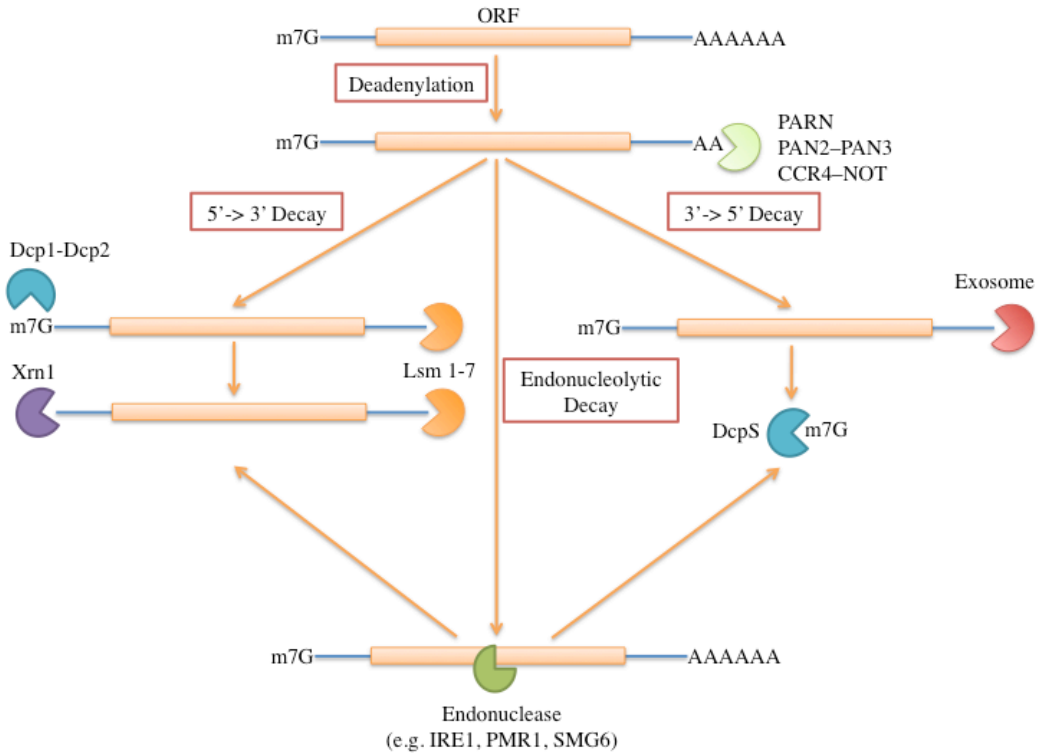


Figure 1-5: mRNA Decay Mechanisms

Most mRNA decay is deadenylation-dependent, where the poly(A)-tail is removed by a deadenylase activity (CCR4–NOT, PAN2-PAN3 or PARN). This is followed by two degradation mechanisms: one is the 5′→3′ decay and the other is 3′→5′ decay. In the first mechanism the mRNA is bound by the Lsm1–7 complex inducing decapping by Dsp1–Dsp2 rendering it susceptible to decay by the 5′→3′ exoribonuclease Xrn1. In the second mechanism the mRNA is degraded in the 3′→5′ direction by the exosome, where DcpS hydrolyses the remaining 5' cap. Another way to initiate mRNA decay is endonucleolytic cleavage by an endonuclease (e.g. IRE1, PRM1 or SMG6), which generates two fragments that a susceptible degradation Xrn1 and the exosome, while still bearing the protective 5'cap and poly(A)-tail (Garneau et al., 2007).

inhibited by nuclear CBC (Brawerman, 1981, Wilusz et al., 2001, Wickens et al., 1997). Enhanced deadenylase activity has been detected for mRNAs that are subject to NMD and contain destabilizing ARE (Wilson and Treisman, 1988, Shyu et al., 1991, Lykke-Andersen and Wagner, 2005) (Figure 1-5).

After deadenylation the mRNA gets 3' to 5' degraded by the exosome, a complex of 10-12 subunits. Each of the core exosome subunits has an RNase PH domain, which is thought to either contribute to catalytic activity or play a role in substrate recognition and placement (Mitchell et al., 1997, Houseley et al., 2006). If the 5' cap is still present after 3' to 5' decay, it is metabolized by the scavenger mRNA-decapping enzyme (DcpS) (Liu et al., 2002).

1.2.4. 5' to 3' mRNA Decay – Decapping and Xrn1

Although 5' to 3' mRNA decay is initiated by the 5' cap removal by the decapping enzymes Dcp1-Dcp2, several accessory factors are required for efficient decapping, e.g. the Lsm1-7 complex, which binds the deadenylated 3' end of the mRNAs promotes decapping (Wilusz et al., 2001, Cougot et al., 2004, Tharun and Parker, 2001). Decapping is then followed by 5' to 3' degradation by the exonuclease Xrn1.

1.2.4.1. Xrn1

The 5' to 3' exonuclease Xrn1 is conserved across eukaryotes and it is involved in RNA transcription, metabolism and interference. Xrn1 (ca. 175 kilo Dalton (kDa)) is the main cytoplasmic RNase, involved in the degradation of decapped mRNAs, NMD and miRNA decay. These take place in the cytoplasm (Parker and Sheth, 2007), where the enzymes Dcp1-Dcp2 decap the mRNA. This generates a 5' monophosphorylated RNA intermediate (Jinek et al., 2011), which irreversibly commits the mRNA for degradation, as Xrn1 recognizes specifically 5' monophosphorylated RNA which is rapidly degraded to mononucleotides without partially degraded intermediates. As a result, Xrn1 plays a central role in the controlled turnover of the mRNA transcriptome (Chang et al., 2011). Xrn1's activity normally requires prior deadenylation and decapping of mRNAs, which are

rate-limiting steps of normal decay, except when endonucleolytic cleavage occurs. Xrn1 also participates in degrading RNA intermediates generated by endonucleolytic mRNA cleavage. Xrn1 has been shown to have an antiviral activity by virtue of its exonuclease activity and to act as a potent suppressor of viral RNA recombination in viruses, such as tomato bushy stunt virus. However, in flaviviruses, Xrn1 is subverted into producing a subgenomic flavivirus RNA, which is essential for viral cytopathogenicity in cells and pathogenicity in mice (Silva et al., 2010). This appears to be facilitated by a highly stable pseudoknot in the viral RNA that stalls Xrn1 and suggests that Xrn1 paradoxically may have important pro and anti-viral roles.

1.2.5. Endonucleolytic Decay

Endonucleolytic cleavage is an efficient mean of destroying mRNAs as it produces two fragments that are susceptible to 5' and 3' exonuleases; e.g. Xrn1 and the exosome. Endonucleolytic cleavage is involved in NMD in mammals, where PTC recognition leads to endonucleolytic cleavage in the vicinity of the aberrant stop codon and induces accelerated deadenylation (Eberle AB et al., 2009, Cao and Parker, 2003, Chen and Shyu, 2003). Certain cellular endonucleases that target mRNA have been characterized, such as SMG6, serine/threonine-protein kinase/endoribonuclease (IRE1) and polysomal ribonuclease 1 (PMR1). As cellular endonucleases are very potent they are highly regulated and/or specific (Hollien and Weissman, 2006, Yang et al., 2004).

IRE1 targets actively translating mRNAs in the endoplasmic reticulum during unfolded protein stress response. IRE1 endonuclease activity was shown to catalyse splicing of the X-box-binding protein 1 (XBP1) mRNA (Hollien and Weissman, 2006, Yoshida et al., 2001).

PMR1 endonucleolytically cleaves actively translating mRNAs on polysomes (Yang and Schoenberg, 2004). The cleavage site on the targeted mRNAs was between the UG dinucleotides of two overlapping repeats of AYUGA, which were found in the loop region of a stem-loop structural element (Chernokalskaya et al., 1997). Mutations to either element did not abrogate endonucleolytic cleavage

providing the bases remained unpaired in the loop. When the sequence was mutated so that these bases became paired to prolong (elongate) the stem, endonucleolytic cleavage was abrogated. This showed that endonucleolytic cleavage in this instance not only requires a specific sequence, but that this sequence has to be in the right structural context (Chernokalskaya et al., 1997). Further research by Brock and Shapiro found that the 3' UTR of the vitellogenin mRNA contained PMR1 cleavage sites, but this transcript was protected from PMR1 endonucleolytic cleavage in the presence of PMR1 (Brock and Shapiro, 1983). The protein vigilin was demonstrated to bind the region of the PMR1 cleavage sites in the 3'UTR of the vitellogenin mRNA thereby blocking the PMR1 cleavage site from being bound and cleaved by PMR1, thus stabilizing the vitellogenin mRNA (Dodson and Shapiro, 1997, Cunningham et al., 2000).

1.2.5.1. Nonsense Mediated Decay and Links to Splicing and Translation

At each step during transcription and maturation of mRNA errors can be introduced into the transcript. To protect the cell from potential deregulation and toxic protein products surveillance mechanisms have evolved that couple translation to degradation pathways, such as non stop decay (NSD), no go decay (NGD) and NMD. These mechanisms occur in the nucleus, while most discovered degradation pathways are translation dependent and occur in the cytoplasm. NMD is the most studied of these pathways (Gebauer and Hentze, 2004, Moore, 2005).

NMD was demonstrated to be crucial in embryogenesis and embryonic viability (McIlwain et al., 2010, Hwang and Maquat, 2011). Hence NMD not only functions in the surveillance of deleterious transcripts, but also plays an important role in the regulation of normal gene expression by degradation (Gardner, 2010). It is thought that alternative splicing can regulate gene expression by targeting mRNAs for NMD (Gardner, 2010). Thus there are different ways in which NMD can be initiated and thus function to facilitate degradation. In mammalian cells, one well-defined NMD is EJC-dependent NMD, which is discussed below.

NMD detects and degrades mRNA transcripts that contain PTCs, which can

be introduced by mutations, inefficient splicing and leaky translation initiation. Core to the NMD complex are the proteins up-frameshift factor 1 (UPF1), UPF2 and UPF3, which are highly conserved (Conti and Izaurralde, 2005). UPF1 is recruited by the release factors to stalled ribosomes, where the transient SURF complex (SMG1 -UPF1-eRF1-eRF3) forms. UPF1 is required for all known NMD pathways, whereas the other components may vary dependent on the NMD substrates and the downstream pathways. The SURF complex associates with EJC through UPF2 leading to assembly of an UPF1-UPF2-UPF3 surveillance complex, allowing initiation of NMD, once UPF1 is phosphorylated (Conti and Izaurralde, 2005, Amrani et al., 2006, Lejeune and Maquat, 2005, Behm-Ansmant et al., 2007). The EJCs are deposited 20-24 nucleotides upstream of every exon junction and thus EJCs are the markers of splicing (Le Hir et al., 2000). Most introns are found in the protein coding region where the EJCs are displaced by the ribosomes during translation. However, when an mRNA transcript is mis-spliced leading to PTCs, EJC remains located 50-55 or more nucleotides downstream from the PTC site, which facilitates recognition of the EJC by the SURF complex. PTC increases the distance between the terminating ribosome to the poly(A) tail, this resultant mRNP conformation is also recognized as abnormal triggering NMD (Gardner, 2010, Gatfield et al., 2003). SMG5, SMG6 and SMG7 recognize phosphorylated UPF1 and are thought to be the link to the mRNA degradation machinery.

In human cells, SMG6 as part of the NMD process, catalyses endonuclease cleavage of PTC-containing mRNA (Eberle AB et al., 2009), followed by exonucleolytic decay, by e.g. Xrn1, of the resultant fragments (Doma and R., 2006). It had thus been now hypothesized that mammalian NMD may be initiated by SMG6-mediated endonucleolytic cleavage (Eberle AB et al., 2009). The endonuclease cleavage sites of SMG6 were found both upstream and downstream of the PTC on the target mRNAs. The molecular basis for the clustering of cleavage sites around the PTC is explained by the specific binding of SMG6 to the EJC (Kashima et al., 2010).

1.2.7. AU-rich Elements in mRNA Stability and Decay

AREs are found in the 3' UTR of many protooncogenes, nuclear transcription factors and cytokines mRNAs. And AREs play a role in gene transcription regulation in neurons and the immune system (Chen and Shyu, 1995, Barreau et al., 2006). They are one of the most common determinants of RNA stability in cells, which is usually determined by regulation of degradation of the transcript (Chen and Shyu, 1995).

Three classes of AREs have been identified with different canonical sequences. The most are characterized by the canonical ARE, containing the core AUUUA ARE motif within a AU rich sequences; WWWU**AUUUA**UUUW (Khabar, 2005). AREs are highly variable and the neighbouring sequences influence AREs effect on mRNA stability (Chen and Shyu, 1995).

AREs are bound by a number of proteins, such as human antigen R (HuR) and AU-rich element RNA-binding protein 1 (AUF1) (Gratacós and Brewer, 2010). AUF1 affects changes in ARE mRNA degradation rates by binding to ARE in a complex with proteins, such as HuR, which is known to stabilize mRNAs and gene expression (Gratacós and Brewer, 2010). Since ARE binding proteins can interact directly or indirectly with mRNA decay machinery, recognizing AREs can lead to enhanced decay of ARE containing transcripts, e.g. AUF1 interacts with the exosome (Chen et al., 2001). ARE binding proteins that stabilize mRNA have been shown to work by two means. The first is the removal of the mRNA from the sites of decay by competing for the binding site of destabilizing factors, which link to the decay machinery. The second is to strengthen the PABP poly(A) interaction thereby inhibiting the deadenylation dependent mRNA decay machinery. HuR has been shown to compete for binding sites with the destabilizing proteins, such as AUF1 (Lal et al., 2004).

1.2.7.1. IL-6 mRNA Transcript

It has long been known that the interleukin-6 (IL-6) mRNA transcript evades turnover in KSHV infected cells and that KSHV also has its own viral version of IL-6; vIL-6 (Hutin et al., 2013, Rezaee et al., 2006). IL-6 is a cytokine that acts both in a pro-inflammatory and an anti-inflammatory way (Moonga et al.,

2002). It can be secreted, bind intracellular and extracellular receptors and affect cells from the innate, adaptive immunity system and other cell types. IL-6 overexpression is linked to cell proliferation, proliferative diseases and plays a prominent role in the pathogenesis of KSHV-induced Diseases, such as Castleman's Disease, Kaposi Sarcoma and Autoimmune Diseases (Burger et al., 1994, Miles et al., 1990, Jones et al., 1999, Oksenhendler et al., 2000, Aoki et al., 2000).

Several transcripts that are known to be processed by SOX preferentially have been shown to contain several UGAAG and GAAGU motifs, within the 5' untranslated region (5' UTR), protein-coding region and often in the 3'UTR (Covarrubias et al., 2011, Clyde and Glaunsinger, 2011). IL-6 contains one UGAAG and one GAAGU motif in the 3'UTR. A recent paper (Hutin et al., 2013), has demonstrated that IL-6 contains a so-called SRE1 (SOX-resistant element 1) in its 3'UTR. This SRE1 contains a non-canonical ARE, which is a stretch of AU-rich sequence, containing the core AUUUA ARE motif, and is bound by AUF1 and HuR. Hutin *et al*. (2013) demonstrated that the protection of IL-6 from SOX mediated decay was observed in the presence of AUF1 and HuR. But when AUF1 and HuR were silenced using siRNA (Hutin et al., 2013), Il-6 was susceptible to SOX-mediated decay.

1.3. RNA Folds and *in silico* Folding

1.3.1. RNA Folds

The biological importance of RNA has grown as more and more functions and roles are discovered such as those in replication, translational regulation and viral propagation. These functions, as in proteins, are transduced by their folds. Most recently, a new field has opened up focused on RNA and viruses. RNA pseudoknots in retroviral mRNAs have been found to cause programmed frameshifts (Silva et al., 2010, Chamorro et al., 1992). These have been shown to produce the correct ratios of proteins required for viral propagation (Chamorro et al., 1992, Tinoco Jr and Bustamante, 1999). In addition, correlations between host temperature and viral infectivity owing to the stability of RNA structure have also been discovered putting knowledge of the way in which RNA folds at the forefront

of addressing key questions in virology (Brower-Sinning et al., 2009).

It has been shown that RNA folds hierarchically and co-transcriptionally (Walter et al., 2009, Brion and Westhof, 1997). This first involves folding of the much more stable secondary structure compared to its denatured state, which acts as an initiator for its assembly into subsystems of increasing size and complexity, leading to the formation of tertiary structure (Zemora and Waldsich, 2010, Brion and Westhof, 1997, Tinoco Jr and Bustamante, 1999). The tertiary folds are less stable and more susceptible to ion concentration and temperature changes than the secondary folds, which are strengthened by base pairing interactions and π stacking. The tertiary fold interactions can also be interrupted by RNA binding proteins, thereby allowing the protein to bind the secondary structures and/or their primary sequence motifs found in loops and bulges (Figure 1-6).

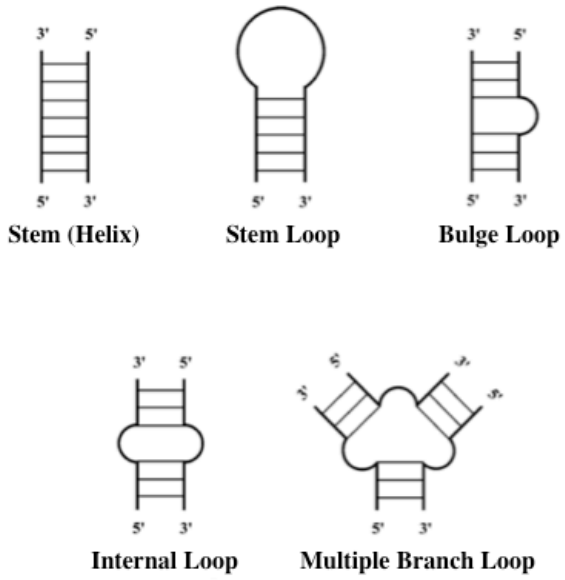


Figure 1-6: RNA Secondary Structure Elements

Five different RNA secondary structure elements are represented: stem (helix), stem loop, bulge loop, internal loop and multi branch loop (Ding and Lawrence, 2003).

1.3.2. *In silico* Folding of RNA

As for proteins, the secondary structure of RNAs confers their function, and fold is more conserved than sequence (Capriotti and Marti-Renom, 2010). Thus

tertiary and secondary structures predictions can be used instead to find or refine regulatory elements that are evolutionarily conserved, and hence, potentially functionally conserved. In genomics, dinucleotide content is often used to predict protein-coding regions and plays an important role in the stability of RNA secondary structures (Birney et al., 2007). Dinucleotide content considerably can affect the stability scores of secondary structure and energy predictions. Dinucleotide content is an important contributing factor in the calculation of the folding energy, because of their stacking energy contributions (Washietl et al., 2005, Workman and Krogh, 1999). With the recent advances in RNA research and increased RNA X-ray crystal and nuclear magnetic resonance (NMR) structures the need for *in silico* RNA fold prediction software has increased in numbers and accuracy.

1.3.2.1. Secondary Structure Prediction - *mfold*

To date the most widely used structure prediction algorithm is the minimum free energy (MFE) method folding single sequences. This method is implemented in one of the longest existing RNA secondary structure prediction programs *mfold* (Tinoco Jr and Bustamante, 1999, Workman and Krogh, 1999). *Mfold* uses the Zuker-Stiegler algorithm for MFE computing using every possible base-pairing, which is forced one-by-one and using empirical estimates of thermodynamic parameters for interactions and loop entropies to score structures (Zuker, 2003b, Zuker and Stiegler, 1981, Mathews et al., 1999). Hence *mfold* secondary structure results are ranked by their MFE.

1.3.2.2. 3D Secondary Structure Prediction - *Mc-Sym*

Current 3D (Three-Dimensional) RNA folding algorithms require manual manipulation or are generally limited to simple structures in terms of size and topology. The prediction accuracy improves with added knowledge from the 2D (Two-Dimensional) structure, but still fails in the prediction of long-range contacts, which are involved in establishment of the tertiary structure (Laing and Schlick, 2011, Mathews et al., 2010). RNA structures that are much longer than 50 nucleotides cannot be predicted with great confidence, because of an increase in

complexity of the probabilistic model, which would have to account for the topologies such as junctions and long-range contacts. 3D structures of RNAs less than 20 nucleotides are better predicted with all-atom knowledge based approaches (such as fragment assembly of RNA with full-atom refinement (FARFAR)), while structures of longer RNAs are better predicted by coarse graining approaches (such as *Mc-Sym*) (Parisien and Major, 2008, Laing and Schlick, 2011). As mentioned, *Mc-Sym* uses a *Mc-Sym* coarse graining approach to allow conformational sampling. It generates 3D structure models from small-residue fragments.

Chapter 2: Biophysics Background Theory

2.1. Macromolecular X-ray Crystallography

3.2.1. X-ray Diffraction to Solve Molecular Structures

X-ray crystallography is a method that is prolifically used to produce atomic or near-atomic resolution structures in academic and industrial research. The immense technological advents in synchrotron tools (such as microfocus high-intensity X-ray source and charge-coupled device (CCD) detectors) (Adams et al., 2010, Arndt et al., 1998, Fischetti et al., 2009, Phillips et al., 2002), in computation power, data storage space and crystallographic software packages (*XDS*, *CCP4 suite*, *Phenix*) have lead to an explosion in X-ray crystal structures of biological macromolecules, macromolecular assemblies and membrane proteins being deposited in the PDB (Adams et al., 2010, Winn et al., 2011, Kabsch, 2009). These 3D atomic or near-atomic resolution models have allowed elucidation of the mechanisms behind biochemical life, such as enzymatic activity, binding of small ligands, of macromolecules and formation of macromolecular assemblies, whose functions are related to their structures.

The distance observed between atoms in macromolecules is between 1-2 10^{-10} m (or 1-2 Ångström (Å)), thus this is the resolution range that is needed to get a better understanding of the biochemical mechanisms. X-rays have wavelengths of up to 10 pm allowing X-ray crystal structures to be solved to resolutions of up to 0.48 Å. To obtain such a crystal structure, a crystal made of exact repeats of the molecules is needed to amplify the signal of the X-ray scattering. The crystal composition has to be characterized from the diffraction data. Finally due to lack of an X-ray lens, the phase problem has to be solved to produce a structure that can be refined and validated so that it can be used to explain the mechanisms of life.

3.2.2. The Real Crystal Lattice

In X-ray crystallography the physical properties of X-rays and crystals are exploited. The diffraction from a single macromolecule in a crystal would be too weak to be measured, but the signal is amplified, as the array of macromolecules diffracts the same image multiple times. A crystal is made up of a symmetrical set

of repeating units, which contain the macromolecule of interest, in a lattice (Figure 2-1). A lattice can be characterised in terms of its dimensions, morphology and the relationship of the unit cells to each other and in relation to sets of parallel planes in space. These are called unit cells and each can be rotated or translated onto another by mathematical operators. The space group describes the morphology and symmetry. The asymmetric unit is the smallest repeating unit the crystal can be divided into using the crystallographic symmetry operations of the space group. This allows the crystallographer to provide a mathematical description of a crystal system and a physical description of the atomic arrangement in the asymmetric unit, which includes solvents, amino acids and nucleic acids (Rhodes, 2006, Drenth, 2007, Rupp, 2010).

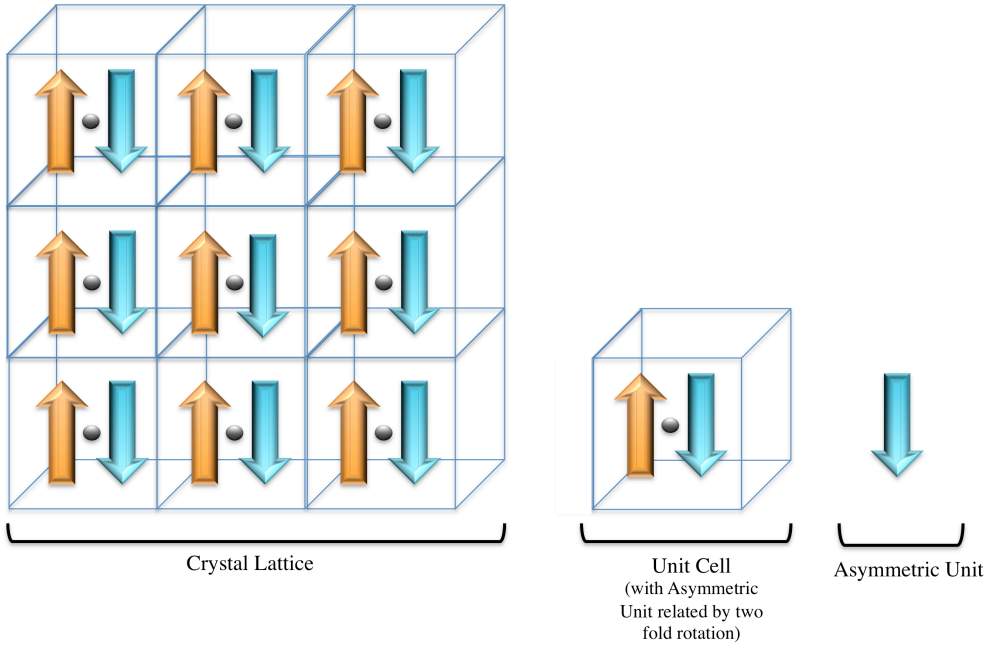


Figure 2-1: Crystal Systems

A crystal lattice is made up of repeating units termed unit cells, which can also contain repeats of a unique object, generally DNA, RNA or protein. This minimal repeat is called an asymmetric unit. The asymmetric units and unit cells within a crystal lattice are symmetrically related and can be superposed using mathematical operators, rotations or translations.

3.2.3. X-ray Scattering and Bragg's Law

X-rays are a form of electromagnetic radiation, which when they pass through matter interact with the electrons around the nuclei. X-rays are generated by accelerating electrons in a vacuum using an electric field directed against an anode (a metal, i.e. Copper (Cu)), leading to multiple collisions (Rhodes, 2006, Rupp, 2010). On collision, some electrons convert their energy via an inverse photoelectric effect into a continuum of X-rays. When the incident X-ray beam enters a crystal, it is partly absorbed by the atoms in its path and will scatter X-ray radiation to produce secondary waves in all directions. Waves are characterized by their frequency, amplitude and phase. Some scattered waves will be subject to constructive interference or destructive interference depending on their relative phase. During the latter case the scattered waves are out of phase by π or any odd multiple of π (180°) and thus cancel out (Rhodes, 2006, Drenth, 2007, Rupp, 2010).

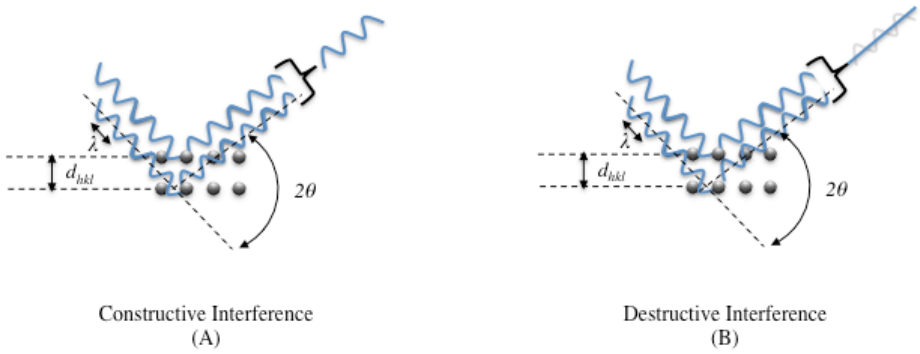


Figure 2-2: Constructive and Destructive Interference of Waves

In both A) and B) the waves have the same amplitudes and frequencies. In A) the waves scattered by the electron density are in phase and constructively interfere, whereas in B) the scattered waves are out of phase by π or 180°. Thus, as they have the same amplitude the waves cancel each other out by destructive interference. These phenomena are described by Bragg's Law and are influenced by the angle θ.

Sir W.H. Bragg (1862-1942) and Sir W.L. Bragg (1890-1971) proved that X-rays reflected from the atoms in a crystal could be treated as if they were reflected by a set of atoms on parallel planes (Figure 2-2). Bragg's law defines the conditions that lead to diffraction by constructive interference. This law states that

only scattered waves where the path length difference corresponds to an integral number of wavelengths will produce diffraction (Equation 2-1).

Equation 2-1: Bragg's Law of Diffraction

$$2d_{hkl} \sin\theta = n\lambda$$

where

θ is the incident angle of the wave

d_{hkl} is the spacing between lattice planes on which atoms reside

n is the integer

λ is the wavelength of incident wave.

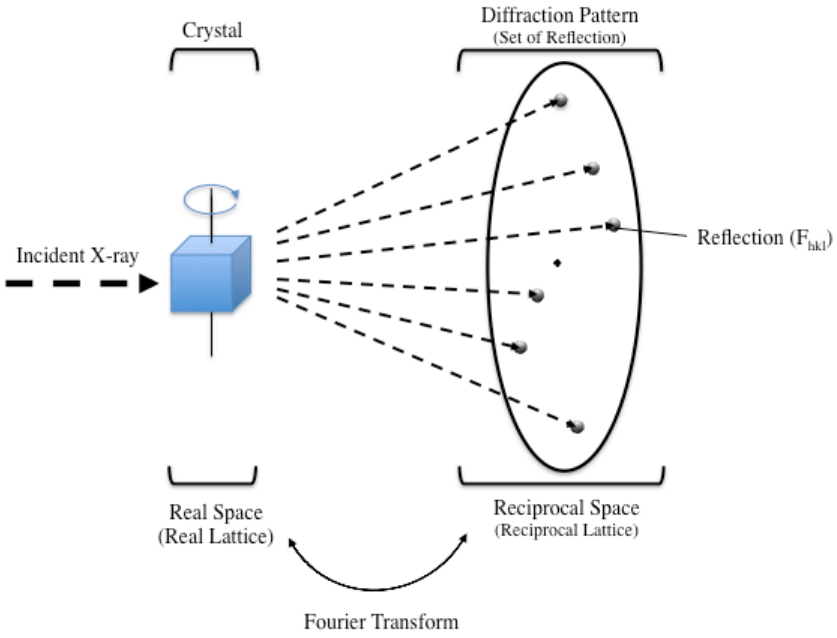


Figure 2-3: X-rays Crystallography Experimental Set Up

A diffraction pattern is made of diffraction maxima that produce a pattern of spots on a detector. These can be viewed as originating from each atom that contributes to the scattering of X-rays from a potentially infinite number of parallel planes (reflections). Thus reflection F_{hkl} corresponds to all atoms that affect the scattering along an identical set of lattice planes hkl in the crystal.

When the scattered waves are subject to constructive interference, a single diffraction spot is the result of positive reinforcement of the scattered waves from

atoms within a set of parallel planes, which outweigh the contributions of random noise atoms (Rhodes, 2006, Rupp, 2010). The diffraction pattern collected (Figure 2-3) when using X-rays in crystallographic experiments is hence the result of contributions from all atoms within the macromolecule that when repeated in 3 dimensions, forms the crystal.

3.2.4. Characterization of the Real and Reciprocal Space

3.2.4.1. Symmetry, Space and Point Groups

The first step in solving a crystal structure after collecting the diffraction data is the determination of a set of indices that describe the sets of lattice planes (hkl, referred to as reflection indices) to calculate the unit cell dimensions (a, b, c and angles α, β, γ) and the point group ultimately leading to space group identification (Rhodes, 2006).

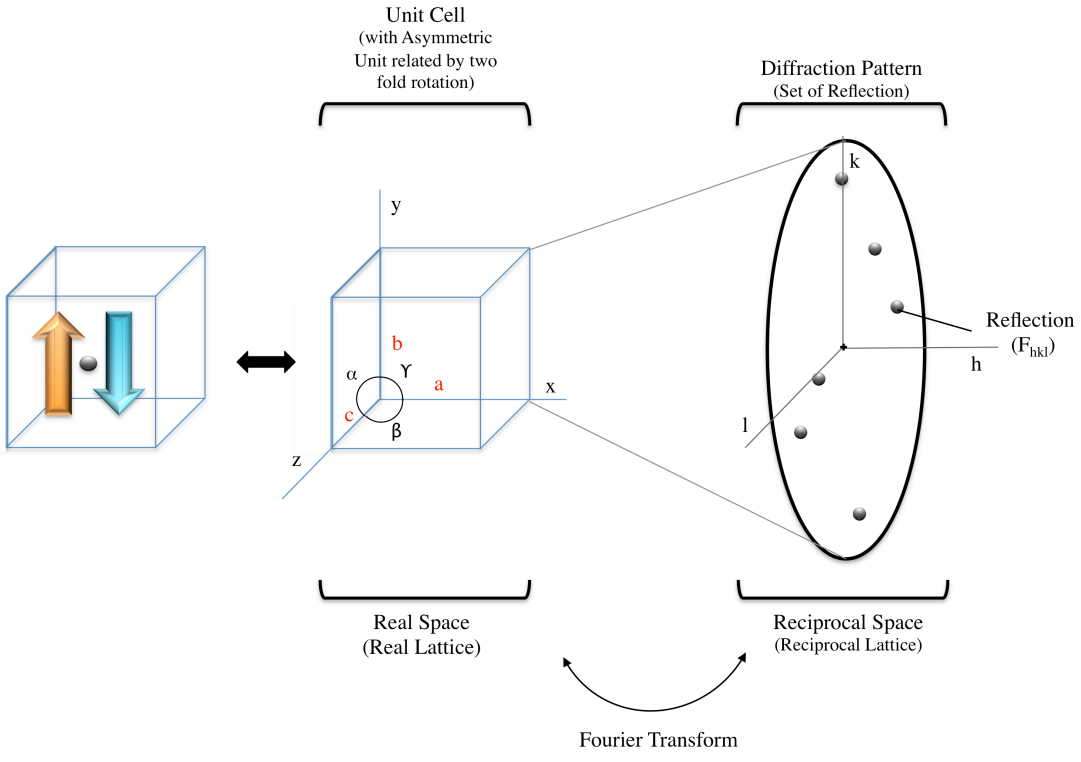


Figure 2-4: The General Geometry of a Unit Cell

A unit cell is characterized by six parameters: three sides (a, b and c) and three angles (α, β and γ).

Each crystal is defined by its geometry; three faces (*a, b* and *c,* one for each spatial dimension, *x*, *y* and *z*) and three angles (α, β and γ) (Figure 2-4). The crystal's geometry and morphology can be associated with one of the seven crystal systems (Table 2-1). Associated with these crystal systems are potential lattice centrings, these are termed face-centred (F)-, C-centred (A, B or C) or body-centred (I) points of symmetry (Figure 2-5). The fourth is referred to as primitive (P) as there are no centres of symmetry. All crystals have at least a primitive Bravais lattice (Rhodes, 2006, Rupp, 2010).

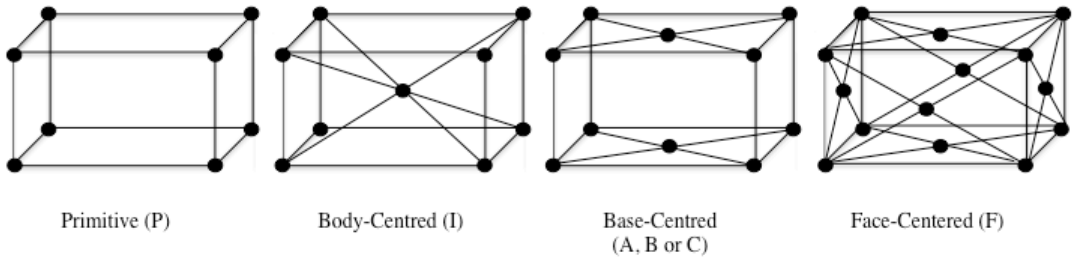


Figure 2-5: The Set of Bravais Lattices of an Orthorhombic Crystal System

A crystal lattice can have: primitive (P) symmetry, where only the corners of the system correspond to elements of symmetry, body-centred (I), where in addition to the corners there is one additional lattice point at the centre of the cell, base-centred (A, B or C), where in addition to the corners there is a further lattice point at the centre of each of one pair of the cell face, (i.e. face A, B or C) and face-centred (F) where in addition to the corners (vertices) there is also a lattice point at the centre of each of the faces of the cell. All A- or B-centred lattices can be described by C-centring.

The crystal is constructed by applying lattice translations to the unit cell contents to fill 3D space. The various asymmetric units can be related by screw axes or pure rotations. In either event, the space group is the combination of lattice translations and any centring if relevant, combined with the operators relating the asymmetric units. The space group notation contains the Bravais lattice type (T_B) and set of mathematical operators (i.e. screw axis) in the Hermann-Mauguin notation ($Am_{/A}$, $Bm_{/B}$, $Cm_{/C}$) along the x (i.e face a), y (i.e face b) and z (i.e face c)

axis (formula 2-1). The screw axis operators combine rotation with fractional translation when applied to the asymmetric units forming the unit cell (Formula 2-1).

Formula 2-1: Space Group in Hermann-Mauguin notation

$$T_{B\,Am/_{A}Bm/_{B}Cm/_{C}}$$

where

T_B is the Bravais lattice type, e.g. P, C, I, …

$Am/_A$ is the screw axis along face a, i.e. the x axis, where A is n

$Bm/_B$ is the screw axis along face b, i.e. the y axis, where B is n

$Cm/_C$ is the screw axis along face c, i.e. the z axis , where C is n

The number n denotes the screw axis, where the angle of rotation is $360°/n$ and the degree of translation along the axis of the face is m/n. E. g. for the space group P4, the Bravais lattice is primitive, with a rotation of 90° (as $360°/4 = 90°$) and no translation as no degree of translation is specified, whereas in $P4_1$ the Bravais lattice is primitive, with a rotation of 90° and a $1/4$ (as $m = 1$) fractional translation along the axis (Figure 2-6).

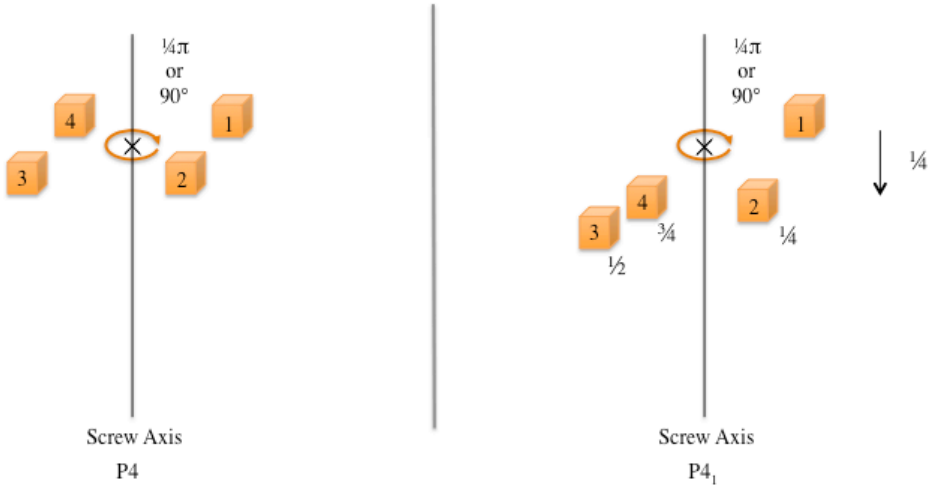


Figure 2-6: Screw Axis and Symmetry Operations

When n is equal 4 then the object is rotated by 90° or $1/4$ around the axis on the same plane and no translation occurs. When n is equal 4 and m equals 1, then each time the object rotates around the axis by 90°, the object is also translated by $1/4$.

Table 2-1: The Seven Crystal Systems, their Unit Cell Dimensions and associated Point and Space Groups

Crystal System	Variations	Unit Cell Faces Angles	Space Groups Numbers	Point Group with associated Space Group
Triclinic	Primitive	$a \neq b \neq c$ $\alpha \neq \beta \neq \gamma \neq 90°$	2	1, (e.g. P1)
Monoclinic	Primitive, Base-Centred	$a \neq b \neq c$ $\alpha = \gamma = 90°$ $\beta \neq 90°$	13	2 (e.g. P2, $P2_1$, C2)
Orthorhombic	Primitive, Body-, Base- and Face-Centred	$a \neq b \neq c$ $\alpha = \beta = \gamma = 90°$	59	222, (e.g. P222, $P222_1$, $P2_12_12$, $P2_12_12_1$, $C222_1$, C222, F222, I222, $I2_12_12_1$)
Tetragonal	Primitive, Body-Centred	$a = b \neq c$ $\alpha = \beta = \gamma = 90°$	68	4 (e.g. P4, $P4_1$, $P4_2$, $P4_3$, I4, $I4_1$), 442 (e.g. P422, $P42_12$, $P4_122$, $P4_12_12$, $P4_222$, $P4_22_12$, $P4_322$, $P4_32_12$, I422, $I4_122$
Trigonal (Rhombohedral)	Primitive	$a = b = c$ $\alpha = \beta = \gamma \neq 90°$	25	3, (e.g. P3, $P3_1$, $P3_2$, R3), 32, (e.g. P312, P321, $P3_112$, $P3_121$, $P3_212$, $P3_221$, R32)
Trigonal/ Hexagonal	Primitive	$a = b \neq c$ $\alpha = \beta = 90°$ $\gamma = 120°$	27	6, (e.g. P6, $P6_1$, $P6_5$, $P6_2$, $P6_4$, $P6_3$), 622, (e.g. P622, $P6_122$, $P6_522$, $P6_222$, $P6_422$, $P6_322$)
Cubic	Primitive, Body-and Face-Centred	$a = b = c$ $\alpha = \beta = \gamma = 90°$	36	23, (e.g. P23, F23, I23, $P2_13$, $I2_13$), 432, (e.g. P432, $P4_232$, F432, $F4_132$, I432, $P4_332$, $P4_132$, $I4_132$)

Hence the space group parameters contain all the information needed to generate the unit cell from the asymmetric unit and gives an idea of the amount of information contained with one unit cell. Thus the determination of the point group (i.e. symmetry in the absence of translational elements) is required prior to data collection, as it gives an indication of how much angular information has to be collected for a complete dataset (Rhodes, 2006, Drenth, 2007, Rupp, 2010).

For determination of the unit cell dimensions and space group, software such as iMOSFLM (Battye et al., 2011) and the *XDS* pipeline (Kabsch, 2009) are used. They are routinely used for data processing, with their initial task being the determination of the unit cell parameters and the identification of potential point groups in a process known as autoindexing where reflection indices are assigned to each spot.

3.2.4.2. Reciprocal Lattice to Real Lattice - Fourier Transform

Once the unit cell parameters and space group have been determined within the crystal (real lattice), by using the space observed between the reflections ($|F_{hkl}|$ or structure factors) on the diffraction pattern (reciprocal lattice) (Figure 2-3 and 2-4), the characterization of the electron density can be undertaken to reconstitute the asymmetric unit.

In a diffraction experiment, the intensities and the position of reflections are measured or recorded. From the position of the reflection, the index triple (h,k,l) can be determined and the appropriate intensity can be assigned. This intensity can be truncated to the structure (factor) amplitude, $|F_{hkl}|$, which is found in the reciprocal space from which the position of the electron density (ρ) within the real space can be deduced. Thus, to reconstitute the asymmetric unit within the unit cell the reflection ($|F_{hkl}|$) need to be assigned and this is achieved using a Fourier sum (Equation 2-2). The Fourier transform and sum were named after J.B.J Fourier (1768 –1830) a French mathematician and physicist, who observed and modelled heat transfer and vibrations, which are also waves. His work has allowed the description of complex waves as a sum of a series of sinusoidal waves (Rhodes, 2006, Drenth, 2007, Rupp, 2010).

As each diffraction spot (reflection) in the diffraction pattern can be represented mathematically as the sum of all the individual trajectories, with the sum of those waves being equivalent to what is ultimately represented in a diffraction spot, the Fourier summation is used to calculate the amplitudes, $|F_{hkl}|$.

Equation 2-2: Structure Factor as a Fourier sum

$$F_{hkl} = \sum_{j=1}^{n} f_j e^{2\pi i\,(hx_j + ky_j + lz_j)}$$

where

F_{hkl} is the sum of every atomic structure factor of every atom along a set of parallel planes, thus the average structure factor

n is the number of terms addressed by the total sum

j is a specific atom related to the position at h,k,l

f_j is the scattering factor of the atom j, which is determined by the size of its electron shell and contributes to the amplitude of a constructive wave

hx_j, ky_j, lz_j are the fractional coordinates of atom j in the summation, and h, k, l the three indices of the corresponding reflection.

Using the structure factors that were obtained using a Fourier sum the electron density (ρ) can be regenerated, with a Fourier transform. This equation is called the electron density equation (Equation 2-3). It is used to calculate electron density and Patterson maps from structure factors.

Equation 2-3: Electron density equation from a Fourier transform

$$\rho(x,y,z) = \frac{1}{V} \sum_{h} \sum_{k} \sum_{l} |F_{hkl}| \,[\cos 2\pi(hx + ky + lz) - \phi(hkl)]$$

where

$\rho(x,y,z)$ is the electron density at point (x,y,z)

V is the unit cell volume

$|F_{hkl}|$ is the intensity or structure factor

$hx + ky + lz$ are the fractional coordinates and h, k, l the three indices of the corresponding reflection

$\phi(hkl)$ is the phase.

The electron density equation allows reconstruction of the real space density, $\rho(x, y, z)$, using the amplitude, $|F_{hkl}|$, which is proportional to the square root of the reflection intensity, where the frequency of each term in the Fourier sum is equivalent to the coordinates h, k and l, and lastly the phases, $\phi(hkl)$ (Drenth, 2007, Rupp, 2010). A wave has three characteristics its intensity, frequency and phase. The diffraction experiment due to its two dimensional setup is able to record the intensity and the frequency, whilst the phase information is lost. And as the electron density equation needs the derivation of phases to calculate the electron density, this is referred to as the phase problem. To solve the phase problem further experimental or computational means have to be undertaken.

3.2.5. Crystallographic Phase Problem

There are two strategies for solving the phase problem, experimental or computational. Each has its own advantages and disadvantages. In experimental phasing, datasets are used for heavy atom substituted variants of the macromolecule of interest for isomorphous replacement (MIR) or alternatively, anomalous scatterers are incorporated for single- or multiple-wavelength anomalous dispersion (SAD/MAD) experiments. Heavy atoms can either be incorporated within the expression system (e.g. selenomethionine), after crystallisation by soaking existing crystals in a heavy atom solution or by co-crystallisation. As the PDB contains increasing numbers of macromolecular structures, which represent a wide variety of folds, macromolecular structures are being solved by molecular replacement (MR). In MR an existing atomic model is used to approximate the phases of an unknown homologue. The unknown phases of the new map are substituted in the Fourier transform with the phase of the MR model to approximate those of the target molecule and thus used to generate electron density. The model is then iteratively refined until no further improvements can be made as the initial phases will all be error prone, especially when obtained using MIR/MAD/SAD. For both methods specialized software is used to obtain the missing phase information.

3.2.5.1. Solving the Phase Problem by Experimental Phasing

Maximum likelihood is discussed in the MR section, but it is an approach that has been successfully applied to all aspects of structure determination and refinement in protein crystallography. Maximum likelihood in MIR/MAD/SAD is used in heavy atom parameter refinement and phasing (Fortelle and Bricogne, 1997, Otwinowsky, 1991), where it is effective in weighting derivatives. In refinement, a maximum likelihood target equation is minimized to maximize the probability of obtaining the known observations given the various parameters derived from the model (e.g. bond lengths, angles, chirality).

5.2.5.1.1. Isomorphous Replacement (MIR)

Experimental phasing techniques exploit the frequent isomorphism of datasets to facilitate the process of obtaining starting phases to build a model. Thus at least two datasets are needed, firstly one native dataset and secondly one dataset carrying heavy atoms. Both datasets must be derived from the same type of crystal. Heavy metals are most frequently used as these have a higher atomic number and thus have a much greater electron density compared to the atoms normally found in biological macromolecules, such as nitrogens, oxygens, carbons and hydrogens. This substantial difference in electron density and thus scattering power leads to significant differences between the native and atom structure factors. Bromide, gold, mercury and platinum are most commonly used for this technique. Mercury or gold target histidines and cysteines may form covalent links, while bromide or iodide target glutamate and aspartate residues through non-covalent links (Rhodes, 2006, Rupp, 2010). Heavy atoms are generally introduced into a protein lattice by soaking existing crystals in a solution of heavy atoms until all the binding sites for the heavy atom are saturated. Unfortunately, this can disrupt the packing of macromolecules in the crystal leading to changes in the relative positions and orientations of monomers within the asymmetric unit, which can render MIR unusable.

In MIR it is assumed that the two crystals used to obtain the differential dataset remain isomorphous after soaking with heavy atoms. In this case, the two datasets are obtained at the same wavelength and their structure factors are first

compared, by looking at the isomorphous difference. Thus, if the structure factor expression for the native (F_N) is subtracted from the derivative heavy atom (F_{HN}) dataset, which in addition contains the structure factor expression for the heavy atoms, information is yielded on the location of the heavy atoms in space (F_N). For this the difference Patterson function is used (Equation 2-4) and it allows the determination of the location of the heavy atoms (Figure 2-7). The Patterson function is a Fourier sums without the phases. The structure factors, which are proportional to the measured reflection amplitudes, in this equation are squared and thus correspond to the intensities. This allows us to calculate a series of intensities, without the phases. In the difference Patterson function the structure factor is $(\Delta F)^2 = (|F_{NH}| - |F_N|)^2$, thus is equal to the lone contribution of the heavy atoms (F_H) (Fortelle and Bricogne, 1997, Otwinowsky, 1991). This allows the production of a vector map that will reveal the location of the heavy atoms within the unit cell.

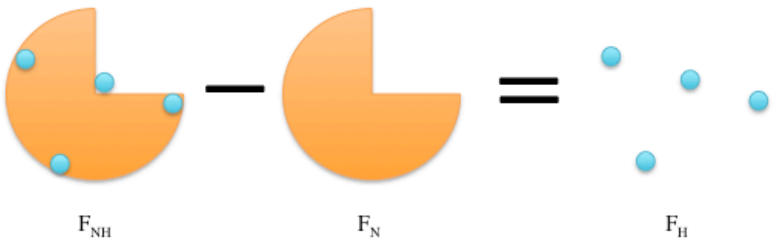


Figure 2-7: Obtaining F_H - Difference Patterson Function

The difference Patterson map allows removing the noise contributions from the native data (F_N) in the derivative dataset (F_{NH}) to obtain F_H and thus to determine the location of the heavy atoms in space.

This vector map was devised by Patterson and corresponds to the convolution of electron density (i.e. the electron density at a point x,y,z multiplied by the electron density at all other points within the asymmetric unit/unit cell). This gives rise to a map in which there are peaks at the ends of interatomic vectors. Ordinarily, it would be too complex to analyse if it were computed using the F_N's or F_{NH}'s alone. Using the isomorphous differences, however, the protein component is subtracted leaving only the heavy atom contribution. The vectors therefore correspond to those between symmetry related heavy atoms from which the x,y,z co-ordinates can be calculated. Another important point is that the phase

ambiguity arising from a single derivative is different to the ambiguity arising from whether the co-ordinate is x,y,z or –x,-y,-z (Rhodes, 2006, Drenth, 2007, Rupp, 2010). The latter is resolved by computing maps in both "hands", the former only by using an additional derivative.

Equation 2-4: Difference Patterson function

$$\Delta P_{(u,v,w)} = \frac{1}{V} \sum_{h} \sum_{k} \sum_{l} \Delta F_{hkl}^{2} \, e^{-2\pi i(hu+kv+lw)}$$

where

$\Delta P_{(u,v,w)}$ is the Patterson function at point (u, v, w) on the Patterson map

V is the unit cell volume

ΔF_{hkl}^{2} is the intensity or square of structure factor F_H

$hu + kv + lw$ are the fractional coordinates and h, k, l the three indices of the corresponding reflection

The Patterson function contains a set of vectors that define all inter-atomic interrelations, thus there are more vectors in the Patterson map than there are atoms. The sets of heavy atoms can be located by trialing all of the possible arrangements of atoms with the distances described by the Patterson map for the arrangements of the heavy atoms in a unit cell (Figure 2-7 and 2-8), this is called the Patterson superposition. Using an arbitrary heavy atom peak as an origin in the Patterson map and then tracing all the remaining vectors for every possible combination of peaks, a match to the peaks in the Patterson function is computationally investigated. This technique can be used on molecules of up to 1000 atoms. This is not enough for proteins but enough to solve the heavy atom substructures for isomorphous replacement and anomalous dispersion, although for large substructures, Patterson techniques have to be replaced by the direct methods approaches used for obtaining small molecule structures. This leads to two possible phase approximations, which were deduced from the centrosymmetric Patterson map using the fixed distances between the heavy atoms. This is the second source of phase ambiguity; the first is due to the nature of the cosine function. The ambiguity is broken by a second or sometimes multiple derivatives that for each hkl will have one phase in common or

by incorporating anomalous data (SIRAS). The requirement for a second derivative is graphically represented in figure 2-9 in what is known as the Harker construction.

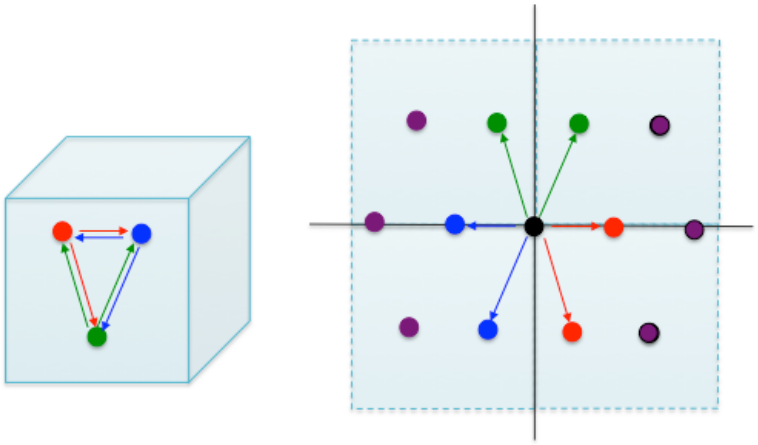


Figure 2-8: Difference Patterson Map of a Three-Atom Coordinate System

Taking all vectors between the heavy atoms and plotting them from the origins of the unit cell, as in a simple coordinate system, can construct a Patterson map. This allows the determination of the absolute and relative locations of the heavy atoms. All of the possible combinations of atoms can be probed for the correct arrangement of the heavy atoms in the unit cell. On the left, 3 heavy atoms found in a unit cell and their 6 inter-atomic vectors are represented. The vectors are coloured by originating atoms. In the Patterson map, on the right, atoms were each placed at the origin, one corner of the unit cell. The Patterson peaks that were overlapping with the atoms obtained using the vectors are coloured according to the vectors used.

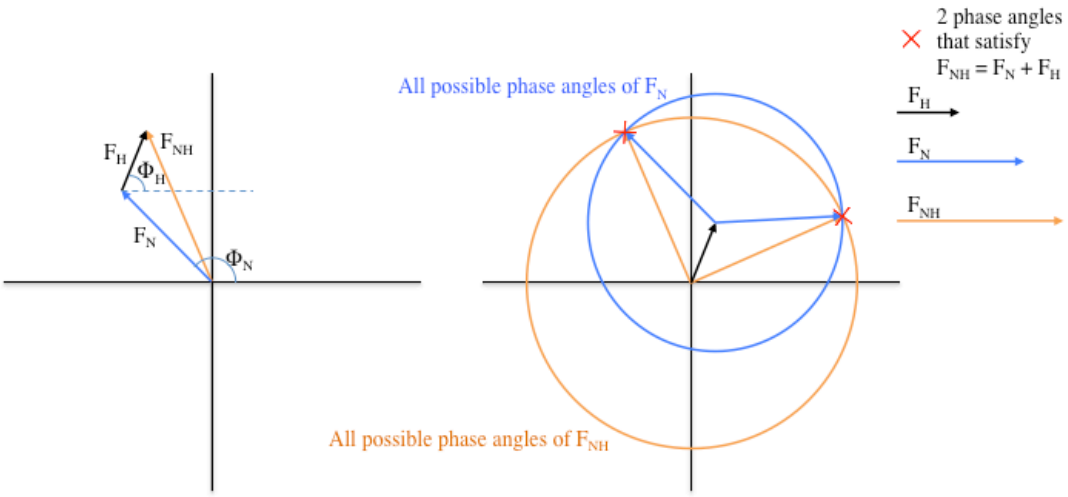


Figure 2-9: Harker Diagram for the Determination of the Two Possible Phase Angles for $F_{NH} = F_H + F_N$

The vectors in the Harker diagrams represent the amplitude, in length, and phase (Φ), in isomorphous replacement. The vector F_{NH} (orange arrow), which represents the heavy atom dataset, is a summation of the vector F_N (blue arrow), of the native dataset and the vector F_H (black arrow) of just the heavy atom contribution. The Harker diagram can be constructed from a single derivative by tracing a circle with a radius equal to F_N (blue arrow and circle) at the origin of the arbitrary origin of a 2D coordinate system, and a circle with a radius equal to F_{NH} (orange arrow and circle), offset from the origin by F_H. The circles represent all the possible phases that the vector could have. The two intersections that the circle produces represent the two possible phase angles of the vector F_N (Drenth, 2007, Rupp, 2010). In order to definitively identify the phase, a second derivative dataset can be obtained.

5.2.5.1.2. Anomalous Scattering (SAD/MAD)

Isomorphous replacement relies on the heavy atom derived crystal remaining isomorphous to the native crystal, despite changes within the crystal. Sometimes, these changes can be so extensive that the difference between F_N - F_{NH} is compromised and even maximum-likelihood based methods cannot compensate for the difference. In these cases, the anomalous signals from atoms such as mercury and selenium, which produce an anomalous scattering signal at specific wavelengths, are obtained at synchrotron X-ray sources. These wavelengths are chosen as their energies cause an electronic transition in the anomalous scatterer in

the range covered by synchrotron radiation, leading to small changes in the scattering intensity, which can be measured and used to solve the phase problem. The most common atom used in this method is selenium, which is incorporated into the protein using recombinant expression in the presence of selenomethionine. Selenomethionine substitutes the native methionines (Hendrickson et al., 1990). Instead of the terminal sulphur group selenomethionine contains a selenium group.

An anomalous dispersion experiment is carried out using a wavelength close to the absorption edge of the specific anomalous scatterer. At this edge appreciable absorption occurs where the emitted X-ray emerges. This results in the loss of centrosymmetry in the diffraction pattern, which is manifested by a breakdown in Friedel's law. Friedel's law states that $|F_{hkl}| = |F_{-h-k-l}|$. This breakdown leads to small but measureable differences between the $|F_{hkl}|$'s measured at different wavelengths. Datasets collected at different wavelengths close to and around the absorption edge therefore can be effectively used as a series of MIR datasets due to changes in the scattering correction where these experiments are referred to as multiple-wavelength anomalous dispersion (MAD) (Rupp, 2010). Single anomalous dispersion where data from only a single wavelength, however, is often sufficient to obtain a preliminary set of phase estimates. Although there is still a phase ambiguity, the phase probability distribution in SAD is often slightly skewed towards the correct phase. Interpretable maps can therefore be frequently obtained using density modification techniques. Due to the technical advances made and when the anomalous scattering signal is strong enough, the use of just a single wavelength is possible to phase a map.

A Friedel's pair of reflections is centrosymmetric under normal scattering conditions (Drenth, 2007, Rupp, 2010). The anomalous scattering leads to subtle but measureable differences in amplitudes. As the wavelength used in these experiments is fixed, the magnitude of the anomalous signal is constant. Thus it can be read from reference table, detailing the information for all atoms. The anomalous signal is represented as the addition of vector F_H to the normal scattering vector F_N, resulting the vector F_{NH} (Drenth, 2007, Rupp, 2010). And thus these differences are used to obtain the starting phases. As in isomorphous replacement, in anomalous

scattering replacement, a Patterson difference map is used to locate the origin of the anomalous scattering, the position of the anomalous scatterer. Using the established standard of absorption magnitude for the anomalous scatterer and the identified change in the phase of the Friedel pair using the Patterson difference map, the approximate phases can be calculated and used for the initial electron density map and initial model building. The model the will then have to go through iterative refinement steps.

5.2.5.1.3. Solving the Phase Problem by Molecular Replacement (MR)

When using MR to solve the phase problem, phases of a structurally related molecule from the PDB are used to substitute for the unknown phases of the target macromolecule. When the crystals share the same crystal system, point symmetry and space group, then the packing of the macromolecule is going to be of close similarity, thus the phases of the structurally related structure can be used directly in completing the electron density function of the unknown target macromolecule (Driessen and Tickle, 1996). However, even if a high sequence identity or a similar tertiary structure is available, this does not necessarily mean that these will crystallize in the same way, as protein concentration, domain truncations and even slight changes in surface residues mediating crystal contacts, affect the crystallization process.

MR is trying to find the orientation and location of the target using the co-ordinates of a closely related molecule (model). This is a six dimensional problem that can be broken down into two, each of 3 dimensions. The first involves finding the correct orientation of the target and traditionally relied on the fact that the model and target have similar atomic distributions and therefore share intra-atomic vectors. As for MIR, a Patterson Function (Equation 2-4) can be used for this (Rhodes, 2006, Rupp, 2010). This, in this case, is a flattening of the 3D structure to allow comparison and superposition of the two structures. The Patterson maps of the two structures with their collections of inter-atomic distance vectors can be matched. Some of the vectors will be the same, while others will be different, due to small structural differences between the structures. Thus a certain level of similarity

has to be present to elevate the signal of the same vectors above the noise of the dissimilar vectors. *Molrep* (Vagin and Teplyakov, 1997) is a program that attempts to match the two Patterson maps through a series of rotations and translations until the two are superimposed. Then the phases of the model are recalculated and used to build an initial electron density map for the unknown structure. As previously mentioned if there is a greater difference between the two structures or the content of the crystal is very complex, this method is less amenable and becomes computationally intensive.

To alleviate this, programs like *Phaser* (McCoy et al., 2007) can be employed. These use maximum-likelihood statistics to increase the probability of finding an MR solution. The maximum-likelihood method targets the MR issue more comprehensively by calculating how well the solution predicts the observations used to obtain it. The higher the equivalence in reflections in each rotation and translation, is the higher the probability that the solution is correct. Rotational or translational matches with poor statistical significance are discarded. To differentiate between all the possible solutions *Phaser* uses Z-scores and log-likelihood gain (LLG). LLG measures randomness and a negative LLG value would indicate that the input search model is worse than a random collection of atoms at describing the target data. But with every rotation or translation step in the right direction the LLG should increase, as randomness decreases. As many of the results may have a high LLG score as they all match some proportion of the data, the Z-score measures the number of standard deviations of these statistical likelihoods above the mean; thus the signal to noise ratio. Z-score compares the LLG values from the rotation or translation search with LLG values of a set of random rotations or translations. The more significant the match is, the greater the Z-score and the better the solution, the higher the LLG.

MR allows solving any structure bearing a certain degree of homology (above 20 % similarity) to a known model. When the fold is thought to be highly conserved homology can be increased by either truncation of all side chains, thus just using the backbone, thereby removing the noise contribution due to poor

sequence identity to the target. Also highly conserved residues between the target and the model can be maintained increasing the possible LLG score.

3.2.6. Refinement and Validation of Macromolecular Models

Once the initial electron density map has been phased, the aim is to improve the model in iterative steps such that it more accurately agrees with the experimental data. This iterative process aims to minimize the differences between the observed and the calculated structure factors and is called refinement. Refinement should combine with maximum likelihood and weighted difference maps. This is needed, as the phases have been produced by either experimental methods or MR and are either inaccurate or biased.

Refinement uses two main aspects, geometry of the atoms and agreement between model and electron density. *Coot* (Emsley et al., 2010) is generally used to manually manipulate the structure residue by residue; using the available tools to regularize the stereochemical properties of peptides, such as bond angles, bond lengths and likely side chain rotamers. The Ramachandran plot (Ramachandran and Sasiskharan, 1968), which plots the relative ψ and ϕ angles of each residue highlighting atomic steric hindrance, is used to further improve the geometry of the model. *Coot*, *PROCHECK* (Laskowski et al., 1993) and *MOLPROBITY* (Chen et al., 2010) all offer graphic and tabulated views of the stereochemical properties of the model.

Real space refinement is very specific and involves fitting atoms to electron density (Diamond, 1985), although most refinement is now performed in reciprocal space. Further constraints or restraints can be used during the refinement process depending on the nature of the molecule(s) being refined. These include non-crystallographic symmetry (NCS) constraints. Molecules, which are related chemically, but not by crystallographic symmetry operators are related in refinement terms by NCS. NCS constraints will constrain areas of density that are similar, but will allow other regions to differ. As such, it can be modified to accommodate conformational differences between molecules within the asymmetric unit.

The success of a refinement strategy is judged using the R_{factor} and the R_{free} regardless of the program used. The R_{factor} is a measure of the agreement between the observed structure factors and the calculated structure factors from the model (Equation 2-5). The R_{free} is by far the most important parameter, but there are other factors such as quality of density, B-value distribution, overall fit to the electron density and co-ordinate error.

Equation 2-5: R_{free} and R_{factor}

$$R = \frac{\sum_{hkl}(|F_{obs}| - |F_{calc}|)}{\sum_{hkl}|F_{calc}|}$$

where

F_{obs} are the observed structure factors
F_{calc} are the calculated structure factors

Using this equation (Equation 2-5), a R_{factor} should be below 50%, otherwise the input model is not any more significant than a random model. There are two types of R-factors that measure the progress of refinement: the R_{factor} and the R_{free} (Brunger, 1992). The R_{factor}'s purpose is to give a sense of the agreement between the diffraction data and the model. The data used to calculate the R_{factor} is derived from the same data that is being refined, it is prone to over fitting. Thus a more sensitive measure of phase error is the R_{free} (Brunger, 1992) as it is calculated using 5% of the experimental reflection that are not used for the refinement and hence is free from model bias. A convergence between the two R-factors of less then 5% is aimed for towards the end of the refinement. At this point the model is said to be ready for deposition in the PDB, providing that the stereochemistry is in the acceptable parameters from e.g. PROCHECK.

2.2. Fluorescence Anisotropy Background Theory

3.2.1. Fluorescence Anisotropy

Fluorescence Anisotropy (FA) is a well-established method for studying binding events between proteins and other macromolecules, such as RNA (Shi and

Herschlag, 2009, Singh et al., 2000). Using fluorescence anisotropy measurements to establish dissociation constants between proteins and other macromolecules is useful as anisotropy is independent of the overall protein concentration and varies with the rotational correlation time (Jameson and Sawyer, 1995). The rotational correlation time is linked to the rotation rate of the molecule, which changes upon binding (Lakowicz, 2006). The rotational correlation time is related to the viscosity of the solvent (η), the molecular volume (V), the gas constant (R) and the temperature (T in kelvin).

When a fluorophore is excited by polarized light the light emitted is also polarized. The degree to which this light is polarized is described as anisotropy (r) (Perrin, 1926, Weber, 1953). The emission can become depolarized by a number of processes, such as binding events, which in turn will change the measured anisotropy (Cantor and Schimmel, 1980).

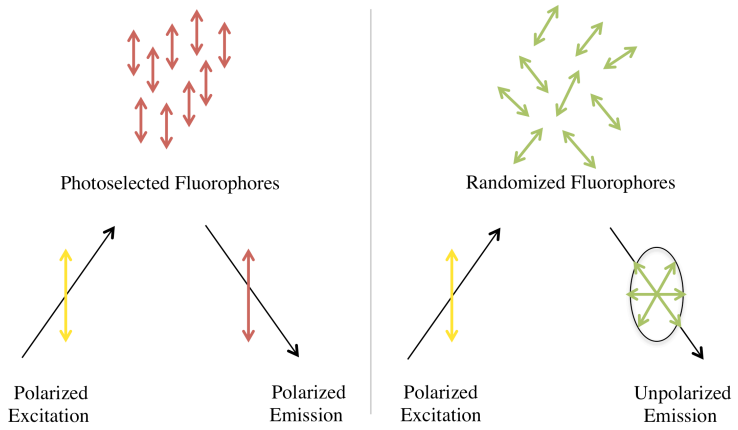


Figure 2-10: Effects of Polarized Excitation and Rotational Diffusion on the Polarization or Anisotropy of the Emission

As a consequence of excitation, there is an angular displacement between absorption and subsequent emission of the photons, due to the rotational diffusion of the molecule. The anisotropy measurements therefore depict the average angular displacement of the excited fluorophore population. This angular displacement is dependent upon the rate and extent of rotational diffusion during the lifetime of the excited state (Perrin, 1926, Weber, 1953, Weber and Hercules, 1966). The rate of rotational diffusion depends on the viscosity of the solvent and the size and shape of the rotating molecule.

In a homogeneous solution, fluorophores are randomly oriented in the ground-state. Following irradiation with polarized light, those that have their absorption transition moments aligned with the electric vector of the light are preferentially excited and are thus no longer randomly oriented; they are then referred to as the excited-state population (Jameson and Sawyer, 1995, Cantor and Schimmel, 1980) (Figure 2-10).

When the fluorophores freely rotate before re-emitting the photons, the degree of polarization of the emitted light will be reduced compared to the original light used for excitation. This difference in anisotropy depends on the fluorophores rotational correlation time (θ) during the fluorescence lifetime (τ) (Alcala et al., 1987, Perrin, 1926, Weber and Hercules, 1966). Assuming no other processes result in loss of anisotropy, the expected anisotropy is given by the Perrin equation (Equation 2-6).

Equation 2-6: Perrin Equation

$$r = \frac{r_0}{1 + \tau/\theta}$$

where

r is the observed anisotropy

r_0 is the intrinsic anisotropy of the molecule

τ is the fluorescence lifetime

θ is the rotational correlation time for the diffusion process

3.2.2. Fluorescence Anisotropy Apparatus Setup

In the experimental set-up, the sample is excited with vertically polarized light. The electric vector of this light is oriented parallel to the vertical (z-axis) (Figure 2-11). The intensity of the emission is measured through a polarizer. When the emission polarizer is oriented parallel (VV) to the direction of the polarized excitation, the observed intensity is I_{VV}. Likewise, when the polarizer is perpendicular (VH) to the excitation, the intensity is I_{VH}. These intensity values are used to calculate the anisotropy experimentally (Equation 2-7) (Lakowicz, 2006, Jameson and Sawyer, 1995).

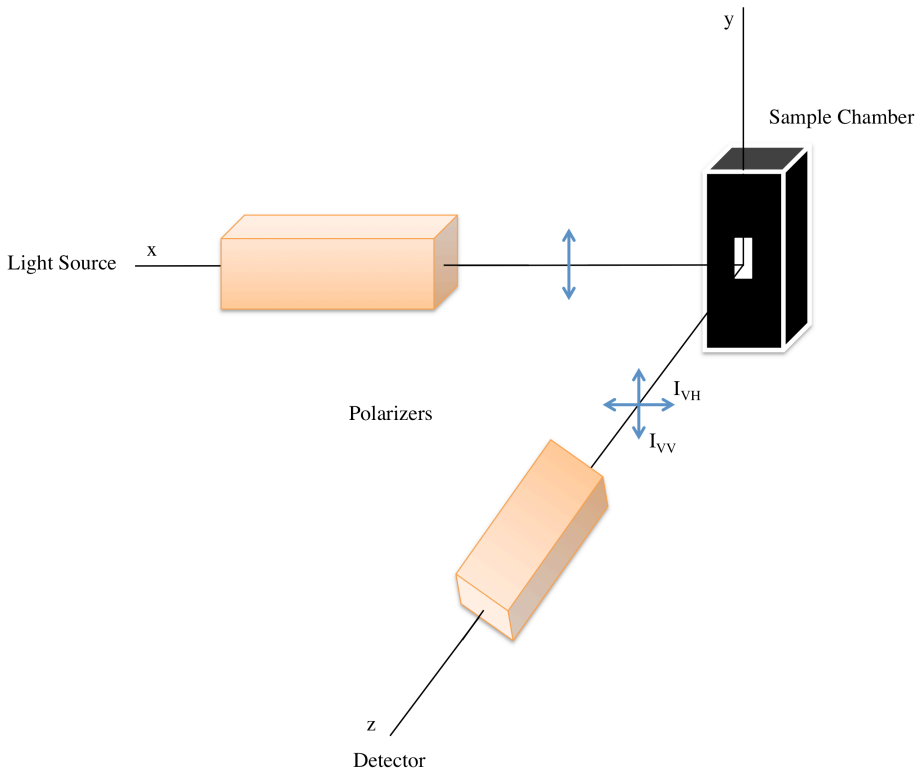


Figure 2-11: Experimental Setup of Fluorescence Anisotropy Apparatus

Equation 2-7: Anisotropy

$$r = \frac{I_{VV} - I_{VH}}{I_{VV} + 2I_{VH}}$$

where

r is the observed anisotropy

I_{VV} is the observed intensity at VV

I_{VH} is the observed intensity at VH

In order to obtain the dissociation constants, the change in anisotropy is measured for samples in a titration series where the ligand concentration is kept constant, but the protein concentration increased (Lakowicz, 2006, Pollard, 2010). The dissociation constant (K_d) is the equilibrium constant of the reversible propensity of a complex to dissociate into 2 components. The K_d (M) corresponds to the concentration of ligand needed so that half of the binding sites of the protein are occupied. Which means, the smaller the K_d value is the higher the affinity of the ligand for the protein. (Equation 2-8) (Pollard, 2010).

Equation 2-8: Dissociation Constant - K_d

$$K_d = \frac{[P][L]}{[C]}$$

where

K_d the dissociation constant (M)

$[P]$ the molar concentration of the Protein (M)

$[L]$ the molar concentration of the Ligand (M)

$[C]$ the molar concentration of the Complex (M)

2.3. Microscale Thermophoresis Background Theory

3.2.1. Microscale Thermophoresis

Microscale thermophoresis (MST) is a recently pioneered method allowing measurement of the dissociation constant of macromolecules in solution on the microliter scale. Thermophoresis is the movement of molecules along a temperature gradient (Duhr and Braun, 2006). Thermodiffusion is labelled "positive" when particles move from a hot to cold region and "negative" when the reverse is true. Thermophoresis depends on changes in size, charge and the solvation shell of molecules. These effects are monitored in a capillary setup outlined in figure 2-12. Prior to thermophoresis, a homogenous molecular distribution is observed inside the capillary. By focusing an Infrared (IR)-Laser onto a specific point on the capillary, a microscopic temperature gradient of 2K-6K (Kelvin) is created (Zillner et al., 2011). The macromolecules respond to this change by moving from the locally heated region to the outer cold regions and thus the concentration of macromolecules in the locally heated region decreases until it reaches a steady state determined by mass diffusion. The movement can also occur towards the locally heated region. This change in concentration can be quantified using the Soret coefficient (S_T) (Equation 2-9) (Duhr and Braun, 2006, Reineck et al., 2010). The Soret coefficient takes into account the size, charge and hydration shell of the macromolecule to be studied at a set temperature.

Equation 2-9: The Concentration Ratio under Thermophoresis

$$\frac{C_{hot}}{C_{cold}} = e^{(-S_T \cdot \Delta T)}$$

where

C_{hot} is the observed concentration of the labelled macromolecule at increased temperature (in μM)

C_{cold} is the observed concentration of the labelled macromolecule at normal temperature (in μM)

S_T is the Soret coefficient (in K^{-1})

ΔT is the difference in temperature between C_{hot} and C_{cold} (in K)

3.2.2. Microscale Thermophoresis Apparatus Setup

In a typical experiment, either the macromolecule of interest or its ligand is labelled with a fluorophore (Figure 2-12). The fluorescent molecules thermophoretic movement is subsequently followed using the fluorescence distribution F inside a capillary (Baaske et al., 2010). As the thermophoresis is monitored using the fluorescence of the labelled molecule, any changes in its thermophoretic mobility on addition of its ligand or target can only arise from changes in the size, charge or solvation entropy due to binding, as the buffer is kept constant (Wienken et al., 2010, Baaske et al., 2010, Duhr and Braun, 2006).

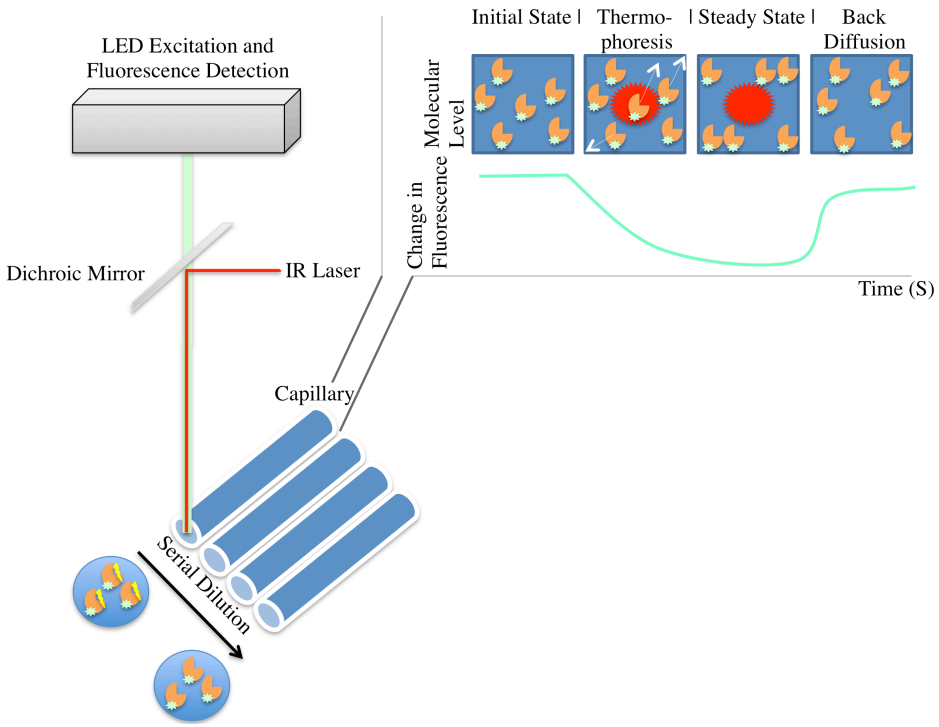


Figure 2-12: Experimental Setup of the Microscale Thermophoresis Apparatus

The K_d is obtained from a serial dilution of the binding substrate. By plotting the measured fluorescence against the logarithm of the different concentrations of the dilution series, a sigmoidal binding curve is obtained. This binding curve can be directly be fitted with the nonlinear solution of the law of mass action, which gives the K_d.

2.4.Aims

To better understand how KSHV infections lead to global and rapid mRNA decay the importance of the UGAAG motif in sequence terms, the involvement of RNA secondary structure and UGAAGs motifs involvement in the secondary structure were investigated via computational, biochemical and biophysical means.

Chapter 3: Materials and Methods

3.1. Computational Analysis

3.1.1. Distribution of UGAAG Motif in Host and Viral Genomes

The EMBOSS web server tool *compseq* (Williams, 2001) was used to calculate the actual and expected frequencies of the UGAAG motif and inverse GAAGU motif in a file containing all mRNAs from the human genome. The files refseq_HHV8 (Rezaee et al., 2010) and refseq _HS (Pruitt et al., 2013) were used, containing respectively the KSHV genes and *Homo sapiens* mRNAs. The *Homo sapiens* mRNAs were retrieved by using the following search parameters (srcdb_refseq[prop] AND biomol_rna[prop] AND ("last 30 days"[PDAT])) AND "Homo sapiens"[porgn:__txid9606].

3.1.2. Alignment and *in silico* Folding of mRNAs

The full mRNA sequences for GFP, DsRed2 and HBB from Genbank (Benson et al., 2012) that were identified by the Glaunsinger Group as targets for SOX were obtained (Covarrubias et al., 2011) (Table 3-1). These were checked against the sequence published by Glaunsinger using nucleotide basic local alignment search tool *(*Blast) (Altschul et al., 1990), as no accession numbers were mentioned in the publication. The three sequences were aligned using the web server *MView* (Brown, 2013, Brown et al., 1998) and *T-coffee* (Notredame, 2013, Notredame et al., 2000).

A cycle of *in silico* RNA folding was then performed using the web server *mfold* (Zuker, 2003a). First, the whole sequence was folded and the number of possible folds in the region targeted for cleavage quantified, thereafter restricting the sequence to 201 nucleotides as used by Covarrubias et al., 2011. Finally, the sequences were truncated to a length that would maintain the identified characteristic fold and would be amenable to biochemical and structural studies.

Table 3-1: Genbank IDs of mRNAs with an Identified Endonucleolytic Cleavage Site

Target Name	Genbank ID
GFP	GI:371926914
DsRed2	AAY25372.1
HBB	GI:28302128

Mutant and shorter engineered variants of the GFP construct were also folded using this protocol. Further RNA sequences and mutant GFP RNA sequences were subsequently *in silico* folded and analysed as mentioned above; the sequence names, accession numbers and compositions can be found in table 3-3 in section 3.2.2.1.

3.1.3. Tertiary Structure Prediction

The web server *MC-Fold | MC-Sym* (Parisien and Major, 2008) was used to obtain tertiary structure predictions for the RNA folds identified . *MC-Fold* | MC-Sym explores probabilistically the conformational space for RNA using input constraints and taking into account Watson-Crick and non Watson-Crick base pairing. The RNA sequences containing the UGAAG motif that had been sequentially reduced in length were used as input (Table 3-1) into *Mc-Fold* (http://www.major.iric.ca/MC-Fold/) for MC-Sym (http://www.major.iric.ca/MC-Sym/) to obtain the 3D RNA structures.

3.1.4. Fitting the RNA in the Active Site of SOX

The 114 tertiary structures for the 51 nucleotides RNA sequence of GFP was then used to model the interaction between SOX and the RNA using as a template the crystal structure of SOX in complex with dsDNA (PDB ID: 3pov) (Bagneris et al., 2011). The *Chimera software* (Pettersen et al., 2004) was used for modelling. The density for the DNA derived from the SOX-DNA structure was used as a template to fit the RNA stem loop into the active site. Atomic clashes were taken into consideration where the most favourable model was retained.

3.1.5. Generation of SOX Conformers using t*CONCOORD*

tCONCOORD (Seeliger and DeGroot, 2009) was used to produce alternative conformations for the native wild type protein (3fhd). t*CONCOORD* builds a library of distance constraints based on the observed interatomic distances in the original structure. Interactions deemed to be stronger are given tighter constraints. The program then produces randomly a large number of potential conformations, and attempts to correct structures with atom-pair distances falling outside the allowed regions. 1000 iterations were applied of the correction algorithm per structure, and the structures were rejected whose interatomic distances violated the original distances by more than 3 nm in total. t*CONCOORD* was set to an output of 250 novel conformations for the native wild type protein (3fhd), which fulfilled the distance constraints. All computationally produced conformers were superimposed on the native wild type (3fhd) using the structalign program.

3.2. Biophysical and Biochemical Characterisation of SOX, Xrn1, SOX:RNA and SOX:Xrn1 Interactions

3.2.1. Protein Expression and Purification

3.2.1.1. Plasmid Purification and Quantitation

3.2.1.1.1. SOX: Plasmid Purification and Quantitation

Plasmid stocks were prepared from 5 mL lysogeny broth-luria (LB) (10 g/L tryptone; 5 g/L yeast extract; 10 g/L sodium chloride at pH 7.5, autoclaved) cultures of *NEB 5-alpha Competent E. coli* grown overnight at 37 °C. DNA purification was conducted using the Wizard® Plus SV Minipreps DNA Purification System (Promega) using the standard operating protocol for centrifugation.

The quantity and purity of DNA yield from plasmid purification was assessed using the 'Nucleic Acid' program in the NanoDrop software using a NanoDrop 1000 spectrophotometer (Thermo Scientific) and the Promega DNase free buffer as blank. Concentrations were measured in μg/μL.

3.2.1.1.2. Xrn1: Plasmid Purification and Quantitation

The same protocol as in section 3.2.1.1.1 was applied to the Xrn1 containing plasmid; pET26b-Xrn1.

3.2.1.2. Transformation of Chemically Competent Cells by Heat Shock

3.2.1.2.1. SOX: Transformation of Chemically Competent Cells by Heat Shock

The SOX plasmid pETM6T1-SOX (Bagneris et al., 2011) was co-transformed into BL21(DE3) Star (Invitrogen) with the plasmid pRARE encoding the rare tRNAs extracted from RosettaTM 2(DE3) cells (Novagen). The plasmid pETM6T1-SOX contained an N-terminal His-tag, followed by a N-utilization substance protein A (NusA) Tag with a tobacco etch virus protein (TEV) cleavage site linker to the full length SOX gene (See Appendix B for plasmid details).

Wild type or mutant SOX proteins (A61T, D221S, E244S, Y373A, H450A, R451A, N458A, R462A, D474N, Y477Stop) were expressed and purified using the same protocol. The mutants were previously produced by Dr. Claire Bagnéris from the Barrett Group.

Transformation of chemically competent of the *E. coli* strain for expression was conducted as follows: 50 µL of competent BL21(DE3) Star cells were thawed on ice for 10 minutes. After thawing, 1 µL of each pRARE (containing a chloramphenicol resistance gene and coding for rare tRNAs) and pETM6T1-SOX (containing a kanamycin resistance gene) plasmids were added and the mixture incubated on ice for 30 minutes. The cells were rapidly transferred to a water bath at 42°C for 30 seconds, then transferred back on ice for 5 minutes. 200 µL of room temperature SOC media were then added to the competent cell mix. The media and transformed cells were placed in an incubator for 1 hours (h) at 37°C and 250 rpm. 250 µL of recovered cells were then spread onto pre-prepared agar plates containing kanamycin at 25 µg/mL and chloramphenicol at 34 µg/mL final concentrations. Plates were incubated at 37°C overnight.

3.2.1.2.2. Xrn1: Transformation of Chemically Competent Cells by Heat Shock

The Xrn1 plasmid pET26b-Xrn1 (a kind gift from Professor Liang Tong) was co-transformed into BL21(DE3) Star (Invitrogen) with a plasmid encoding the rare tRNAs extracted from RosettaTM 2(DE3) cells (Novagen). The plasmid pET26b-Xrn1 contained residues 1–1,245 of *K. lactis* Xrn1 with a C-terminal hexahistidine tag.

The same protocol as in section 3.2.1.2.1 was applied for the transformation of pET26b-Xrn1.

3.2.1.3. Recombinant Expression

3.2.1.3.1. SOX: Recombinant Expression

A colony of BL21(DE3) Star containing the pETM6T1-SOX and pRARE plasmids was picked from the agar plate and used to inoculate 100 mL of LB to which chloramphenicol (34 μg/mL) and kanamycin (25 μg/mL) were added to yield a seed culture, which was grown overnight at 37°C, at 225 rpm (Certomat BS1 Shaker incubator). 5 mL of this culture were then inoculated into each of 12 x 500 mL volumes of LB/chloramphenicol/kanamycin in 2 L shaker flasks. These cultures were grown as for the seed culture until they reached an optical density at 600 nm (OD_{600}) of 0.8. At this stage, synthesis of recombinant SOX was induced by the addition of 1 mM isopropyl-b-D-thiogalactopyranoside (IPTG) and the cultures were grown overnight at 18 °C. Cells were harvested by centrifugation (5000 *g*, 40 mins, 4 °C on the Beckman Avanti J-20 I rotor then at 4000 *g* for 20 minutes at 4 °C on the Hettich Rotina 420R to collect the cell pellets in 50 mL centrifuge tube for snap freezing). The pellet was stored at -80 °C.

3.2.1.3.2. Xrn1: Recombinant Expression

The expression protocol was adapted from previous publications (Bagneris et al., 2011, Chang et al., 2011). Recombinant Xrn1 was overexpressed in *E. coli* BL21(DE3) Star. 50 μL of these cells stored in 10% glycerol were inoculated into 100 mL of LB (10 g/L tryptone; 5 g/L yeast extract; 10 g/L sodium chloride at pH

7.5, autoclaved), chloramphenicol (34 μg/mL) and kanamycin (25 μg/mL), and a seed culture grown overnight (37°C, 225 rpm - Certomat BS1 shaker incubator). 5 mL of this culture were then inoculated into each of 12 x 500 mL volumes of LB/ chloramphenicol/kanamycin in 2 L shaker flasks. These cultures were grown as for the seed culture until they reached an OD_{600} of 0.8 . At this stage, synthesis of recombinant Xrn1 was induced by the addition of 0.5 mM IPTG and the cultures were grown for 4 hrs at 20 °C. Cells were harvested by centrifugation (5000 *g*, 40 mins, 4 °C on the Beckman Avanti J-20 I rotor then at 4000 *g* for 20 minutes at 4 °C on the Hettich Rotina 420R to collect the cell pellets in 50 mL centrifuge tube for snap freezing). The pellet was stored at -80 °C.

3.2.1.4. Purification Protocols

3.2.1.4.1. SOX: Purification

All chromatography columns described in this thesis were used according to the manufacturer's specifications. All were attached to an ÄKTA automated chromatography system for purification using the UNICORN control software.

Samples were loaded on to equilibrated chromatography columns (except the gel filtration column) using a peristaltic pump at flow rates between 0.5 and 1.0 mL/min and the unbound fractions were collected. For those involving gel filtration, samples were loaded using the ÄKTA inbuilt syringe injection loop system.

The purification protocol was adapted from previous publications (Bagneris et al., 2011). Cell pellets were first resuspended in nickel buffer A (300 mM NaCl, 25 mM Tris pH 7.2) supplemented with DNase I (10 μg/mL final concentration (NEB)) and an Ethylenediaminetetraacetic acid (EDTA)-free protease inhibitor cocktail tablet (Roche). After the cells were lysed using a cell disruptor system (3C High Pressure Homogeniser) on ice, the lysates were clarified by centrifugation (46 000g for 1 h at 4 °C, Beckman Avanti J-20 XP rotor) and the supernatant filtered through a 0.45mm filter prior to loading onto two 5 mL HisTrap FF column (GE-Healthcare). The columns were washed with 20 column volumes (CVs) of buffer A containing 50 mM imidazole and the protein eluted using 60 CVs of buffer B (300

mM NaCl, 25 mM Tris pH 7.2, 500mM imidazole). The eluate was diluted to a concentration of 200 mM NaCl and a pH of 8.5 by using a dilution buffer (0 mM NaCl, 25 mM Tris pH 8.5) and then loaded onto a 5 mL HiTrap Q HP column (GE-Healthcare) previously equilibrated in 25 mM Tris, 50 mM NaCl, pH 8.5. The fusion protein was then eluted using a 50-1000 mM NaCl linear gradient. Eluted fractions were assessed for their protein composition by Sodium Dodecyl Sulphate - Polyacrylamide Gel Electrophoresis (SDS-PAGE) gels. Fractions containing protein with a molecular weight consistent with recombinant His-NusA-SOX (114 kDa) were pooled. The tag was removed by the addition of TEV protease to the pooled fractions during overnight dialysis in Pierce snakeskin membrane (3 kDa molecular weight cut off (MWCO)) in a buffer comprising 25 mM Tris pH 8.5, 200 mM NaCl, 1 mM dithiothreitol (DTT). The solution was diluted to a concentration of 200 mM NaCl using pH of 8.5 Tris-HCl buffer and then applied to a 5 mL HiTrap Q HP column. The untagged protein eluted using a 50–500 mM NaCl linear gradient. Fractions containing the purest protein were pooled and concentrated using a 30 kDa cut-off Vivaspin centrifugal concentrator (Vivascience) (3000 *g*, Hettich Rotina 420R). The protein was then loaded onto a gel filtration column (Superdex200 HR 26/60) that had been pre-equilibrated in a buffer consisting of 32 mM Tris pH 8.5, 189 mM NaCl, 1.6 ug bovine serum albumin (BSA) and 10 mM DTT. Fractions were analysed on SDS–PAGE gels and those containing pure protein were concentrated to 4 mg/mL using a 30 kDa cut-off spin cartridge (Vivascience) (3000 *g*, Hettich Rotina 420R), aliquoted and stored at -80 °C.

3.2.1.4.2. Xrn1: Purification

Cell pellets were resuspended in nickel buffer A (20 mM Tris pH 7.5, 100 mM NaCl and 10 mM β-mercaptoethanol) supplemented with DNase I (10 μg/mL final concentration) and an EDTA-free protease inhibitor cocktail tablet (Roche). After lysis using a cell disruptor system (3C High Pressure Homogeniser) on ice, the lysates were clarified by centrifugation (46 000g for 1 h at 4 °C, Beckman Avanti J-20 XP rotor) and the supernatant filtered through a 0.45mm filter prior to loading onto one 5 mL HisTrap FF column (GE-Healthcare). The column was washed with 20 CVs of buffer A containing 50 mM imidazole and the protein eluted using 20 CVs of buffer B (20 mM Tris (pH 7.5), 100 mM NaCl, 10 mM β-

mercaptoethanol and 500 mM imidazole). Fractions containing the purest protein were pooled and concentrated using a 30 kDa cut-off Vivaspin centrifugal concentrator (Vivascience) (3000 *g*, Hettich Rotina 420R). The protein was then loaded onto a gel filtration column (Superdex200 HR 26/60) that had been pre-equilibrated in a buffer consisting of 20 mM Tris pH 7.5, 200 mM NaCl, 2 mM DTT and 5% (v/v) glycerol. Fractions were analysed using SDS–PAGE gels and those containing protein, concentrated to 4 mg/mL using a 30 kDa cut-off spin cartridge (Vivascience) (3000 *g*, Hettich Rotina 420R), aliquoted, flash frozen with liquid nitrogen and then stored at –80 °C.

3.2.1.5. Quantitation of Protein Yield from Recombinant Expression and Sample Concentration

The protein concentration, for the final eluent pool from the purification and for each experiment, was measured using the 'Protein A280' program in the NanoDrop software using a NanoDrop 1000 spectrophotometer (Thermo Scientific) using the Gel Filtration Buffer as blank. Concentrations were measured in mg/mL.

Each sample was measured three times and the average value of these was then used in the calculation of the protein concentration.

The protein concentration was calculated using the rearranged Beer-Lambert equation (Equation 3-1).

Equation 3-1: Rearranged Beer-Lambert Equation

$$[C]\ (\text{mg/mL}) = \left(\frac{A_{280}}{\varepsilon_{0.1\%}.l}\right) \Rightarrow \left(\frac{A_{280}}{\varepsilon_{0.1\%}}\right)$$

where

[C] is the protein concentration in mg/mL

$\varepsilon_{0.1\%}$ is the wavelength-dependent molar absorptivity coefficient for each specific protein in $(\text{mg/mL})^{-1}\text{cm}^{-1}$

A_{280} is the measured absorbance value at a wavelength of 280 nm

l is the path length (cm), which was set to 1 cm

3.2.1.6. Analysis of Proteins by Sodium Dodecyl Sulphate-Polyacrylamide Gel Electrophoresis (SDS-PAGE)

The purified protein was tested for its identity and its purity using SDS-PAGE gel analysis. The purified protein was positively identified on the basis of size and in comparison to previous purified samples. The SDS-PAGE gels were either pre-cast Invitrogen, 1 mm, 4-12% (v/v) acrylamide Bis-Tris SDS gels or hand-cast 1 mm SDS-PAGE gels, see table 3-2 for composition. The reagents for the hand-cast SDS-PAGE gels were combined to produce the gel by adding tetramethylethylenediamine (TEMED) and 10% Ammonium Persulfate (APS) solutions last prior to pouring.

Both types of gels were prepared and run as instructed in the Invitrogen manufacturer's guide book for SDS-PAGE analysis using NuPAGE® MES (2-(N-morpholino)ethanesulfonic acid) SDS Running Buffer. The pre-cast gels were run in an Invitrogen-specific tank and the hand-cast gels in the Tetra Electrophoresis system (BioRad).

Table 3-2: Composition of Hand-cast Polyacrylamide Gels for SDS-PAGE

	Resolving Gel	Stacking Gel
Gel Percentage (%)	11.4%	4.0%
Reagent		
30% Polyacrylamide Solution	3.8 mL	670 µL
1.5 M Tris-HCl pH 8.8	2.5 mL	-
1.0 M Tris-HCl pH 6.8	-	630 µL
10% APS	100 µL	50 µL
10% SDS	100 µL	50 µL
dH_2O	3.5 mL	3.6 mL
TEMED	12 µL	8 µL
Total Volume (mL)	10.0 mL	5.0 mL

The crude soluble, insoluble and flow through fractions were diluted to 1:20 to a volume of 12 µL to which 3 µL of 5X NativePAGE™ sample loading buffer (Invitrogen) were added. 12 µL of partially purified protein samples were used in a 1:1 volumetric ratio of sample to which 3 µL of 5X NativePAGE™ sample loading buffer were added. 15 µL were loaded into the wells of SDS-PAGE gels and 5-10 µL page ruler prestained protein ladder (Thermo Scientific) were used as molecular weight markers. Electrophoretic separation was undertaken using a PowerPac™

basic power supply (BioRad) at 180 V typically for 35-40 minutes depending on the percentage of the resolving gel. Gels were stained with Instant Blue (Expedeon). Gels were visualised on a light box or scanned for electronic visualization.

3.2.2. SOX:RNA Binding and Activity Assays

3.2.2.1. RNA Preparation

The RNA oligonucleotide sequences (Table 3-3) were purchased from Eurogentec (Belgium). The lyophilized RNA was dissolved in RNA-annealing buffer (100 mM NaCl, 20 mM $MgCl_2$, 20 mM Tris HCL at pH 7.4; RNase and DNase Free) to a concentration of 1 mM. The RNA was annealed using the Peqlab Primus 96 Gradient PCR-machine by heating the mixture to 90 °C for 1 minute, followed by a decrease of 1 °C every minute until 4 °C was reached.

All RNAs were synthesized, so that they did not contain a 5' monophosphate or fluorescent tag, since they would then be targets for single stranded exonucleolytic cleavage by SOX. For the Tris/Borate/EDTA (TBE), TBE-Urea and Fluorescence anisotropy experiments, all synthesised RNAs were substituted with a 3' *6*-carboxyfluorescein (6-FAM) label. The sequences used in the studies described are given in table 3-3.

Table 3-3: RNA Sequences of Identified Structured Target Fold

RNA Name	RNA Sequence	Purification Method
51 GFP	5'-UAC-GGC-AAG-CUG-ACC-CUG-AAG-UUC-AUC-UGC-ACC-ACC-GGC-AAG-CUG-CCC-GUG- 3'	HPLC-IEX
51 GFP UCUCU	5'-UAC-GGC-AAG-CUG-ACC-CUC-UCU-UUC-AUC-UGC-ACC-ACC-GGC-AAG-CUG-CCC-GUG-3'	HPLC-IEX
23 GFP	5'-AGC-UGA-AGU-UCA-UCU-GCA-CCA-GC-3'	HPLC-IEX
58 HBB	5'-AGG-UGA-AGG-CUC-AUG-GCA-AGA-AAG-UGC-UCG-GUG-CCU-UUA-GUG-AUG-GCC-UGG-CUC-ACC-U-3'	HPLC-IEX

N.B.: 51 GFP UCUCU is a construct in which the UGAAG motif was mutated into UCUCU.
23 GFP is a construct in which the original GFP sequence was shorted.
High-Performance Liquid Chromatography Ion Exchange (HPLC-IEX)

3.2.2.2. TBE Gel – RNA Electrophoretic Mobility Shift Assays

Electrophoretic mobility shift assays were performed to assess protein-RNA binding using a TBE gel system. The TBE gels were either pre-cast (Invitrogen), 1 mm, 6% (v/v) acrylamide TBE gels or hand-cast 1 mm TBE gels, see table 3-4 for composition. The reagents for the hand-cast TBE gels were combined to produce the gel by adding TEMED and 10% APS solutions last prior to pouring (Hellman and Fried, 2007).

Both types of gels were prepared and run as instructed in the Invitrogen manufacturer's guide book for TBE analysis using 4% ficoll as a loading buffer and Orange-G in a separate lane, as a running marker. 1 times TBE (89 mM Tris-Base, 89 mM Boric Acid, 2 mM EDTA) buffer was used as running Buffer. The pre-cast gels were run in an Invitrogen-specific tank and the hand-cast gels in the Tetra Electrophoresis system (BioRad) (Hellman and Fried, 2007).

Table 3-4: Composition of Hand-cast TBE Gels

	Resolving Gel
Gel Percentage (%)	4.0%
Reagent	
30% Polyacrylamide Solution	1.6 mL
5X TBE	2.4 mL
10% APS	120 µL
dH_2O	7.8 mL
TEMED	14.4 µL
Total Volume (mL)	12.0 mL

Binding assays were performed using the structured 51 nucleotides GFP, mutated UCUCU GFP and 58 nucleotides HBB RNA (Table 3-3 for sequences). 20 pmol of RNA were incubated with 150 pmol of SOX for 1 h in a buffer (Table 3-5) at room temperature. First the optimal binding condition was assessed using a range of NaCl concentration and pH range (Table 3-5), thereafter buffer B. TBE D was used. Subsequently, 4% ficoll and 1 mM EDTA were added to each sample prior to loading onto 6% TBE gels or hand-cast 4% TBE gels (Invitrogen) (Buisson et al., 2009, Bagneris et al., 2011, Hellman and Fried, 2007).

15 µL of RNA, EDTA and ficoll mixture were loaded into the wells of TBE gels and 10 µL of 4% ficoll. A solution of 2 % Orange-G was used as a running marker. Electrophoretic separation was undertaken using a PowerPac™ basic power supply (BioRad) at 50 V typically for 120-180 minutes depending on the percentage of the gel. Gels were scanned using a FLA300 (Fujitsu) Imager, at an excitation wavelength (λex) of 490 nm and emission at λem of 520 nm. 6-FAM has an excitation and emission wavelength of 492 nm and 517 nm respectively.

Table 3-5: TBE Binding Buffers

Buffer Name	Buffer Composition	pH
B. TBE St.	50 mM Tris–HCl, 200 mM NaCl, 20 mM, β-mercaptoethanol, 50 mM EDTA, 10 mM MgCl2	9
B. TBE A	100 mM NaCl, 50 mM bis-tris, 15 mM DTT, 50 mM EDTA	6
B. TBE B	100 mM NaCl, 50 mM bis-tris, 15 mM DTT, 50 mM EDTA	6.5
B. TBE C	100 mM NaCl, 50 mM bis-tris, 15 mM DTT, 50 mM EDTA	7
B. TBE D	100 mM NaCl, 50 mM Tris, 15 mM DTT, 50 mM EDTA	7.5
B. TBE E	100 mM NaCl, 50 mM Tris, 15 mM DTT, 50 mM EDTA	8.5
B. TBE F	300 mM NaCl, 50 mM bis-tris, 15 mM DTT, 50 mM EDTA	6
B. TBE G	300 mM NaCl, 50 mM bis-tris, 15 mM DTT, 50 mM EDTA	6.5
B. TBE H	300 mM NaCl, 50 mM bis-tris, 15 mM DTT, 50 mM EDTA	7
B. TBE I	300 mM NaCl, 50 mM Tris, 15 mM DTT, 50 mM EDTA	8.5
B. TBE J	400 mM NaCl, 50 mM bis-tris, 15 mM DTT, 50 mM EDTA	6
B. TBE K	400 mM NaCl, 50 mM bis-tris, 15 mM DTT, 50 mM EDTA	6.5
B. TBE L	400 mM NaCl, 50 mM bis-tris, 15 mM DTT, 50 mM EDTA	7
B. TBE M	400 mM NaCl, 50 mM Tris, 15 mM DTT, 50 mM EDTA	8.5

3.2.2.3. TBE-Urea Gel – RNA Electrophoretic Activity Assays

Endonuclease activity assays were performed using TBE-Urea gel analysis. The assays were performed using the structured 51 nucleotides GFP, mutated UCUCU GFP and 58 nucleotides HBB RNA. 20 pmol of RNA were incubated with 150 pmol of SOX for 1 h in a buffer (Table 3-6) at 37 °C. First the optimal binding condition was assessed using a range of NaCl concentration and pH range (Table 3-6), then the B. TBE St.' Buffer was subsequently used. The RNA reactions were halted by the addition of 50 mM EDTA and 7.5 µl of each reaction mixture were combined with 7.5 µl of Novex TBE-Urea sample buffer (Invitrogen), heated at 70°C for 3 mins, prior to loading onto pre-cast 1 mm, 15% TBE-Urea gels

(Invitrogen) (Hellman and Fried, 2007, Buisson et al., 2009, Bagneris et al., 2011). Electrophoretic separation was undertaken using a PowerPac™ basic power supply (BioRad) at 100 V typically for 120 minutes. Gels were scanned using a FLA300 (Fujitsu) Imager, at an excitation wavelength (λex) of 490 nm and emission at λem of 520 nm.

Table 3-6: TBE-Urea Buffers

Buffer Name	Buffer Composition	pH
B. TBE St.'	50 mM Tris–HCl, 200 mM NaCl, 20 mM β-mercaptoethanol, 10 mM MgCl2	9
B. TBE A'	100 mM NaCl, 50 mM bis-tris, 15 mM DTT, 50 mM EDTA	6
B. TBE B'	100 mM NaCl, 50 mM bis-tris, 15 mM DTT, 50 mM EDTA	6.5
B. TBE C'	100 mM NaCl, 50 mM bis-tris, 15 mM DTT, 50 mM EDTA	7
B. TBE D'	100 mM NaCl, 50 mM Tris, 15 mM DTT, 50 mM EDTA	7.5
B. TBE E'	100 mM NaCl, 50 mM Tris, 15 mM DTT, 50 mM EDTA	8.5
B. TBE F '	300 mM NaCl, 50 mM bis-tris, 15 mM DTT, 50 mM EDTA	6
B. TBE G'	300 mM NaCl, 50 mM bis-tris, 15 mM DTT, 50 mM EDTA	6.5
B. TBE H'	300 mM NaCl, 50 mM bis-tris, 15 mM DTT, 50 mM EDTA	7
B. TBE I'	300 mM NaCl, 50 mM bis-tris, 15 mM DTT, 50 mM EDTA	8.5
B. TBE J'	400 mM NaCl, 50 mM bis-tris, 15 mM DTT, 50 mM EDTA	6
B. TBE K'	400 mM NaCl, 50 mM bis-tris, 15 mM DTT, 50 mM EDTA	6.5
B. TBE L'	400 mM NaCl, 50 mM bis-tris, 15 mM DTT, 50 mM EDTA	7
B. TBE M'	400 mM NaCl, 50 mM Tris, 15 mM DTT, 50 mM EDTA	8.5

3.2.2.4. Fluorescence Anisotropy – SOX:RNA Interaction (RNA K_d)

Fluorescence anisotropy assays were conducted at 25 °C using a fluoromax-3 spectrofluorimeter (Jobin Yvon Horiba), and the data fitted using GraFit (Erithacus Software). Serial dilutions of SOX were incubated for 30 min with 50 nM of structured 51 nucleotides RNA, in a binding buffer comprising 25 mM Tris–HCl pH 8.5, 200 mM NaCl and 10% glycerol in a final volume of 60 μL (Bagneris et al., 2011). These were then transferred to 60 μL quartz cuvettes for anisotropy measurements that were repeated six times at each concentration of SOX. Experiments were conducted with a slit width of 5 nm with the excitation (λex: 492 nm) and emission (λem: 515 nm). The changes in anisotropy following titration of SOX into the RNA were used to calculate the affinity constants. All data were fitted to a one site binding equation.

3.2.3. SOX:Xrn1 Interaction

3.2.3.1. Pull Down Assay – SOX:Xrn1 Interaction

Purified his-tagged Xrn1 was incubation with a 4 times excess of purified nonhis-tagged SOX in buffer A (20 mM Tris pH 7.5, 100 mM NaCl and 10 mM β-mercaptoethanol). The solution was loaded onto a 1 mL HisTrap column (GE Healthcare) and the protein-complex eluted using a 0-500mM Imidazole linear gradient in buffer A over 20 CVs. The complex eluted at 500 mM imidazole, where the presence of both Xrn1 and SOX was verified via SDS-Page Gel (Pollard, 2010).

3.2.3.2. Microscale Thermophoresis – SOX:Xrn1 Interaction (RNA K_d)

A titration series of 16 dilutions was prepared, where KHSV-SOX was kept constant at 20 nM, while Xrn1's concentration was varied from 2 nM - 17.5 μM. KHSV-SOX was fluorescently labelled with blue n-hydroxysuccinimide (NanoTemper Technologies) that specifically targets lysine. 10 μl of the serial dilution of the non-labelled molecule were mixed with 10 μl of the diluted fluorescently labelled molecule (Jerabek-Willemsen et al., 2011, Wienken et al., 2010). Mixed samples were loaded into glass capillaries and the MST-analysis was performed using a Monolith.N115 Series spectrometer (NanoTemper Technologies).

3.2.4. Crystallization of SOX complexes

3.2.4.1. SOX WT/SOX 244:RNA

The RNA oligonucleotides 51 nucleotides GFP, 23 nucleotides GFP and 58 nucleotides HBB were added to SOX WT/SOX 244 at a concentration of 2 mg/mL and concentrated to 8 mg/mL. The resulting RNA-SOX WT/SOX 244 complexes were formed in a 1.2:1 ratio and left to incubate for 1 h at 4 °C. Complexes were also formed by adding the RNA to SOX WT/SOX 244 at high concentration to yield complexes at 8 mg/mL without the requirement for further concentration. As previously described, all complexes were subjected to crystallisation trials using commercially available, sparse matrix suites (Proplex and JCSG+ from Molecular

dimensions) in sitting drop 96 well plates via vapour diffusion at 16 °C (100 nL drops in a 1:1 and 1:2 (v/v) ratios of mother liquor and sample). Preliminary hits were subsequently optimised in 96 well plates using the same complex:precipitant ratios (See Appendix C).

3.2.4.2. SOX WT/SOX 244:Xrn1

The SOX WT/SOX 244 and Xrn1 complex was buffer exchanged into crystallisation buffer (189 mM NaCl, 32 mM Tris pH 8.5, 10 mM DTT and 1.6 ug BSA) at concentrations of 2 and 6 mg/mL respectively and were used in vapour diffusion crystallisation trials as previously described with optimisation performed on initial "hits" (See Appendix C).

3.2.4.3. SOX WT/SOX 244:Xrn1:RNA

The SOX 244 and Xrn1 complex was prepared as above 2.2.4.3. A complex was formed for 51 nucleotides GFP, 23 nucleotides GFP and 58 nucleotides HBB. The SOX 244:Xrn1:RNA complex was used for a sparse matrix screen using commercially available suites (Proplex and JCSG+ from Molecular dimensions) in sitting drop 96 well plates via vapour diffusion at 16 °C with a SOX WT/SOX 244 - Xrn1 (100 nL drops in a 1:1 and 1:2 (v/v) ratios of mother liquor and sample). A fine screen of initial hit conditions was undertaken in 96 well formats using the same drop ratios (See Appendix C).

3.2.4.4. Cryo-cooling Protocol for Macromolecular Crystals

Several cryoprotectant solutions were screened initially for the various complexes, but ethylene glycol was found to be the most effective and therefore used in cryo-cooling all crystals obtained throughout this study. For crystals harvested from commercial screens (JCSG, PACT, Classics or pH screen (Jena Biosciences)), 0.5 μL of reservoir solution was added directly to the crystallisation drop to aid in the recovery of crystals for freezing using appropriately sized nylon or grid loops.

Otherwise, a cryoprotectant solution was made up using the components of the reservoir solution and 20% (w/v) ethylene glycol where 5 µL drops were placed onto 22 mm circular cover slip. Crystals were taken from their drops and soaked in the cryoprotectant for 10-30 seconds by depositing the crystal into the solution, then rapidly recovered into a loop and flash frozen in liquid nitrogen. Crystals were stored in liquid nitrogen until transported to a synchrotron X-ray source.

3.2.4.5. Collection and Processing of Macromolecular Diffraction

The diffraction data were collected at the Diamond Light Source synchrotron. Following data collection autoindexing, data integration, scaling and merging were performed using the *XDS Suite* (Kabsch, 2009). MR for the dataset of the Xrn1 monomer, KSHV WT monomer, SOX 244 monomer, SOX 244 dimer were conducted using *Phaser* (McCoy et al., 2007) as part of the *CCP4 GUI* (Winn et al., 2011) , using the deposited co-ordinates of apo Xrn1 (PDB ID: 3pie) (Chang et al., 2011), of SOX WT with dsDNA (PDB ID: 3pov) (Bagneris et al., 2011) and the apo X-ray crystal structure of SOX WT (PDB ID: 3fhd) (Dahlroth et al., 2009). Refinement was performed using *Buster* (Bricogne, 1993, Bricogne, 1997) and *Phenix* (Adams et al., 2010) and *Coot* (Emsley et al., 2010) was used for manual re-building.

Chapter 4: Computational Analysis of the UGAAG Motif and SOX and RNA Interaction

The computational work was undertaken to investigate the discovery of the Glaunsinger group that a UGAAG motif was found in the vicinity of the endonucleolytic cleavage sites of SOX and the apparent need for this sequence to be encompassed within a 25 to 201 nucleotide stretch. Initial indications seemed to suggest both sequence and structural components were important to RNA processing (Gaglia and Glaunsinger, 2010, Covarrubias et al., 2011). Towards investigating these two elements, it was first established whether this sequence was over or underrepresented in the host and viral genomes, using the program *compseq* from the *EMBOSS software suite*. This analysis was performed to ascertain how prevalent this motif is in human mRNA transcripts and to establish the extent to which KSHV transcripts would also be susceptible to SOX cleavage. The human interleukin-6 (IL-6) gene, known to evade SOX mediated degradation, was specifically analysed for the presence of UGAAG motifs (Hutin et al., 2013, Chandriani and Ganem, 2007). To address the issue of structured elements being key to RNA processing, sequences of the known SOX targets were sequence aligned and analysed for their propensity to fold into secondary and tertiary structures, on the basis of hierarchical folding and local structuring. The most energetically favourable of these folds was then assessed on its ability to be successfully fitted into the active site of SOX. To investigate the roles that the UGAAG motif played within the sequence, a UCUCU mutant was created. A shorter GFP stem loop was engineered for crystallography purposes. These were also subjected to folding and analysis.

4.1. Distribution of UGAAG Motif in Genomes

4.1.1. Representation of UGAAG in Host and Viral Genomes

The web server *compseq* (Williams, 2001) was used to calculate the observed frequency of the UGAAG and GAAGU motifs as well as the expected frequency based on the nucleotide content of the submitted genome sequences in the *Homo sapiens* mRNA Reference Sequence (RefSeq) and KSHV genomes (Table 4-1). This means that the program uses the frequency of the each nucleotide

in the original sequence to then calculate the likely hood of a motif made up of 5 nucleotides. The observed frequency of UGAAG in the *Homo sapiens* mRNA RefSeq set was ~1.93 times higher than expected (expected = 0. 0010486; observed = 0. 0020203). Thus the UGAAG motif appears to be nearly two-fold overrepresented in the human genome in the 5' to 3' direction. The frequency of the UGAAG motif in reverse (i.e. GAAGU) was 1.15 times more frequent than expected (expected = 0. 0010486; observed = 0. 0012062). In light of the suggestion that SOX targets this sequence motif, one might expect its frequency to be lower in the viral genome KSHV. Indeed, UGAAG is only 1.07 times more frequent than expected in the viral genome, and the GAAGU motif is 0.76 times less frequent than expected. Hence, when looking at the ratio of observed/expected (Table 4-1), it can be concluded that overall the UGAAG and GAAGU are overrepresented in the *Homo sapiens* genome and less well represented in the KSHV genome, as would be expected for their genomic content.

Table 4-1: Observed versus Expected Frequency of UGAAG and GAAGU Motif in *Homo sapiens* and KSHV Genomes

Genome	Motif	Observed Frequency	Expected Frequency	Observed / Expected
Homo sapiens, mRNA RefSeq				
	UGAAG	0.0020203	0.0010486	1.9267378
	GAAGU	0.0012062	0.0010486	1.1503152
KSHV, complete genome (NC_009333.1)				
	UGAAG	0.0009060	0.0008501	1.0657304
	GAAGU	0.0006451	0.0008501	0.7588000

The difference in frequencies may explain partly why this motif has been selected as a target for endonucleolytic cleavage. But the lone ratio of expected to observed occurrence, does not necessarily represent the actual occurrence of the motif between the two genomes; viral and host. For this, the observed frequency of the UGAAG and GAAGU motifs were compared between *Homo sapiens* and KSHV. As these frequencies are the result of a ratio of observed counts over total counts of pentameric sequences, they represent the relative frequency within the genome, thus allowing the comparison of frequency between the two species. From this it can be observed that the UGAAG motif is present 2.23 times more often in the host genome than the viral genome (Table 4-2). Further, the GAAGU motif is

1.87 times more often present in the host RNAs then in the viral one.

Table 4-2: UGAAG and GAAGU overrepresentation within the *Homo sapiens* Genomes compared to the KSHV Genomes

Motif	Obs_V/Obs_H
UGAAG	2.23
GAAGU	1.87

4.1.2. UGAAG and the IL-6 mRNA

It has long been known that the interleukin-6 (IL-6) mRNA transcript evades turnover in KSHV infected cells and that KSHV also has its own viral version of IL-6; vIL-6 (Hutin et al., 2013, Rezaee et al., 2006). A recent paper (Hutin et al., 2013), has demonstrated that IL-6 contains a SRE1 in its 3'UTR. This SRE1 contains a non-canonical ARE.

>gi|224831235|ref|NM_000600.3| Homo sapiens interleukin 6 (interferon, beta 2) (IL6), mRNA
AAUAUUAGAGUCUCAACCCCCAAUAAAUAUAGGACUGGAGAUGUCUGAGGCUCAUU
CUGCCCUCGAGCCCACCGGGAACGAAAGAGAAGCUCUAUCUCCCCUCCAGGAGCCCA
GCUAUGAACUCCUUCUCCACAAGCGCCUUCGGUCCAGUUGCCUUCUCCCUGGGGCUG
CUCCUGGUGUUGCCUGCUGCCUUCCCUGCCCCAGUACCCCCAGGAGAAGAUUCCAAA
GAUGUAGCCGCCCCACACAGACAGCCACUCACCUCUUCAGAACGAAUUGACAAACAA
AUUCGGUACAUCCUCGACGGCAUCUCAGCCCUGAGAAAGGAGACAUGUAACAAGAGU
AACAUGUGUGAAAGCAGCAAAGAGGCACUGGCAGAAAACAACCUGAACCUUCCAAAG
AUGGCUGAAAAAGAUGGAUGCUUCCAAUCUGGAUUCAAUGAGGAGACUUGCCUGGU
GAAAAUCAUCACUGGUCUUUUGGAGUUUGAGGUAUACCUAGAGUACCUCCAGAACA
GAUUUGAGAGUAGUGAGGAACAAGCCAGAGCUGUGCAGAUGAGUACAAAAGUCCUG
AUCCAGUUCCUGCAGAAAAAGGCAAAGAAUCUAGAUGCAAUAACCACCCCUGACCCA
ACCACAAAUGCCAGCCUGCUGACGAAGCUGCAGGCACAGAACCAGUGGCUGCAGGAC
AUGACAACUCAUCUCAUUCUGCGCAGCUUUAAGGAGUUCCUGCAGUCCAGCCUGAGG
GCUCUUCGGCAAAUGUAGCAUGGGCACCUCAGAUUGUUGUUGUUAAUGGGCAUUCC
UUCUUCUGGUCAGAAACCUGUCCACUGGGCACAGAACUUAUGUUGUUCUCUAUGGAG
AACUAAAAGUAUGAGCGUUAGGACACUAUUUUAAUUAUUUUUAAUUUAUUAAUAUU
UAAAUAUGUGAAGCUGAGUUAAUUUAUGUAAGUCAUAUUUAUAUUUUUAAGAAGU
ACCACUUGAAACAUUUUAUGUAUUAGUUUUGAAAUAAUAAUGGAAAGUGGCUAUGC
AGUUUGAAUAUCCUUUGUUUCAGAGCCAGAUCAUUUCUUGGAAAGUGUAGGCUUAC
CUCAAAUAAAUGGCUAACUUAUACAUAUUUUUAAAGAAAUAUUUAUAUUGUAUUUA
UAUAAUGUAUAAAUGGUUUUUAUACCAAUAAAUGGCAUUUUAAAAAAUUCAGCAAA
AAAAAAAAAAAAAAA

Figure 4-1: IL-6 mRNA with 3'UTR SRE1, containing ARE, UGAAG and GAAGU Motif

The full mRNA sequence for IL-6. The 5'and 3'UTR are in black the protein-coding region is in grey. Within the 3'UTR the core ARE motifs AUUUA, which are bound by AUF1 and HuR, are highlighted in yellow and bold. The UGAAG and GAAGU motifs are found within the ARE, both highlighted in cyan and bold.

Although Hutin et al. (Hutin et al., 2013) demonstrated that the protection of IL-6 from SOX mediated decay was observed, when AUF1 and HuR were silenced using siRNA (Hutin et al., 2013), analysis of the gene reveals one 5'-UGAAG-3' and one 5'-GAAGU-3' motif in the 3'UTR embedded within the ARE region (Figure 4-1). This suggests that the SOX cleavage sites are obscured when the IL-6 mRNA ARE sites are sequestered by the HuR and AUF1 complexes. Further verification, however, would require the *in vivo* analysis of IL-6 transcripts containing mutations within the UGAAG and GAAGU motifs.

4.2. *In silico* RNA Folding and SOX-RNA Interaction Modelling

4.2.1. *In silico* Folding of the UGAAG Target Constructs

To investigate whether the three identified sequences had also other characteristics and motifs in common, they were aligned using *T-coffee* and *MView*. To explore whether cleavage after the consensus sequence was structure dependent, *in silico* folding was performed first on the entire sequence of the mRNA. This folding yielded 13 to 46 solutions for the full-length mRNAs. The number of solutions increased with the sequence length. For each construct a local secondary structure around the UGAAG site was identified. For the GFP sequence the same secondary structure centred on the UGAAG motif was identified in 10 out of the 13 solution folds for the full-length sequence; with the main solution being dominant in the population corresponding to the low energy folds. The HBB full-length mRNA folding yielded 14 solutions. Within these 14 solutions the same fold was identified 10 times around the UGAAG motif, again these folds populated the lower energy fold solution. In contrast, for the DsRed2 the highest repeat of the same local fold around the UGAAG motif was 5 times out of 39 solutions for the full-length sequence. When a stable substructure was identified, the sequence was cut to a 201 nucleotides sequence encompassing the UGAAG motif, as experiments suggested that 201 nucleotides were sufficient to maintain the structure. This reduced sequence also showed a preference for the same fold. Using several cycles of *mfold* on the identified sequences allowed us to cut down the RNA sequence to a minimal nucleotide length, while maintaining the structure originally observed in the full length mRNA. This minimal sequence was also needed to comply with the

current limitations in RNA synthesis, which are firstly sequence length, quantity and cost.

4.2.2. Comparison of the β-globin, DsRed2 and GFP RNA Sequence and Folds

When the three sequences were aligned with *MView* and *T-coffee* web server tools it became appeared that these three sequences shared more than just the UGAAG motif. As can be seen in figure 4-2 D, when HBB, GFP and DsRed2 are aligned along the UGAAG motif without gaps, the sequence contains a GRYG—RYY---RR-R--CY-A---UGAAG motif upstream of the cleavage site, which contains 12 conserved purines and 6 conserved pyrimidines. When aligned using *T-coffee* along the UGAAG motif, but allowing for gaps (Figure 4-2 E), 15 strictly conserved nucleotides are purines and 5 were pyrimidines. The nucleotides that conserved between DsRed2 and GFP of the sequences are made up of 15 purines and 9 pyrimidines in the upstream region of the UGAAG motif. Taken together these alignments suggest that conservation of purines is an additional sequence characteristic in connection with the UGAAG motif.

The fold of the GFP, HBB and DsRed2 mRNA was investigated. From this folding it could be deduced that all 3 sequences had only two characteristics in common. Firstly the presence of the UGAAG motif in a stem loop and secondly the fact that in all three, the two guanines were always involved in base pairing (See secondary structure prediction Figure 4-2 A, B and C). The secondary structure prediction (Figure 4-2) agreed with the sequencing results from the Glaunsinger Group (Covarrubias et al., 2011), with the cleavage site located towards loops/bulges. The cleavage site for the GFP stem loop was found in the upper loop bulge (Figure 4-2 C). As the secondary structure prediction for the GFP RNA was the most stable and it had the most supporting data in the context of SOX induced degradation (Covarrubias et al., 2011), it was subsequently the focus of our investigation.

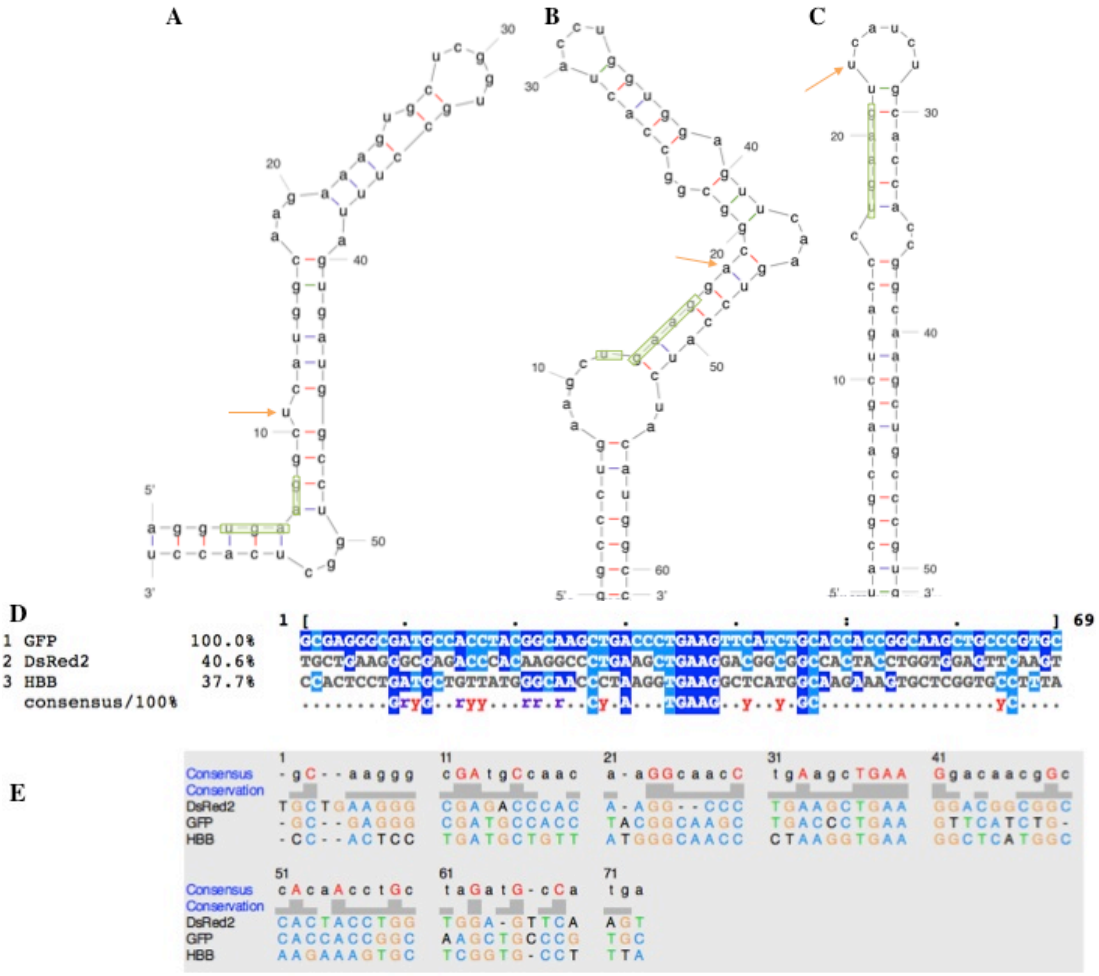


Figure 4-2: Comparison of HBB, DsRed2 and GFP Stem Loops Features

The folded RNA in A) is the stem loop of HBB, in B) the stem loop for DsRed2 and in C) the stem loop for GFP. The UGAAG motif for each of these is shown in a green box and the cleavage site is indicated with a orange arrow. D) T-coffee aligement of HBB, GFP and DsRed2 69 nucleotide seqeunces along the central UGAAG motif without gaps. E) MView aligment of the same sequences allowing for gaps.

4.2.3. 3D Structure Prediction of the GFP Stem Loop Fits in Active Site

An ensemble of 250 SOX conformations was generated from the native wild type structure 3fhd using the distance constraints-based method within tCONCOORD (Seeliger and DeGroot, 2009). This allowed visualising the flexibility of SOX. As it can be seen from the figure 4-3 A, SOX has high flexibility in the loop regions, which could be expected. In addition it can be seen that SOX

presence a high variability in the N-terminal and C-terminal region. The variability of the N-terminal was also seen in the many solved x-ray crystal structures obtain during this PhD.

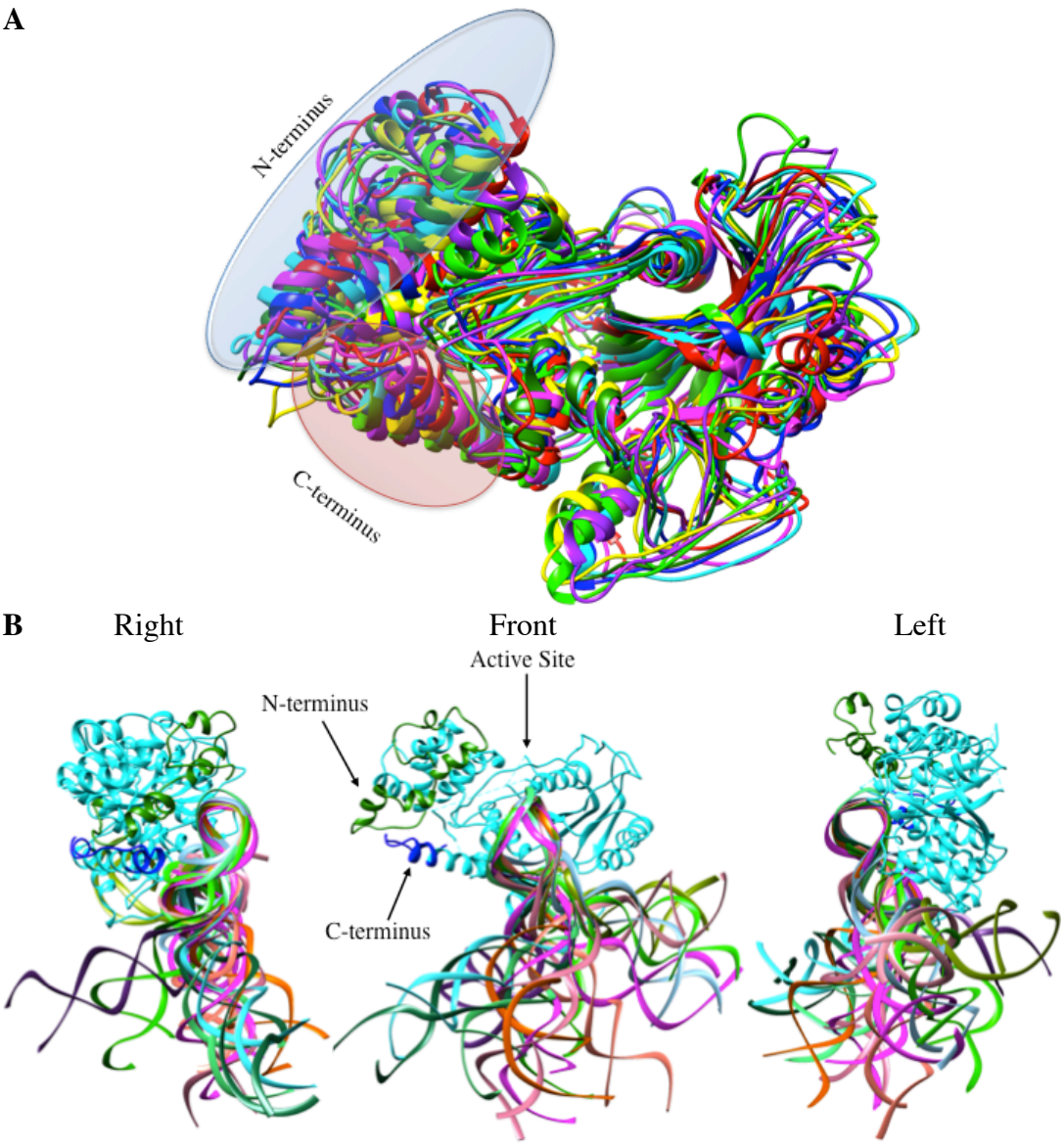


Figure 4-3: Exploration of Conformational Space of the RNA Stem Loop and SOX

A) tCONCOORD-generated conformers from a native wild type SOX structure (PDB: 3fhd). Eight selected conformers depicting the extent to which structural variation was simulated. B) Selected GFP 51 nucleotides stem loop structures from the 114 samples fitted into the active site of SOX, showing varying degrees of curvature.

The 3D structure of the GFP stem loop containing the UGAAG motif was predicted using *Mc-Fold*/*-Mc-Sym* (Figure 4-3 B and 4-4 B-C). The resultant structures were then fitted using the *Chimera Suite* into a map of the double stranded DNA of the SOX PDB structure (PDB ID: 3pov), which was co-crystallised with double stranded DNA.

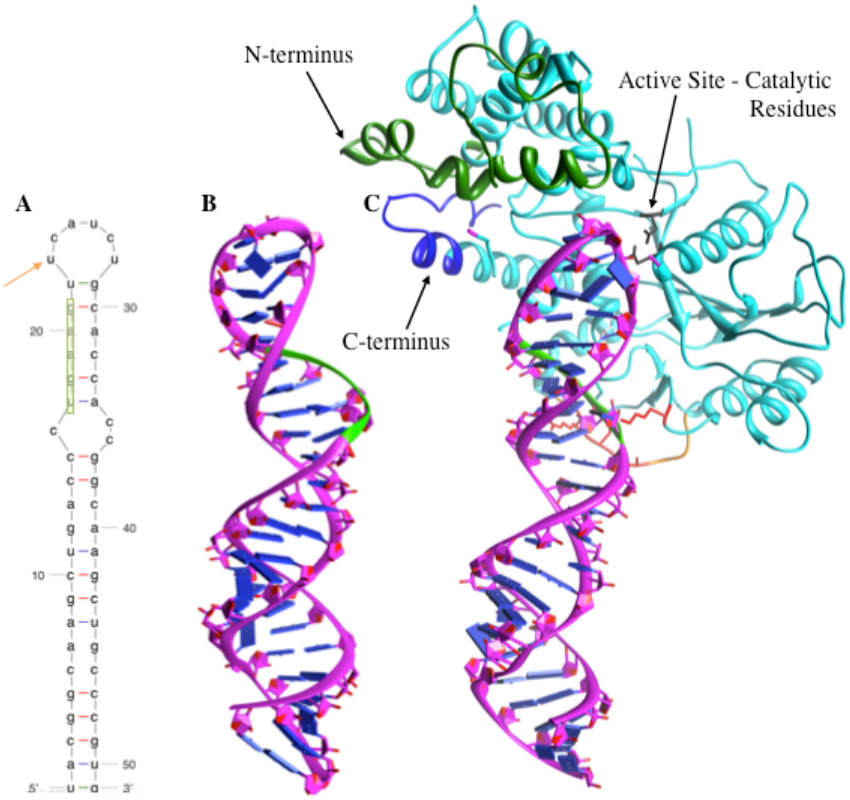


Figure 4-4: Predicted 3D Structure Fits into the Active Site of SOX

A) The 2D secondary structure prediction for the 51 nucleotides GFP stem loop. B) MC-Fold / MC-Sym 3D structure prediction of the 51 nucleotides GFP sequence. The UGAAG sequence is highlighted in light green. C) The 3D 51 nucleotides GFP RNA structure fitted into the PDB structure of the SOX-DNA complex (PDB ID: 3pov). Highlighted in light green is the UGAAG motif, the N-terminus is in dark green (containing HSO residues T24 and A61), the C-terminus in blue (containing HSO residues D474 and Y477Stop) and the NLS loop in orange, with arginines and lysines in red. The active site residues are highlighted in black (D221 and E244) and these are typically involved in RNA and DNA turnover.

The 114 solutions for the GFP sequence showed varying degree of curvature and also fitted with more or fewer atomic clashes into the active site (Figure 4-4 A). The solution that best fitted is shown in figure 4-4 D. The RNA followed the straight pattern of the DNA, with the UGAAG pattern facing the NCL residues 315–320. This protein sub-sequence contains a number of lysines and arginines (Figure 4-4 D).

4.2.4. Mutagenesis and Engineering

To investigate the role that the UGAAG motif played in the recognition and cleavage motif was substituted with UCUCU. The UCUCU was chosen as this sequence was pyrimidine rich versus the purine rich UGAAG. This mutated GFP UCUCU sequence showed two alternative folds when folded *in silico* using *mfold*. These did not maintain the same overall structure as the wild type GFP (Figure 4-5 A and B). But in both folds one or two stem loops are formed. In the fold in figure 4-5 A and B UGA is found just before the top loop forms, followed by GAA upstream, which again is reminiscent of the wild-type structure and purine rich neighbourhood (Figure 4-2 C). The loop downstream of then UGA is bigger and thus presents more free base pairs.

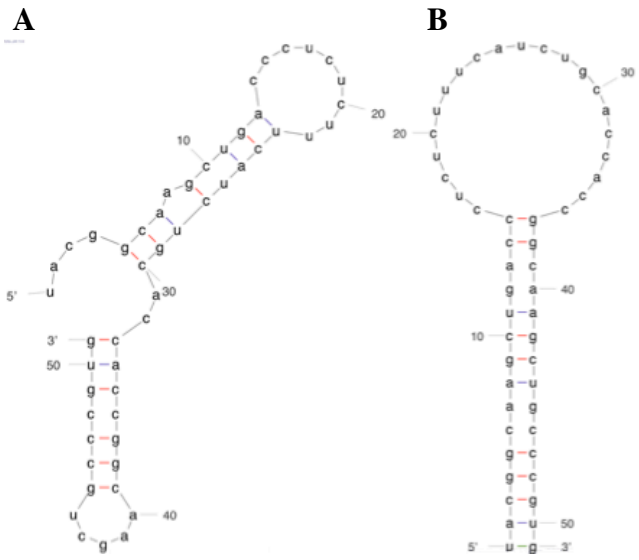


Figure 4-5: Effects of the UCUCU Mutation on the GFP Stem Loop Sequence

The UCUCU mutation did not maintain the same structure as the 51 nucleotides wild type GFP Stem Loop and presented 2 alternative folds, A) and B) .

A shorter version of the 51 nucleotides GFP sequence containing the UGAAG motif was engineered. This version was 23 nucleotides long and contained an introduced 2 GC base pairs at the bottom of the stem loop to stabilize the structure and an adenine overlap at the bottom, which had been reported to help with crystallisation (Hoggan et al., 2003) (Figure 4-6). This shorter version was later used in crystallisation trials.

Figure 4-6: Maintained Stem Loop Structure of the Engineered 23 nucleotides GFP Sequence

The 23 nucleotides version of the GFP stem loop maintained the characteristic fold.

Chapter 5: Biochemical and Biophysical Characterisation of SOX, Xrn1 and RNA Interaction

5.1. SOX WT binds Xrn1

The Xrn1 and SOX proteins, wild type and mutants, used in the biochemical and biophysical experiments were all recombinantly produced and purified to a high level of purity; Xrn1 was purified to a lesser degree (See Appendix D and E for more details).

5.1.1. Pull Down Assay of Xrn1 and SOX WT

Despite SOX's intrinsic endo/exonuclease activities, the rapid and global decay of host mRNA transcripts observed cannot be accounted for on the basis of these activities alone (Covarrubias et al., 2011, Kronstad and Glaunsinger, 2012). Whilst Xrn1 has been shown to be involved in this pathway in other viruses, the nature of its involvement in KSHV infection is unclear (Gaglia et al., 2012). Collaborators, however, were able to demonstrate a physical interaction between the N-terminal region of Xrn1 and SOX using a yeast two-hybrid system (Ebrahimi, B.; unpublished work) consistent with the N-terminal region of Xrn1 being highly conserved and predicted to be involved in protein-protein interaction. This interaction was therefore tested *in vitro* using a pull down assay. His-tagged *K. lactis* Xrn1 preincubated with untagged SOX was applied to a HisTrap column. Both proteins co-eluted following the application of an imidazole gradient (Figure 5-1) confirming that they are indeed capable of forming a complex *in vitro*. The controls of SOX and Xrn1 on their own were also performed to exclude non-specific binding to the column. To further validate this interaction, MST experiments were undertaken (Figure 5-2).

Having shown that SOX and Xrn1 physically interact, the effects of this association on formation of the SOX 51 nucleotides GFP RNA complex and SOX's turnover ability in presence of Xrn1 were subsequently investigated (Figure 5-3 and Figure 5-5 respectively).

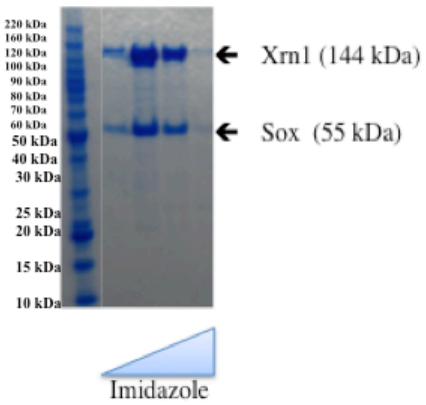


Figure 5-1: Pull Down of Xrn1 and SOX

With the increasing imidazole concentration increased protein bands for Xrn1 and SOX became visible; increasing and decreasing stoichiometrically.

5.1.2. Xrn1 and SOX WT bind with µM K_d

To further validate the predicted and experimentally demonstrated interaction between Xrn1 and SOX, MST was used. The binding data for the SOX and Xrn1 interaction showed the characteristic sigmoidal binding curve that was fitted with the nonlinear solution of the law of mass action, giving a K_d of 0.865 µM (Figure 5-2).

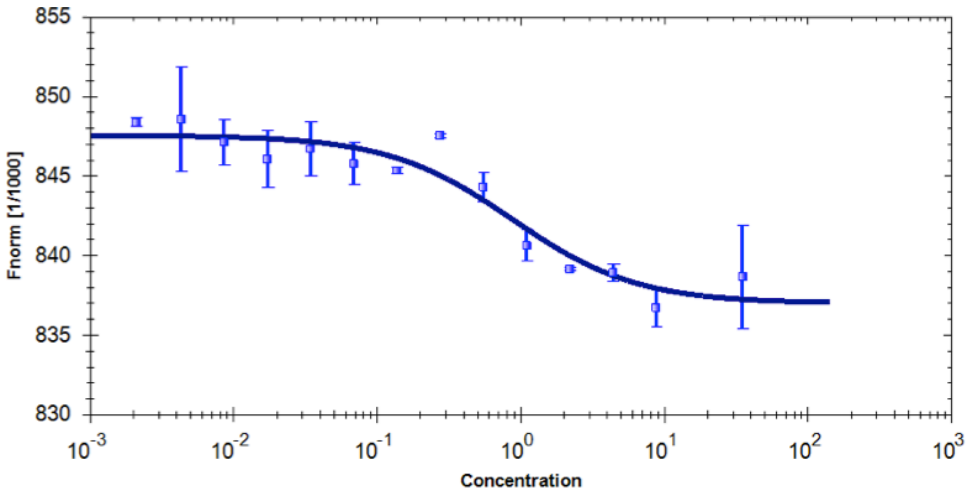


Figure 5-2: Microscale Thermophoresis Binding Curve of Xrn1 and SOX

5.2. SOX Binds UGAAG Stem Loop Structures

Based on the TBE gel shift protocol from (Bagneris et al., 2011) preliminary gel shifts with the 51 nucleotides GFP RNA were undertaken to test the prediction from the *in silico* folding studies. Due to recurrent issues in co-crystallisation of SOX and the RNA stem loop and suggestions from the literature, that pHs below the proteins pI and lower pHs were favourable for binding and RNA stabilisation, a range of pHs were tested. It was also noted that SOX's endonuclease activity would potentially be inhibited by a lower pH, as it is known as the alkaline exonuclease (Bujnicki and Rychlewski, 2001). Salt concentrations are also known to affect binding and as most crystallisation solution contain salts in sometimes very high concentrations, salt concentrations were also screened (Lohman, 1986). These gels helped to establish conditions for crystallisation trials that would increase the RNA stability and the protein-RNA complex (See Appendix F, Figure F-1). From here on, the crystallisation and TBE gel buffers contained 100 mM NaCl and a pH 7 (known as buffer TBE C).

Having established the optimal conditions for binding using the 51 nucleotides GFP RNA, the ability of SOX to associate with the UGAAG containing stem loop structures identified *in silico* were next investigated. SOX WT also binds the 58 nucleotides HBB RNA (Figure 5-3 D). Interestingly, the different complexes show different stabilities. The tightest bands and thus the most stable complexes are formed between the WT SOX and the 51 nucleotides GFP RNA (Figure 5-3).

Xrn1 does not bind the stem loop. When WT SOX and Xrn1 are present together a major band for SOX binding the stem loop is visible and a smaller higher molecular weight band is also forming; the latter could be a complex of SOX, Xrn1 and RNA (Figure 5-3 B). These gels were also stained for with Instant Blue, the sites of RNA shifts corresponded to sites stained for protein by the Instant Blue. However titration experiments with Xrn1 should be undertaken to enhance the band thought to contain Xrn1-SOX-RNA band.

It was then established whether the mutants identified as having a profound effect on HSO (Goldstein and Weller, 2004, Glaunsinger et al., 2005) were defective in their ability to associate with 51 nucleotides GFP RNA (Figure 5-3 A).

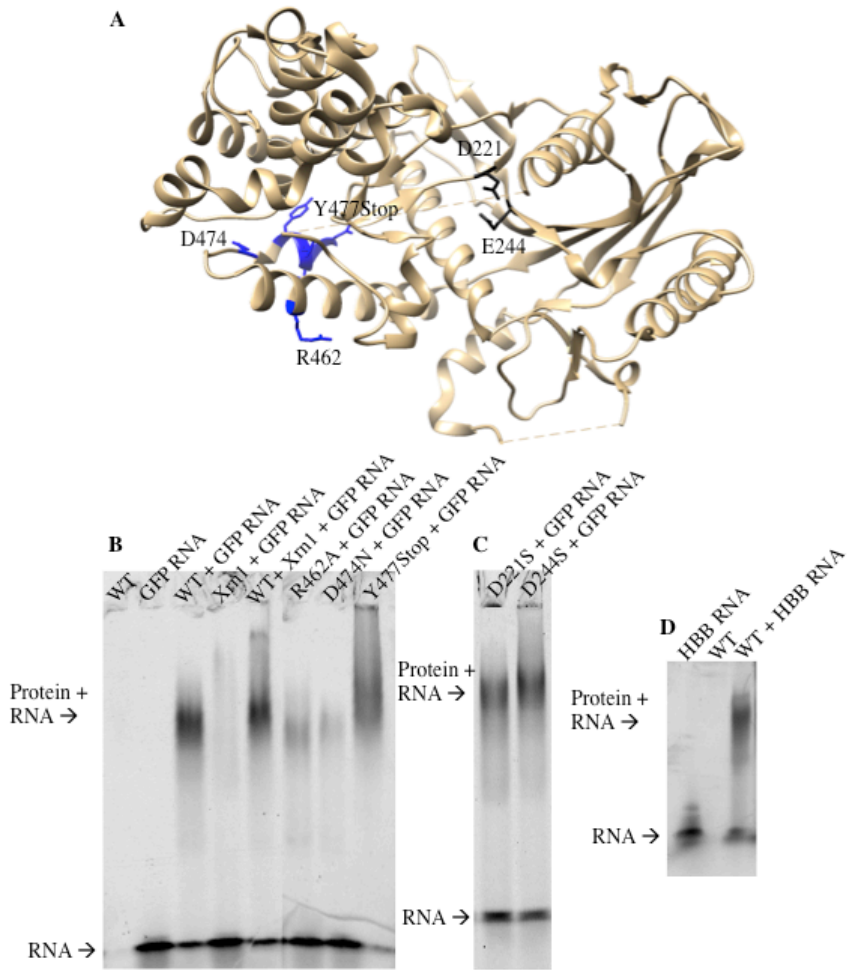


Figure 5-3: TBE Gel Shift of SOX and RNA Binding

A) Representation of SOX, with the active site residues highlighted in black (D221 and D244) and the C-terminal residues involved in HSO in blue (R462, D474 and Y477Stop). B) In the SOX control lane no fluorescence was detected and at the RNA control lane fluorescence was detected at the bottom of the gel. It can be observed that bands are narrower for the 51 nucleotides GFP RNA complex and less defined for the HSO mutants; n=3. C) The active site mutants are not affecting the binding ability of SOX to the 51 nucleotides GFP RNA; n=3. D) Whereas the bands for the 58 nucleotides HBB RNA appear broader, thus indicating a less stable complex. This was an individual experiment; n=2.

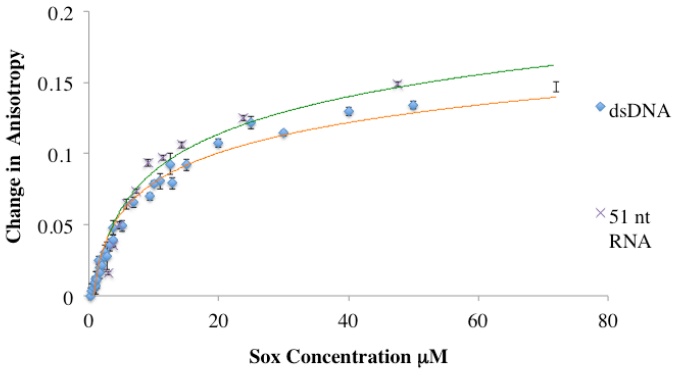
Whilst the catalytic mutants D221S and E244S appear to be completely unperturbed in their capacity to associate with the 51 nucleotides GFP RNA as would be expected (Figure 5-3 C), D474N and Y477 are significantly disrupted.

These two mutants are well known HSO mutants. Whilst D474N has significantly reduced affinity for the 51 nucleotides GFP RNA compared to WT, the Y477stop shifted is highly smeared suggestive of an inability to form a stable complex (Figure 5-3 B). Interestingly, both D474 and Y477 are located on the C-terminal helix at the bottom of the active site. This region contains a number of lysine and arginine residues consistent with an additional nucleotides binding site (Figure 5-3 A). To further investigate the role of this region in RNA stem loop recognition, R462, a residue in the C-terminal helix was mutated to alanine and its capacity to bind the 51 nucleotides GFP RNA tested. This revealed that R462A does affect RNA stem loop binding and that potentially a greater part of the C-terminal helix is involved in binding.

5.3. SOX has Enhanced Affinity for UGAAG Stem Loop Structures

Previous research had shown that SOX's was able to bind double stranded DNA, single stranded and double stranded RNA using FA (Bagneris et al., 2011).

A



B

Substrate	K_d in µM
ds-DNA	12.0
ds-DNA-5'P	6.0
GFP 51	4.6

Figure 5-4: Fluorescence Anisotropy Binding Assay of SOX Involving the 51 Nucleotides RNA Stem Loop

A) The binding curve of SOX for 51 nucleotide RNA stem loop and for ds-DNA, the later was obtained by Dr. Bagneris and published in Bagneris et al., 2011. B) The K_d's obtained for the 51 nucleotide stem loop and those reported for ds-DNA and ds-DNA-5'P in Bagneris et al., 2011.

To establish its affinity for the 51 nucleotides GFP RNA stem loop, similar studies were performed, as FA is superior to EMSA quantitatively. As for the

EMSA experiments titration with non-specific and unlabelled RNA should be undertaken to exclude non-specific binding. These experiments reveal that SOX had a similar affinity for the stem loop as for ds-DNA-5'P and ds-DNA (Figure 5-4) (See figure 1-2 B for sequences). The measured K_d of 4.6 μM was more consistent with the K_d observed for ds-DNA-5'P (6 μM) and was ~20 fold greater compared to ss-RNA-5'P and ds-RNA-5'P (See Bagneris et al., 2011).

5.4. SOX Turns Over RNAs

Having ascertained that SOX was able to bind to the stem loop structures identified, it was next established whether the 51 nucleotides GFP RNA was a target for endonucleolytic cleavage. Once cleavage had been confirmed (Figure 5-5), RNA turnover conditions were ascertained since it could not be assumed that they would be consistent with those determined for ss-RNA-5'P. In order to ascertain the optimal conditions for RNA cleavage and additionally with a view to its inhibition for structural studies, different salt concentrations and lower pHs were investigated. It was found that higher pHs and lower salt concentrations favoured turnover. Based on these results and the previous experimental protocol used by Dr. Bagneris, all subsequent cleavage assays were performed using the condition 50 mM Tris–HCl pH 9, 200 mM NaCl, 20 mM β-mercaptoethanol, 10 mM $MgCl_2$ (Buffer TBE' St.) for these gels. Lower pH and higher NaCl concentrations were shown to inhibit the activity of SOX, which then lead to lower pHs being used in the crystallisation trials.

5.4.1. SOX WT Turnover of HBB and GFP RNA

SOX turns over the 51 nucleotides GFP and 58 nucleotides HBB RNA (Figure 5-5 A). Although HBB RNA appears to be more susceptible to degradation than the 51 nucleotides GFP RNA in the absence of SOX, the reduction in intensity of the highest molecular weight band nonetheless suggests that it also is a substrate (Figure 5-5 B). The UCUCU mutant GFP RNA is also turned over by SOX (Figure 5-5 C). SOX has an intrinsic RNA endonuclease activity and RNA exonuclease activity. The later relies on a 5' phosphate for activity and all oligonucleotides were synthesised so not to include a 5'phosphate and with the FAM label on the 3' end.

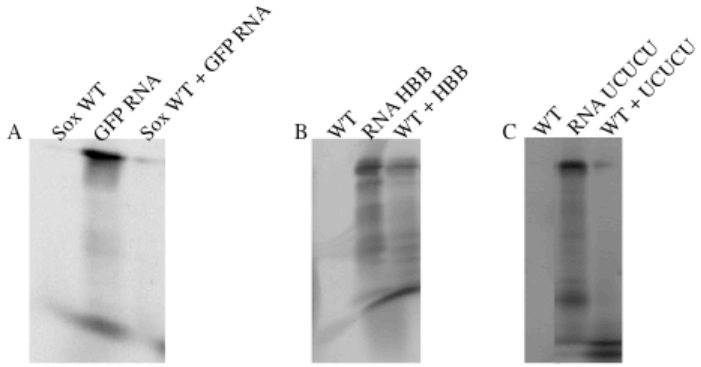


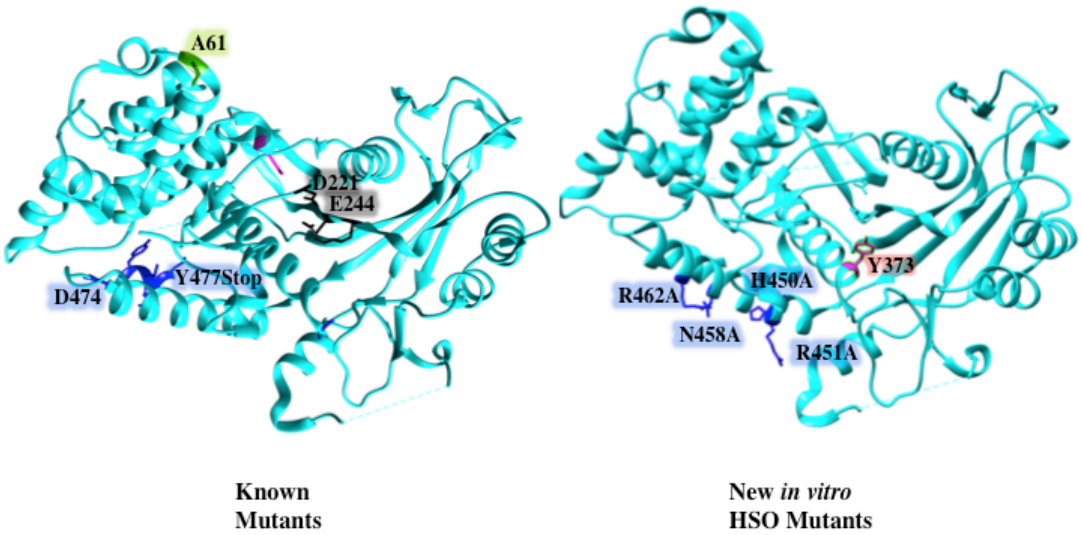
Figure 5-5: TBE-Urea of SOX and RNA Binding

In this figure the RNase activity of SOX can be seen for (A) 51 nucleotides GFP (B) the 58 nucleotides HBB and (C) the 51 nucleotides UCUCU mutant RNA. The RNA contained no 5' phosphate, thus the cleavage had to be induced via endonucleolytic cleavage by SOX, followed by exonucleolytic cleavage by SOX's exonuclease activity. Though these gels cannot be absolutely conclusive it can be seen that in the third lane when SOX is present the RNA on the top and bottom of the gel diminished; n=3.

5.4.2. SOX HSO Mutants Abrogate or Decrease Turnover of GFP RNA

SOX WT turns over the RNA stem loop, while Xrn1 by itself is not able to turnover the stem loop RNA. The turnover of the 51 nucleotides GFP RNA by SOX appears to be attenuated in the presence of Xrn1 (Figure 5-6). This could indicate that either Xrn1 occludes the active site of SOX, or that Xrn1 catalysed cleavage involves re-modelling of the cleaved 51 nucleotides GFP RNA substrate facilitated by conformation re-arrangement in one or both proteins for handover. This could also be the indication of the possible participation of an additional co-factor (Uetz et al., 2006). As the optimal conditions for Xrn1 activity only strongly differ in the NaCl concentration from 50 mM to 200 mM NaCl, otherwise the $MgCl_2$, pH 8.0 and reducing agent concentration were comparable (Chang et al., 2011), it is possible, but unlikely that the experimental condition interfered with Xrn1's activity.

A



B

SOX Mutants	Impact on Activity
A61T	Abrogates HSO *in vivo*
D221	Abrogates DNA and RNA Catalysis *in vivo* and *in vitro*
E244S	Abrogates DNA and RNA Catalysis *in vivo* and *in vitro*
Y373A	DNA contact based of Crystal Structure 3pov
H450A	C-terminal Mutant hypothesized to impact RNA binding
R451A	C-terminal Mutant hypothesized to impact RNA binding
N458A	C-terminal Mutant hypothesized to impact RNA binding
R462A	C-terminal Mutant hypothesized to impact RNA binding
D474N	Abrogates HSO *in vivo*
Y477Stop	Abrogates HSO *in vivo*

C

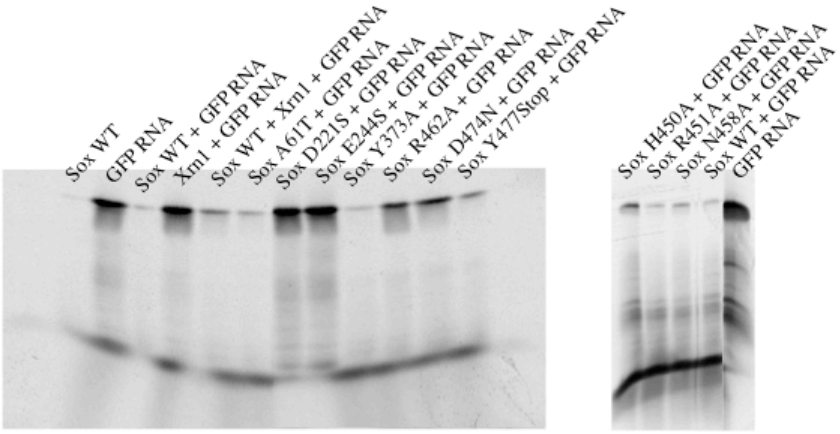


Figure 5-6: Impact of HSO Mutations and C-terminal Helix Mutation on SOX Endonucleolytic Activity

The location of the various mutations (A), the documented effect of the mutants (B) and the effects of these on RNA stem loop degradation (C) are shown in this figure.

The mutants A61T and Y373A both show the least attenuation in RNase activity, A61T is a known HSO mutant, while Y373A is a DNA contact mutant (on the basis on the crystal structure PDB ID: 3pov). The mutants D221A, E244S, R462A, D474N and Y477Stop inhibited SOX RNase activity. D221A and E244S are catalytic mutants, while D474N and Y477Stop are known to be important for HSO and RNase activity. Y477Stop is a C-terminal truncation. Further mutants were created along the C-terminal helix, which contains the HSO mutants D474N and Y477Stop and is situated at the bottom of the active site. R462A found closer to the known HSO mutants D474N and Y477Stop shows as strong an inhibition of RNase activity as D474N. While N458A and R451A do not appear to have an impact on the RNase activity, H450A has a more marked impact on RNase activity. This indicates that different areas of the SOX protein participate in the endonucleolytic cleavage mechanism, such as certain residues in the C-terminal loop associated with HSO and newly identified residues in the two C-terminal helices (H450 and R462).

5.5. Crystallization

5.5.1. SOX and Xrn1

As SOX and Xrn1 were shown to interact (See Section 5.1), crystallisation trials were undertaken. Despite crystals growing overnight in several conditions, the typical morphology is shown in figure 5-7 A and B. All crystals, however, diffracted only to ~ 6 Å (Figure 5-7 C), which is just within the limitations for MR. Once the data was collected and analysed, no complex was observed, and only the presence of Xrn1 monomers could be observed. Other crystals grown from the complex conditions, but of different morphology, were found to contain SOX monomers.

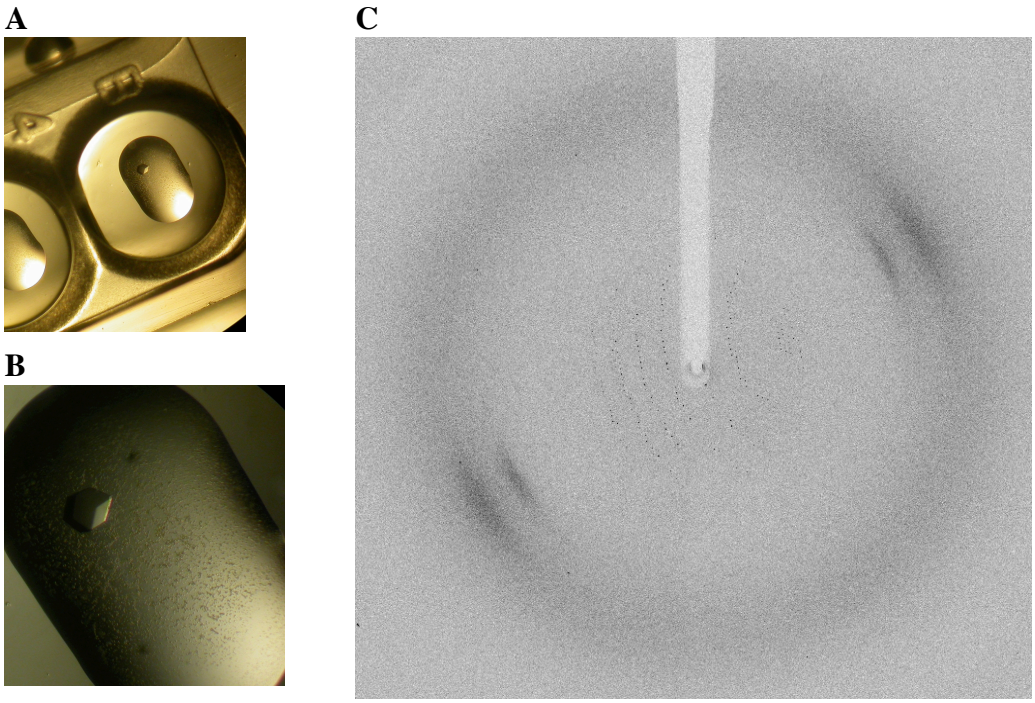


Figure 5-7: Crystal obtained when setting up with Xrn1 and SOX

A crystal of Xrn1 and SOX, A) and B), from the following condition 0.1 M Na/-thiocyanate and 20 % PEG 3350 diffracted to 5.87 Å C).

5.5.2. SOX, 51 nucleotides Structured RNA and Other RNAs

Crystals were obtained in many conditions when set up with SOX WT and SOX 244 in complex with 51 nucleotides GFP, 23 nucleotides GFP and the 58 nucleotides HBB RNA oligonucleotides. However, only several apo structures of SOX WT and SOX 244 were obtained. All of these contained a single SOX molecule in the asymmetric unit whose active site was occluded by neighbouring molecules in the crystal lattice. By contrast, crystals grown from 150 mM Malic Acid (pH 7) and 20% Peg 3350, using the low salt and low pH crystallisation buffer, revealed two SOX monomers in the asymmetric unit (Table 5-1). Following MR that confirmed the asymmetric unit composition, it could be seen that the active sites of the KSHV monomers were no longer obstructed and able to accommodate the stem loop. Furthermore, in the Fo-Fc difference density map (Figure 5-8), a density could be observed in the active site region consistent in size with RNA nucleotides. Attempts were made to fit either nucleotides or an RNA stem loop into the density followed by refinement. However, subsequent 2Fo-Fc maps failed to

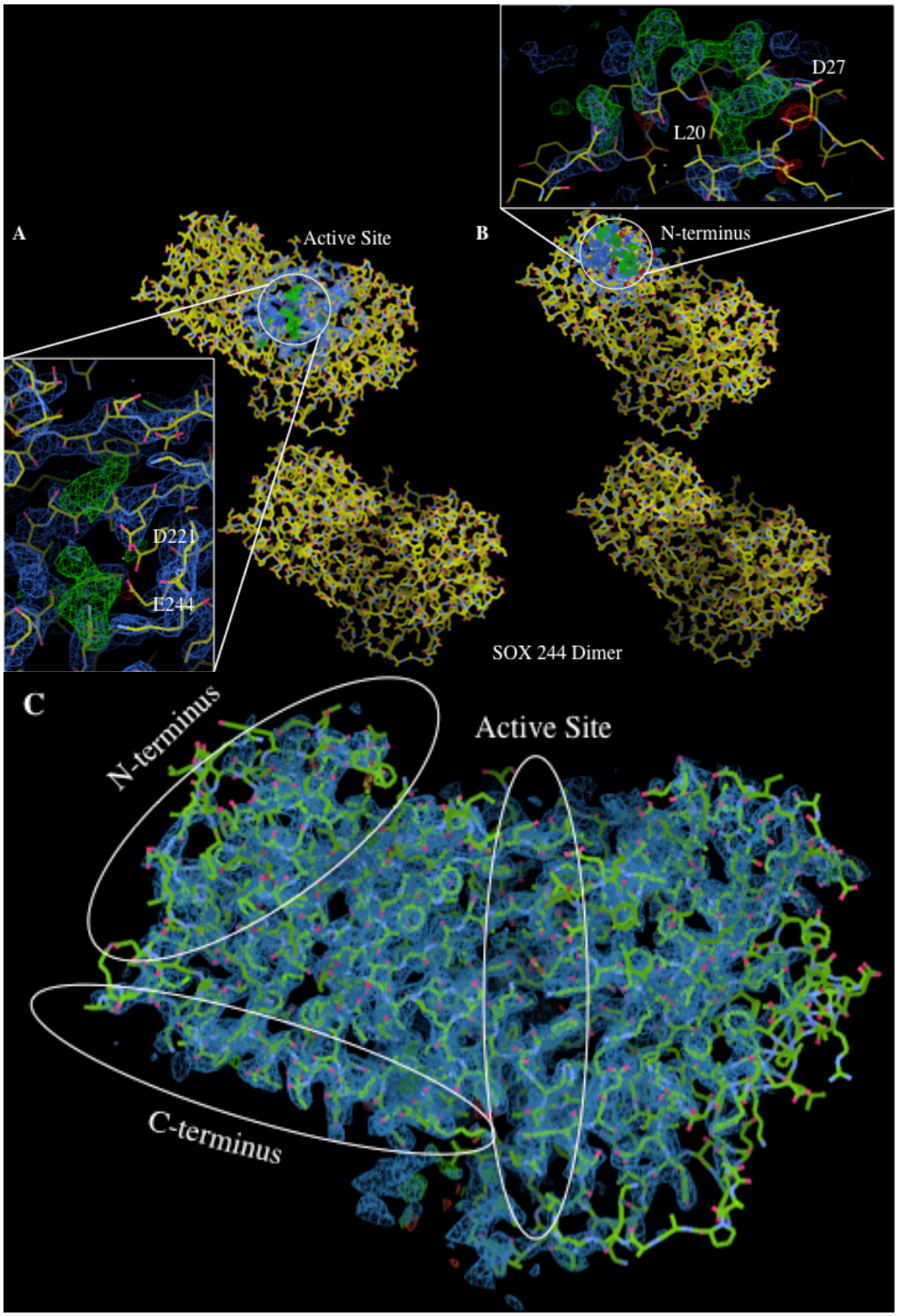


Figure 5-8: Dimer Crystal Structure of SOX 244

A) Fo-Fc map (green) is visible within the active site of the protein SOX 244 dimer with the catalytic residues D221 and E244 highlighted. B) Fo-Fc map (green) is visible at the N-terminus with residues that do not fit the map highlighted. C) SOX 244 after refinement with absent Fo-Fc map in the N-terminus and active site.

show any convincing density when contoured above 1σ. The original difference density could therefore have either originated from the solvent, though this seems to be unlikely based on its size and shape or be due to partially occupied/highly mobile nucleotides. The N-terminal loop of SOX had a conformation that was similar to the N-terminal conformation observed previously in the apo structure (PDB ID: 3fhd) and thus differed from the conformation seen in then dsDNA bound structure (PDB ID: 3pov). On the basis of these results, a 23 nucleotides stem loop comprising nucleotides 16 to 34 was engineered for co-crystallisation. It was again possible to produce crystals, but these did not contain any ligand bound structures.

Table 5-1: X-ray Data Collection and Processing Statistics for SOX 244 Dimer

Space group	**P 1 2_1 1**
Cell dimensions (Å, °)	a=62.7, b=78.3, c=111.3, β = 98.6
Resolution (Å)	48.59 – 2.96 (3.1-2.96)
No. of reflections	Total 66491 Unique 22013
$\mathbf{R}_{merge}$	0.17(0.42)
Completeness (%)	98.6 (96.6)
Multiplicity	3.3
$(I/\sigma(I))$	8.26 (2.98)
$R_{\text{-cryst}}$ (%)	22.6
R_{free} (%)	29.9

After the computational analysis of the UGAAG containing RNA sequences it was possible to determine experimentally using EMSA and FA that SOX binds the RNA stem loops with similar affinity to ds-DNA. SOX is able to induce cleavage of the RNA stem loop. Additionally it was demonstrating that the HSO mutants D474N and Y477Stop and R462A affect binding and turnover of the RNA stem loop by SOX. Using a pull down assay and microscale electrophoresis it was demonstrated that SOX and Xrn1 interact directly. And finally a crystal structure of SOX containing the E244S mutation was obtained.

Chapter 6: Discussion, Conclusion and Future Work

6.1. Discussion

To conquer the cell and its RNA and protein expression machinery, viruses have to tap into the regulatory clockwork of the cells that they target (Kronstad and Glaunsinger, 2012, Abelson, 1979, Walsh and Mohr, 2011). This regulation occurs at the RNA, DNA and protein level. For this the viruses have copied host protein coding genes and DNA regulatory elements, adjusted to fit their purposes and timing. The DNA and protein level of regulation had been focused upon for many years, but with the recent breakthroughs in the RNA field the key role that RNA plays in regulating the cell has been shown; via the discovery of RNA and *cis*, *trans*, stability and degradation regulating RNA elements (Licatalosi and Darnell, 2010). As more host RNA regulatory mechanisms are elucidated and more viral mechanisms are discovered that work on the RNA level, it might help to elucidate how viruses with their limited genomes are able to so fundamentally change the fate of a cell (Kronstad and Glaunsinger, 2012, Walsh and Mohr, 2011). It is important to know how KSHV takes over the host degradation machinery during the lytic cycle to facilitate viral proliferation and evasion from the immune system.

The Ganem group first established that the onset of the lytic phase was accompanied by the rapid and global degradation of host mRNA transcripts and that this also correlated with the expression of SOX (Glaunsinger and Ganem, 2004b, Glaunsinger et al., 2005, Glaunsinger and Ganem, 2006). The Glaunsinger group described the presence of an UGAAG motif in the vicinity of the cleavage sites of mRNA transcripts that were targeted by SOX (Covarrubias et al., 2011). If SOX was targeting this pentameric sequence, a reasonable expectation would be for the motif to be over-represented in human transcripts whilst under-represented in those originating from KSHV. To investigate this, the ratio of observed over expected frequency of the UGAAG and GAAGU in the host genome was calculated and found to be 1.93 and 1.15 respectively (Table 4-1, Section 4.1.1). UGAAG thus occurs twice as frequently than would be expected at random given the nucleotide content of the host transcriptome, whilst the GAAGU has a slight increase in occurrence compared to expected values. By contrast, UGAAG in the viral genome has an observed to expected ratio of 1.07 consistent with random occurrence

whereas GAAGU appears less frequently (0.76) (Table 4-1, Section 4.1.1). The KSHV sequence that was used was the KSHV genome, as it was not possible to retrieve 86 mainly intronless genes (Rezaee et al., 2006). However as the KSHV genome contains mainly intronless genes and little intergenic sequence this was seen as acceptable as a comparison as SOX operates in the cytoplasm and thus would encounter either these sequences as they are found in the genome. These results are consistent with the preferential degradation of the host genome observed but nonetheless reveal that the viral genome would still be a significant target for degradation. This overrepresentation that was found could also be explained by the presence of ESE signals in the coding regions of host mRNAs, due to the splicing process that the majority of host transcripts undergo. These ESE signals have recently been shown to be highly overrepresented in the mRNA transcripts in the vicinity of exon-exon junctions. ESEs are motifs that are degenerate, but contain GA rich sequence and have been associated with UGAAG and GAAG sequences (Chasin, 2007, Pertea et al., 2007, Tacke and Manley, 1995, Thomas et al., 2012). The binding of EJC complex on the spliced mRNA has been linked to GAAG and GA rich binding elements (Saulière et al., 2012). It is possible that the lack of EJC complex association with viral mRNAs would help the viral transcript to evade SOX induced degradation in addition to the lower UGAAG ratio, if SOX was to target host mRNAs by also binding EJC. Although frequency analysis does indicate that the host genome would be preferentially degraded, it does not take into account any possible nucleotide changes that could be occurring within the pentameric sequence that may be tolerated by SOX in terms of binding, similar to the GA rich sequences associated with binding of splicing factors, such as ESS, ESEs and EJC. The alignment of the three identified endonucleolytically cleaved mRNAs around the UGAAG motif showed that upstream of the cleavage site GA rich elements were present (Figure 4-2 D and E). In addition, as this ratio of observed versus expected on the basis of the genomes nucleotide content is not the same as observed overall frequency, the ratio of host observed versus viral observed frequency may shed light into how likely it is that SOX encounters viral or host transcripts. This ratio for the UGAAG motif was 2.23 and 1.87 for the GAAGU motif (Table 4-2, Section 4.1.1). These ratios assume that the cell would contain equal quantities of host and viral transcripts. In the lytic phase the viral genome is highly expressed,

both in terms of the number of transcripts per gene and the total number of genes expressed (Schumann et al., 2013). Overexpression of SOX occurs at the beginning of the lytic cycle, when the viral genome is just beginning to undergo replication and is thus competing with host transcripts for the replication machinery. Preferential elimination of host transcripts at this early stage could thus be beneficial to the virus further down the lytic cycle. The ratio of host to viral transcripts in the cytosol, however, has yet to be reported. Thus it is possible that SOX is cleaving its own RNA, if it is not located via protein-protein interaction to spliced host mRNA transcripts that are transnationally competent; e.g. via association with EJC. It was shown that SOX targets ribosome bound and translationally competent mRNAs (Clyde and Glaunsinger, 2011).

Several host transcripts have been identified that evade endonucleolytic cleavage by SOX; one of these transcripts is IL-6. IL-6, when membrane bound, has an anti-inflammatory effect and is known to evade degradation in several γ-herpesvirus infections. A single UGAAG motif has been identified in the 3'UTR of the IL-6 gene, however, research published in 2013 (Hutin et al., 2013) also revealed the presence of a SRE-1 motif in this region. This sequence contains non-canonical AREs, which play critical roles in mRNA regulation since they can be targeted for both degradation and stabilisation by several protein factors. Interestingly, studies have shown that the SRE-1 motif of IL-6 forms a complex with HuR and AUF1 (Figure 4-1, Section 4.1.2). This would lead to sequestration of the UGAAG motif in the 3'UTR and protection of IL-6 from the endonucleolytic activity of SOX. Consistent with this, it has been reported by the Glaunsinger group that deletion of AUF1 and HuR renders IL-6 transcripts susceptible to SOX induced degradation. These studies support the hypothesis that SOX utilises specific motifs to degrade the host mRNA, but also illustrates that the global and rapid degradation of host mRNA transcripts by SOX is complex and likely to require several host and viral target proteins. This is reminiscent of PMR1, which endonucleolytically cleaves actively translating mRNAs on polysomes (Yang and Schoenberg, 2004). The cleavage site on the targeted mRNAs was between the UG dinucleotides of two overlapping repeats of AYUGA, which needs not to be base paired and be part of a loop in a stem-loop structure (Chernokalskaya et al., 1997). Further more it was shown that binding of the PMR1 endonuclease to its target stem loop could be

abrogated by binding of other proteins to sequences overlapping the stem loop structure (Brock and Shapiro, 1983, Dodson and Shapiro, 1997, Cunningham et al., 2000).

Given the fact that the UGAAG motif alone appears to be insufficient for SOX mediated cleavage, the possibility that target recognition also involves binding of a structured RNA element was investigated. Thus to identify whether the pentameric sequence was embedded in a structured element, *in silico* folding studies were undertaken on the three identified mRNA sequences (Covarrubias et al., 2011). Based on these studies, a persistent fold was observed for the GFP transcript, which gave the most consistent stem loop structure when tested over a range of nucleotide lengths. As a result, most subsequent biochemical and crystallization studies focused on the minimal 51 nucleotides stretch that retained the overall stem loop structure in the vicinity of the UGAAG motif. From these *in silico* studies, the cleavage site was predicted to reside within a loop directly downstream of the UGAAG motif located in a Watson Crick based paired stem (Figure 4-2 C, Section 4.2.2). The 3D structure predictions for this stem loop produced 114 structures thus reflecting its overall flexibility. The comparison of these structures with the ds-DNA in the SOX-DNA complex allowed the RNA stem loop to be fitted into the catalytic region using the DNA co-ordinates as a guide (Figure 4-4 A, Section 4.2.3) (Bagneris et al., 2011). The most energetically favourable SOX complex model was obtained with a straight stem loop conformation, however, this is very much based on the DNA co-ordinates and thus likely to be biased. The contacts that are made between SOX and the stem loop are at the NLS loop, which is known to bind DNA and could possibly bind RNA as well, but the stem loop does not make contact with the C-terminal helix, that contains residues identified as key in HSO. These modelling studies nonetheless provide evidence that RNA stem loops can be accommodated within the active site of SOX and that the loop containing the cleavage site would be appropriately positioned relative to the catalytic residues D221 and E244 for processing (Figure 4-4 D, Section 4.2.3). In order to investigate whether the 51 nucleotides GFP and 58 nucleotides HBB RNAs were indeed substrates for SOX, EMSAs were performed that revealed the formation of shifted complexes in both cases using wild-type SOX and the D221S and E244S catalytic mutants (Figure 5-3 B). These were followed by RNase assays, which showed

unequivocally that the 51 nucleotides GFP stem loop is endonucleolytically degraded by wild-type SOX whilst this activity is abrogated in the case of the D221S and E244S mutants. Although there is evidence for cleavage of HBB, its high susceptibility to degradation renders these results more speculative. On aggregate, however, these findings are in agreement with the literature confirming that the same catalytic residues are required for both the DNase and RNase activities of SOX (Glaunsinger et al., 2005, Glaunsinger and Ganem, 2004). To ascertain whether SOX has enhanced affinity for 51 nucleotides GFP stem loop over ssRNA and ssDNA substrates, FA was used to determine its K_d. This was found to be 4.6 μM (Figure 5-4) and thus represented a 20 fold increase in affinity compared to ssRNA-5'P and dsRNA-5'P, which are exonucleotically cleaved by SOX (Bagneris et al., 2011). The affinity of SOX for 51 GFP is comparable to the affinity of SOX for the ds-DNA-5'P; the latter being the target of SOX during viral genome packaging, for which the same catalytic machinery is used (Bagneris et al., 2011, Glaunsinger et al., 2005).

Several mutants were reported by the Ganem and Glaunsinger labs respectively, that negatively impact on HSO, but are otherwise proficient in DNA processing. These mutants include A61T, D474N and Y477Stop (Glaunsinger et al., 2005). RNase cleavage TBE-Urea gel assays performed using D474N and Y477Stop, which are located towards the C-terminus of the molecule, revealed that these mutants are attenuated in their ability to associate with 51 nucleotides GFP stem loop (Figure 6-1 and Figure 5-6).

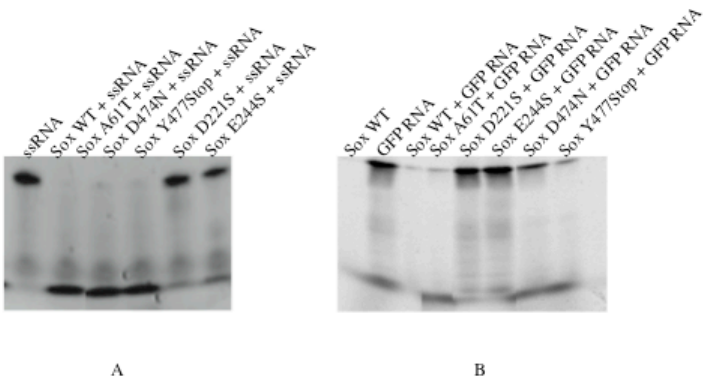


Figure 6-1: HSO SOX Mutants do not Abrogate ssRNA Turnover and do Abrogate RNA GFP Stem Loop

A) TBE-Urea gel taken from the (Bagneris et al., 2011) showing that the HSO mutants (A61T, D474N and Y477Stop) do not abrogate the turnover of ssRNA. B) TBE-Urea gel from this study showing that the HSO mutants do abrogate turnover of the 51 nucleotides GFP stem loop RNA.

By contrast, A61T, situated at the N-terminus is only marginally perturbed. Previously published data from the Barrett Group revealed that these mutants had no impact on the ability of SOX to exonucleolytically degrade ssRNA substrates. (Figure 6-1) (Bagneris et al., 2011). This is a further indication that a stem loop structure is the cognate substrate for SOX, when inducing HSO *in vivo*.

As mentioned D474N and Y477Stop were shown to be defective in their capacity to endonucleolytically cleave 51 nucleotides GFP stem loop whilst A61T has near wild-type activity. These results suggest that the C-terminal helix is required for endonucleolytic cleavage. To further investigate the involvement of this helix, additional mutations along the helix were constructed that comprised H450A, R451A, N458A and R462A. H450A and R462A substantially reduced the ability of SOX to turnover the stem loop, while H450A and R451A did not (Figure 5-6). From the EMSA experiments it can be seen that R462A, D474N and Y477Stop affect binding of the RNA stem loop (Figure 5-3), indicating that the majority of this helix is involved in binding the RNA stem loop. In order to achieve this, the stem loop would have to be substantially bent and would thus have to be intrinsically highly flexible, which is in keeping with the tertiary structure predictions, in which the RNA property to bend was illustrated (Figure 4-4 A).

To investigate the importance of the UGAAG motif within the GFP sequence the three identified endonucleolytically cleaved mRNA targets containing the UGAAG motif were aligned (Figure 4-2 D and E). This allowed the identification of conserved GA rich nucleotides in the sequence upstream of the cleavage site and the UGAAG motif. Thus a mutant GFP 51 sequence was constructed in which UGAAG was replaced by UCUCU to investigate whether the UGAAG motif is indispensible. This sequence folded into two possible stem loops, containing larger top loops. On contained one stem loop and the other contained two stem loops (Figure 4-5 A and B, Section 4.2.4.). The UCUCU construct was

efficiently degraded by SOX (Figure 5-5 C). These results suggest that SOX is likely to recognise not just the UGAAG motif, but more likely a combination of GA rich and structural elements. Many RNA elements are degenerate and thus hard to identify. With deeper computational analysis a better motif could be identified using published data on SOX induced mRNA depletion in cells. This could also mean that SOX recognised structural features that were still maintained despite the mutation. This would be at least partially supported by the capacity of SOX to degrade the HBB and DsRed2 mRNA transcripts that despite containing stem loop structures have quite distinct folding patterns. But in addition to the UGAAG motif they have in common a UGA followed by a GAA in the upstream regions of the cleavage site. To test this hypothesis more fully, other structured RNA molecules lacking the UGAAG motif and/or GA rich elements should be tested for their capacity to be endonucleolytically processed by SOX. These experiments were not possible within the time constraints of the project. Interestingly UGAAG is a motif associated with the splicing mechanism. UGAAG motif is overrepresented in both in mice and constitutive and cryptic exons in alternative splicing in Arabidopsis (Zavolan et al., 2003, Cech, 1990, Thomas et al., 2012). UGAAG is found in the 3′ splice site, as e.g. in the meiotic recombination protein (REC102) gene of *Trichinella spiralis* (Ma and Xuhua Xia, 2011, Pettitt et al., 2008) and exon where it is associated with the EJC and likely to be overrepresented.

SOX is known to solely target mRNAs and to contain a NLS (Glaunsinger et al., 2005). Its DNase activity is associated with its export to the nucleus. SOX is thus found in both the cytoplasm and the nucleus. Its RNase activity is thought to occur in the cytoplasm (Lee and Glaunsinger, 2009). KSHV genome contains mainly intronless genes, which are then translated in the cytoplasm by the host ribosomes (Rezaee et al., 2006, Schumann et al., 2013). However, the processes of transcription, splicing and mRNA nuclear export are intimately linked in the host. Thus, this poses a significant barrier to the viral RNA transcript export, translation and hence viral replication (Schumann et al., 2013). The issue is overcome by the ORF57 KSHV protein, which allows the efficient export of intronless viral mRNA from the nucleus to the cytoplasm (Jackson et al., 2012, Malik and Schirmer, 2006, Schumann et al., 2013).

It had been suggested, in the literature and by collaborators, that the cytoplasmic exonuclease Xrn1 was involved in HSO (Covarrubias et al., 2011). Xrn1 is the most common and highly conserved 5' to 3' exonuclease in eukaryotes. The highest conservation is present in the N-terminal region of the protein, which was predicted to be the region of Xrn1 most likely to interact with SOX from yeast two hybrid experiments conducted by collaborators (data not shown). *K. lactis* Xrn1 has a 48% sequence identity with human Xrn1 (Chang et al., 2011). As it had been suggested that though SOX does have an intrinsic 5' to 3' exonuclease activity this was not sufficient to account for the rapid mRNA decay, thus Xrn1s involvement was investigated (Covarrubias et al., 2011, Bagneris et al., 2011). Via the EMSA and RNase activity experiments, it could be demonstrated that Xrn1 was not able to bind the 51 nucleotides stem loop nor cleave it (Figure 5-3 and Figure 5-6). This is in agreement with Xrn1 being a 5' to 3' exonuclease that requires a 5' monophosphate. It was possible to show that SOX and *K. lactis* Xrn1 did interact directly using a pull down experiment (Figure 5-1) and that this interaction could be independently confirmed using MST where a binding affinity of 0.865 μM was obtained (Figure 5-2). In the RNA binding TBE gel assays, a higher molecular weight band appeared when SOX, Xrn1 and the GFP stem loop RNA were simultaneously present, which could represent Xrn1 in a complex with SOX and the GFP stem loop decay intermediate (Figure 5-3). In the RNase assays, the presence of Xrn1 appeared to reduce the endonucleolytic activity of SOX. This could indicate that when the SOX-Xrn1 complex forms the handing over of the cleaved 5' end of the stem loop to Xrn1 is the rate-limiting step. Thus shorter mRNAs would lead to a slower turnover. But as SOX and Xrn1 are likely to encounter longer RNAs in the cell the slowing hand over will lead to a more rapid degradation of the longer cellular RNAs by Xrn1 after initial endonucleolytic cleavage by SOX. This could also be an indication that SOX and Xrn1 are part of a wider degradation complex, e.g. linked to NMD.

In an attempt to elucidate the nature of the Xrn1, SOX and stem loop RNA complex crystallization trials were undertaken. Unfortunately no complex could be obtained. Various complex combinations were set up involving Xrn1-SOX, Xrn1-SOX-Stem Loop RNA (23 and 51 nucleotides GFP RNA and 58 nucleotides HBB

RNA), SOX-Stem Loop RNA and SOX 244-Stem Loop RNA. Crystals were obtained from set up complex crystallization trials and shot at Diamond Light Source (Figure 5-7). The mutant SOX 244 was used to inhibit any potential cleavage at 16 °C and the 23 GFP stem loop was engineered as to reduce flexibility in the RNA and allow different or tighter packing. It was possible to obtain a dimer crystal structure of SOX 244 at 3 Å from trials that contained the 51 nucleotides RNA stem loop. Initially positive missing density was visible in the active site of one monomer. The active site was not obstructed and would have been able to accommodate a shorter version of the stem loop. However, during subsequent refinement rounds the 2Fo-Fc maps failed to show any convincing density when contoured above 1σ. The original difference density could therefore have either originated from the solvent, though this seems to be unlikely based on its size and shape or be due to partially occupied/highly mobile nucleotides.

6.2. Conclusion

It was possible to determine that the UGAAG motif is overrepresented in the host genome. Both UGAAG and UCUCU motifs lead to endonucleolytic cleavage of stem loops. The endonucleolytic cleavage of stem loop RNA is inhibited by HSO mutants, which do not interfere with exonucleolytic activity seen with dsDNA-5'P and ssRNA-5'P. SOX has a 20 fold high affinity to the RNA stem loop then to ssRNA-5'P and this affinity is similar to its affinity to dsDNA-5'P. SOX interact directly with Xrn1 the 5' to 3' exonuclease with an affinity of 0.865 μM.

6.3. Future Work

To further substantiate these findings and hypothesis further work is needed. Firstly to further strengthen these SOX and RNA stem loop binding results titration with other unlabelled RNAs known to not contain the UGAAG motif and not to specifically bind SOX should be undertaken to exclude non-specific binding. In addition using the FLA300 it should be possible to quantify the shifted RNA in a titration experiments, such as titrating Xrn1 to SOX with the RNA stem loop and Sox to the RNA stem loop. In attempt to clarify the results a trials with different EMSA gel matrices should be undertaken. Secondly, further stem loops need to be

tested for their ability to be bound and endonucleolytically cleaved by SOX. It would be advantageous to be able to analyse the binding mode of SOX and the UGAAG and UCUCU containing stem loops using protein-RNA foot printing, as binding of the RNA to the protein is likely to also influence its three dimensional structure. Further, efforts will be required to obtain crystal structures of complexes. The use of a chimeric RNA-DNA sequence based on the top half of the stem loop containing the UGAAG motif and the bottom half of the dsDNA that was previously co-crystallized may be advantageous since it may promote favourable crystal packing for complex formation. To further the hypothesis that SOX targets the mRNAs on the basis of the exonic splicing signals and the linked to UGAAG and GA rich motifs, an in depth computational analysis of the host genes and mRNAs that are known to be targeted need to be undertaken as well as a detailed analysis of the distribution of the UGAAG motif within the viral genome and host genome. To further substantiate the Xrn1-SOX complex size-exclusion chromatography should be undertaken to test the binding and strength of the complex. And finally *in vivo* experiments should be conducted to localize Xrn1 in the presence of SOX and to identify any other possible binding partners that might be involved in the host mRNA maturation process, such as EJC.

Reference:

ABELSON, J. 1979. RNA processing and the intervening sequence problem. *Annual Review of Biochemistry,* 48**,** 1035-1069.

ADAMS, P. D., AFONINE, P. V., BUNKÓCZI, G., CHEN, V. B., DAVIS, I. W., ECHOLS, N., HEADD, J. J., HUNG, L. W., KAPRAL, G. J., GROSSE-KUNSTLEVE, R. W., MCCOY, A. J., MORIARTY, N. W., OEFFNER, R., READ, R. J., RICHARDSON, D. C., RICHARDSON, J. S., TERWILLIGER, T. C. & ZWART, P. H. 2010. PHENIX: a comprehensive Python-based system for macrmolecular structure solution. *Acta Crystallographica Section D: Biological Crystallography,* 66**,** 213-221.

ALCALA, J. R., E., G. & PRENDERGAST, F. G. 1987. Fluorescence lifetime distributions in proteins. *Biophysical Journal,* 51**,** 597-604.

ALLMANG, C., PETFALSKI, E., PODTELEJNIKOV, A., MANN, M., TOLLERVEY, D. & MITCHELL, P. 1999. The yeast exosome and human PM–Scl are related complexes of $5' \rightarrow 3'$ exonucleases. *Genes & Development,* 13**,** 2148-2158.

ALTSCHUL, S. F., GISH, W., MILLER, W., MYERS, E. W. & LIPMAN, D. J. 1990. *Basic local alignment search tool (Blast)* [Online]. Available: http://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&BLAST_PROGRAMS=megaBlast&PAGE_TYPE=BlastSearch&SHOW_DEFAULTS=on&LINK_LOC=blasthome [Accessed 20 February 2012].

AMRANI, N., SACHS, M. S. & JACOBSON, A. 2006. Early nonsense: mRNA decay solves a translational problem. *Nature Reviews Molecular Cell Biology,* 7**,** 415-425.

AOKI, Y., YARCHOAN, R., BRAUN, J., IWAMOTO, A. & TOSATO, G. 2000. Viral and cellular cytokines in AIDS-related malignant lymphomatous effusions. *Blood,* 96**,** 1599-1601.

ARNDT, U. W., LONG, J. V. P. & DUNCUMB, P. 1998. A microfocus X-ray tube used with focusing collimators. *Journal of Applied Crystallography,* 31**,** 936-944.

ARVANITAKIS, L., MESRI, E. A., NADOR, R. G., SAID, J. W., ASCH, A. S., KNOWLES, D. M. & CESARMAN, E. 1996. Establishment and characterization of a primary effusion (body cavity-based) lymphoma cell line (BC-3) harboring kaposi's sarcoma-associated herpesvirus (KSHV/HHV-8) in the absence of Epstein-Barr virus. *Blood,* 88**,** 2648–2654.

BAASKE, P., WIENKEN, C. J., REINECK, P., DUHR, S. & BRAUN, D. 2010. Optical Thermophoresis for Quantifying the Buffer Dependence of Aptamer Binding. *Angewandte Chemie International Edition* 49**,** 2238 –2241.

BAGNERIS, C., BRIGGS, L. C., SAVVA, R., EBRAHIMI, B. & BARRETT, T. E. 2011. Crystal structure of a KSHV–SOX–DNA complex: insights into the molecular mechanisms underlying DNase activity and host shutoff. *Nucleic Acids Research,* 39**,** 5744-5756.

BARREAU, C., PAILLARD, L. & OSBORNE, H. B. 2006. AU-rich elements and associated factors: are there unifying principles? *Nucleic Acids Research,* 33**,** 7138-7150.

BATTYE, T. G., KONTOGIANNIS, L., JOHNSON, O., POWELL, H. R. & LESLIE, A. G. 2011. iMOSFLM: a new graphical interface for diffraction-image

processing with MOSFLM. *Acta Crystallographica Section D: Biological Crystallography,* 67**,** 271-281.
BEELMAN, C. A. & PARKER, R. 1995. Degradation of mRNA in eukaryotes. *Cell,* 81**,** 179-183.
BEHM-ANSMANT, I., GATFIELD, D., REHWINKEL, J., HILGERS, V. & IZAURRALDE, E. 2007. A conserved role for cytoplasmic poly(A)-binding protein 1 (PABPC1) in nonsense-mediated mRNA decay. *EMBO Journal,* 26**,** 1591-1601.
BENSON, D. A., CAVANAUGH, M., CLARK, K., KARSCH-MIZRACHI, I., LIPMAN, D. J., OSTELL, J. & SAYERS, E. W. 2012. GenBank. 2009 ed.: Nucleic Acids Research.
BERGET, S. M. 1995. Exon recognition in vertebrate splicing. *Journal of Biological Chemistry,* 270**,** 2411-2414.
BIRNEY, E., STAMATOYANNOPOULOS, J. A., DUTTA, A. & AL., E. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature,* 447**,** 799-816.
BLACK, D. L. 2003. Mechanisms of alternative pre-messenger RNA splicing. *Annual Review of Biochemistry,* 72**,** 291-336.
BOSHOFF, C. & CHANG, Y. 2001. Kaposi's sarcoma-associated herpesvirus: a new DNA tumor virus. *Annual Review of Medicine,* 52**,** 453-470.
BOSHOFF, C. & WEISS, R. A. 1998. Kaposi's sarcoma-associated herpesvirus. *Advances in Cancer Research,* 75**,** 57–86.
BOYNE, J. R., COLGAN, K. J. & WHITEHOUSE, A. 2008. Recruitment of the complete hTREX complex is required for Kaposi's sarcoma-associated herpesvirus intronless mRNA nuclear export and virus replication. *Plos Pathogens,* 4**,** e1000194.
BOYNE, J. R., JACKSON, B. R., TAYLOR, A., MACNAB, S. A. & WHITEHOUSE, A. 2010. Kaposi's sarcoma-associated herpesvirus ORF57 protein interacts with PYM to enhance translation of viral intron-less mRNAs. *EMBO Journal,* 29**,** 1851-1864.
BOYNE, J. R. & WHITEHOUSE, A. 2006. Nucleolar trafficking is essential for nuclear export of intron-less herpesvirus mRNA. *Proceedings of the National Academy of Sciences,* 103**,** 15190-15195.
BOYNE, J. R. & WHITEHOUSE, A. 2009. Nucleolar disruption impairs Kaposi's sarcoma-associated herpesvirus ORF57-mediated nuclear export of intron-less viral mRNAs. *FEBS Letter,* 583**,** 3549-3556.
BRAWERMAN, G. 1981. The role of the poly(A) sequence in mammalian messenger RNA. *Critical Reviews in Biochemistry and Molecular Biology,* 10**,** 1-38.
BRICOGNE, D. 1993. Direct phase determination by entropy maximization and likelihood ranking: status report and perspectives. *Acta Crystallographica Section D: Biological Crystallography,* 49**,** 37-60.
BRICOGNE, G. 1997. The Bayesian Statistical Viewpoint on Structure Determination: Basic Concepts and Examples. *Methods in Enzymology,* 276A**,** 361-423.
BRION, P. & WESTHOF, E. 1997. Hierarchy and dynamics of RNA folding. *Annual Review of Biophysics and Biomolecular Structure,* 26**,** 113-137.
BROCK, M. L. & SHAPIRO, D. J. 1983. Estrogen stabilizes vitellogenin mRNA against cytoplasmic degradation. *Cell,* 34**,** 207-214.

BROWER-SINNING, R., CARTER, D. M., CREVAR, C. J., GHEDIN, E., ROSS, T. M. & BENOS, P. V. 2009. The role of RNA folding free energy in the evolution of the polymerase genes of the influenza A virus. *Genome Biology,* 10**,** R1-8.

BROWN, J. C. & NEWCOMB, W. W. 2011. Herpesvirus Capsid Assembly: Insights from Structural Analysis. *Current Opinion in Virology,* 1**,** 142-149.

BROWN, N. P. 2013. *MView* [Online]. Available: http://www.ebi.ac.uk/Tools/msa/mview/ [Accessed 25 September 2013].

BROWN, N. P., LEROY, C. & SANDER, C. 1998. MView: a web-compatible database search or multiple alignment viewer. *Bioinformatics,* 14**,** 380-381.

BRUNGER, A. T. 1992. The Free R Value: a Novel Statistical Quantity for Assessing the Accuracy of Crystal Structures. *Nature,* 355**,** 472-474.

BUISSON, M., GEOUI, T., FLOT, D., TARBOURIECH, N., RESSING, M. E., WIERTZ, E. J. & BURMEISTER, W. P. 2009. A bridge crosses the active-site canyon of the Epstein-Barr virus nuclease with DNase and RNase activities. *Journal of Molecular Biology,* 391**,** 717-728.

BUJNICKI, J. M. & RYCHLEWSKI, L. 2001. The herpesvirus alkaline exonuclease belongs to the restriction endonuclease PD-(D/E)XK superfamily: insight from molecular modeling and phylogenetic analysis. *Virus Genes,* 22**,** 219–230.

BURGE, C. B., TUSCHL, T. & SHARP, P. A. 1999. *Splicing of precursors to mRNAs by the spliceosomes,* Cold Spring Harbor, NY, Cold Spring Harbor Press.

BURGER, R., WENDLER, J., ANTONI, K., HELM, G., KALDEN, J. R. & GRAMATZKI, M. 1994. Interleukin-6 production in B-cell neoplasias and Castleman's disease: evidence for an additional paracrine loop. *Annals of Hematology,* 69**,** 25–31.

CANTOR, C. R. & SCHIMMEL, P. 1980. *Biophysical Chemistry - Part II: Techniques for the Study of Biological Structure and Function,* San Francisco, W. H. Freeman.

CAO, D. & PARKER, R. 2003. Computational modeling and exper- imental analysis of nonsense-mediated decay in yeast. *Cell,* 113**,** 533-545.

CAPRIOTTI, E. & MARTI-RENOM, M. A. 2010. Quantifying the relationship between sequence and three-dimensional structure conservation in RNA. *BMC Bioinformatics,* 11**,** 1-10.

CASELLI, E., ZATELLI, M. C., RIZZO, R., BENEDETTI, S., MARTORELLI, D., TRASFORINI, G., CASSAI, E., DEGLI UBERTI, E. C., DI LUCA, D. & DOLCETTI, R. 2012. Virologic and Immunologic Evidence Supporting an Association between HHV-6 and Hashimoto's Thyroiditis. *Plos Pathogens,* 8**,** e1002951.

CASTLE, J. C., ZHANG, C., SHAH, J. K., KULKARNI, A. V., KALSOTRA, A., COOPER, T. A. & JOHNSON, J. M. 2008. Expression of 24,426 human alternative splicing events and predicted cis regulation in 48 tissues and cell lines. *Nature Genetics,* 40**,** 1416-1425.

CECH, T. R. 1990. Self-Splicing Of Group I Introns. *Annual Review of Biochemistry,* 59**,** 543-568.

CESARMAN, E. & KNOWLES, D. M. 1999. The role of Kaposi's sarcoma-associated herpesvirus (KSHV/HHV-8) in lymphoproliferative diseases. *Seminar in Cancer Biology,* 9**,** 165–174.
CHAMORRO, M., PARKIN, N. & VARMUS, H. E. 1992. An RNA pseudoknot and an optimal heptameric shift site are required for highly efficient ribosomal frameshifting on a retroviral messenger RNA. *Proceedings of the National Academy of Sciences,* 89**,** 713-717.
CHANDRIANI, S. & GANEM, D. 2007. Host Transcript Accumulation during Lytic KSHV Infection Reveals Several Classes of Host Responses. *PLoS ONE,* 8**,** 1-10.
CHANG, J. H., XIANG, S., XIANG, K., MANLEY, J. L. & TONG, L. 2011. Structural and biochemical studies of the 5′ →3′ exoribonuclease Xrn1. *Nature Structural & Molecular Biology,* 18**,** 270-276.
CHANG, Y. F., IMAM, J. S. & WILKINSON, M. F. 2007. The nonsense-mediated decay RNA surveillance pathway. *Annual Review of Biochemistry,* 76**,** 51-74.
CHASIN, L. A. 2007. Searching for splicing motifs. *advances in Experimental Medicine and Biology* 623**,** 85-106.
CHEN, C.-Y. A. & SHYU, A.-B. 2003. Rapid deadenylation triggered by a nonsense codon precedes decay of the RNA body in a mammalian cytoplasmic nonsense-mediated decay pathway. *Molecular and Cellular Biology,* 23**,** 4805-4813.
CHEN, C.-Y. A. & SHYU, B.-A. 2010. Mechanisms of deadenylation- dependent decay. *Wiley Interdisciplinary Reviews: RNA,* 2**,** 167-183.
CHEN, C. Y., GHERZI, R., ONG, S. E., CHAN, E. L., RAIJMAKERS, R., PRUIJN, G. J., STOECKLIN, G., MORONI, C., MANN, M. & KARIN, M. 2001. AU binding proteins recruit the exosome to degrade ARE-containing mRNAs. *Cell,* 107**,** 451-464.
CHEN, C. Y. & SHYU, A. B. 1995. AU-rich elements: characterization and importance in mRNA degradation. *Trends In Biochemical Science,* 20**,** 465-470.
CHEN, J., CHIANG, Y. C. & DENIS, C. L. 2002. CCR4, a 3′ → 5′ poly(A) RNA and ssDNA exonuclease, is the catalytic component of the cytoplasmic deadenylase. *EMBO Journal,* 21**,** 1414-1426.
CHEN, V. B., ARENDALL, W. B. I., HEADD, J. J., KEEDY, D. A., IMMORMINO, R. M., KAPRAL, G. J., MURRAY, L. W., RICHARDSON, J. S. & RICHARDSON, D. C. 2010. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D: Biological Crystallography,* 66**,** 12-21.
CHENG, H., DUFU, K., LEE, C.-S., HSU, J. L., DIAS, A. & REED, R. 2006. Human mRNA export machinery recruited to the 5′ end of mRNA. *Cell,* 127**,** 1389-1400.
CHERNOKALSKAYA, E., DOMPENCIEL, R. E. & SCHOENBERG, D. R. 1997. Cleavage properties of a polysomal ribonuclease involved in the estrogen-regulated destabilization of albumin mRNA. *Nucleic Acids Research,* 25**,** 735-742.
CHI, B., WANG, Q., WU, G., TAN, M., WANG, L., SHI, M., CHANG, X. & CHENG, H. 2013. Aly and THO are required for assembly of the human TREX

complex and association of TREX components with the spliced mRNA. *Nucleic Acids Research,* 41**,** 1294-1306.
CLYDE, K. & GLAUNSINGER, B. A. 2011. Deep Sequencing Reveals Direct Targets of Gammaherpesvirus-Induced mRNA Decay and Suggests That Multiple Mechanisms Govern Cellular Transcript Escape. *PLoS ONE,* 6.
COLLER, J. & PARKER, R. 2004. Eukaryotic mRNA decapping. *Annual Review of Biochemistry,* 74**,** 861-890.
CONRAD, N. K. & STEITZ, J. A. 2005. A Kaposi's sarcoma virus RNA element that increases the nuclear abundance of intronless transcripts. *EMBO Journal,* 24**,** 1831-1841.
CONTI, E. & IZAURRALDE, E. 2005. Nonsense-mediated mRNA decay: molecular insights and mechanistic variations across species. *Current Opinion in Cell Biology,* 17**,** 316-325.
COUGOT, N., VAN DIJK, E., BABAJKO, S. & SÉRAPHIN, B. 2004. Cap-tabolism. *Trends In Biochemical Science,* 29**,** 436-444.
COVARRUBIAS, S., GAGLIA, M. M., KUMAR, G. R., WONG, W., JACKSON, A. O. & GLAUNSINGER, B. A. 2011. Coordinated destruction of cellular messages in translation complexes by the gammaherpesvirus host shutoff factor and the mammalian exonuclease Xrn1. *PLoS Pathogens,* 7**,** e1002339.
CUNNINGHAM, K. S., DODSON, R. E., A., N. M., SHAPIRO, D. J. & R., S. D. 2000. Vigilin binding selectively inhibits cleavage of the vitellogenin mRNA 3′-UTR by the mRNA endonuclease PMR-1. *Proceedings of the National Academy of Sciences,* 97**,** 12498-12502.
DAHLROTH, S.-L., GURMU, D., HAAS, J., ERLANDSEN, H. & NORDLUND, P. 2009. Crystal structure of the shutoff and exonuclease protein from the oncogenic Kaposi's sarcoma-associated herpesvirus. *FEBS Journal,* 276**,** 6636–6645.
DANCKWARDT, S., NEU-YILIK, G., THERMANN, R., FREDE, U., HENTZE, M. W. & KULOZIK, A. E. 2002. Abnormally spliced beta-globin mRNAs: a single point mutation generates transcripts sensitive and insensitive to nonsense-mediated mRNA decay. *Blood,* 99**,** 1811-1816.
DESAILLOUD, R. & HOBER, D. 2009. Viruses and thyroiditis: an update. *Journal of Virology,* 6.
DIAMOND, R. 1985. Real Space Refinement. *Methods in Enzymology,* 115**,** 237-252.
DIMAANO, C. & ULLMAN, K. S. 2004. Nucleocytoplasmic transport: Integrating mRNA production and turnover with ex- port through the nuclear pore. *Molecular and Cellular Biology,* 24**,** 3069-3076.
DING, Y. & LAWRENCE, C. E. 2003. A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Research,* 31**,** 7280-7301.
DOCKRELL, D. H. 2003. Human herpesvirus 6: molecular biology and clinical features. *Journal of Medical Microbiology,* 52**,** 5-18.
DODSON, R. E. & SHAPIRO, D. J. 1997. Vigilin, an ubiquitous protein with 14 KH domains, is the estrogen-inducible vitel- logenin mRNA 3′-untranslated region binding protein. *Journal of Biological Chemistry,* 272**,** 12249-12252.
DOMA, M. K. & R., P. 2006. Endonucleolytic cleavage of eukaryotic mRNAs with stalls in translation elongation. *Nature,* 440**,** 561-564.

DONELLO, J. E., LOEB, J. E. & HOPE, T. J. 1998. Woodchuck hepatitis virus contains a tripartite posttranscriptional regulatory element. *Journal of Virology,* 72**,** 5085-5092.
DOURMISHEV, L. A., DOURMISHEV, A. L., PALMERI, D., SCHWARTZ, R. A. & LUKAC, D. M. 2003. Molecular Genetics of Kaposi's Sarcoma-Associated Herpesvirus (Human Herpesvirus 8) Epidemiology and Pathogenesis. *Microbiology and Molecular Biology Reviews,* 67**,** 175-212.
DRENTH, J. 2007. *Principles of Protein X-Ray Crystallography,* New York, Springer.
DRIESSEN, H. P. C. & TICKLE, I. J. 1996. *Molecular Replacement using known structural information,* Totowa, New Jersey, Humana Press.
DU, H. & ROSBASH, M. 2002. The U1 snRNP protein U1C recognizes the 5 splice site in the absence of base pairing. *Nature,* 419**,** 86-90.
DUHR, S. & BRAUN, D. 2006. Why molecules move along a temperature gradient. *Proceedings of the National Academy of Sciences,* 103**,** 19678–19682.
EBERLE AB, LYKKE-ANDERSEN, S., MUHLEMANN, O. & JENSEN, T. H. 2009. SMG6 promotes endonucleolytic cleavage of non-sense mRNA in human cells. *Nature Structural & Molecular Biology,* 16**,** 49-55.
EMSLEY, P. & COWTAN, K. 2004. Coot: model‐building tools for molecular graphic. *Acta Crystallography,* D60**,** 2126‐2132.
EMSLEY, P., LOHKAMP, B., SCOTT, W. G. & COWTAN, K. 2010. Features and development of Coot. *Acta Crystallographica Section D: Biological Crystallography,* 66**,** 486-501.
FASKEN, M. B. & CORBETT, A. H. 2005. Process or perish: Quality control in mRNA biogenesis. *Nature Structural & Molecular Biology,* 12**,** 482-488.
FAUSTINO, N. A. & COOPER, T. A. 2003. Pre-mRNA splicing and human disease. *Genes & Development,* 17**,** 419-437.
FISCHETTI, R. F., XU, S., YODER, D. W., BECKER, M., NAGARAJAN, V., SANISHVILI, R., HILGART, M. C., STEPANOV, S., MAKAROV, O. & SMITH, J. L. 2009. Mini-beam collimator enables microcrystallography experiments on standard beamlines. *Journal of synchrotron radiation,* 16**,** 217-225.
FORTELLE, E. & BRICOGNE, G. 1997. Maximum-Likelihood Heavy-Atom Parameter Refinement for Multiple Isomorphous Replacement and Multiwavelength Anomalous Diffraction Methods. *Methods in Enzymology,* 276**,** 472-494.
GAGLIA, M. M., COVARRUBIAS, S., WONG, W. & GLAUNSINGER, B. A. 2012. A common strategy for host RNA degradation by divergent viruses. *Journal of Virology,* 86**,** 9527-9530.
GAGLIA, M. M. & GLAUNSINGER, B. A. 2010. Viruses and the cellular RNA decay machinery. *Interdisciplinary Reviews: RNA* 1**,** 47-59.
GARDNER, L. B. 2010. Nonsense-Mediated RNA Decay Regulation by Cellular Stress: Implications for Tumorigenesis. *Molecular Cancer Research,* 8**,** 295-308.
GARNEAU, N. L., WILUSZ, J. & WILUSZ, C. J. 2007. The highways and byways of mRNA decay. *Nature Reviews Molecular Cell Biology,* 10**,** 113-126.
GATFIELD, D., UNTERHOLZNER, L., CICCARELLI, F. D., BORK, P. & IZAURRALDE, E. 2003. Nonsense-mediated mRNA decay in Drosophila:

at the intersection of the yeast and mammalian pathways. *EMBO Journal,* 22**,** 3960-3970.
GEBAUER, F. & HENTZE, M. W. 2004. Molecular mechanisms of translational control. *Nature Reviews Molecular Cell Biology,* 5**,** 827-835.
GIORGI, C. & MOORE, M. J. 2007. The nuclear nurture and cytoplasmic nature of localized mRNPs. *Seminars in Cell & Developmental Biology,* 18**,** 186-193.
GLAUNSINGER, B., CHAVEZ, L. & GANEM, D. 2005. The exonuclease and host shutoff functions of the SOX protein of Kaposi's sarcoma-associated herpesvirus are genetically separable. *Journal of Virology,* 79**,** 7396-7401.
GLAUNSINGER, B. & GANEM, D. 2004. Lytic KSHV infection inhibits host gene expression by accelerating global mRNA turnover. *Molecular Cell,* 13**,** 713–723.
GLAUNSINGER, B. A. & GANEM, D. E. 2006. Messenger RNA turnover and its regulation in herpesviral infection. *Adv Virus Res,* 66**,** 337-94.
GOLDSTEIN, J. N. & WELLER, S. K. 2004. The exonuclease activity of HSV-1 UL12 is required for in vivo function. *Virology,* 244**,** 442–457.
GRABOWSKI, P. J. 1998. Splicing regulation in neurons: Tinkering with cell-specific control. *Cell,* 92**,** 709-712.
GRATACÓS, F. M. & BREWER, G. 2010. The role of AUF1 in regulated mRNA decay. *Interdisciplinary Reviews: RNA,* 1**,** 457–473.
GRAY, K. S., COLLINS, C. M. & SPECK, S. H. 2012. Characterization of Omental Immune Aggregates during Establishment of a Latent Gammaherpesvirus Infection. *PLoS ONE* 7**,** 1-10.
HARDY, W. R. & SANDRI-GOLDIN, R. M. 1994. Herpes simplex virus inhibits host cell splicing, and regulatory protein ICP27 is required for this effect. *Journal of Virology,* 68**,** 7790-7799.
HASTINGS, M. L. & KRAINER, A. R. 2001. Pre-mRNA splicing in the new millennium. *Current Opinion in Cell Biology,* 13**,** 302-309.
HELLMAN, K. M. & FRIED, M. G. 2007. Electrophoretic mobility shift assay (EMSA) for detecting protein–nucleic acid interactions. *Nature Protocols,* 2**,** 1849-1861.
HENDRICKSON, D. G., HOGAN, D. J., MCCULLOUGH, H. L., MYERS, J. W., HERSCHLAG, D., FERRELL, J. E. & BROWN, P. O. 2009. Concordant regulation of translation and mRNA abundance for hundreds of targets of a human microRNA. *PLoS Biology,* 7**,** e1000238.
HENDRICKSON, W. A., HORTON, J. R. & LEMASTER, D. M. 1990. Selenomethionyl proteins produced for analysis by multiwavelength anomalous diffraction (MAD): a vehicle for direct determination of three-dimensional structure. *EMBO Journal***,** 1665-1672.
HOGGAN, D. B., CHAO, J. A., PRASAD, G. S., STOUTA, C. D. & WILLIAMSON, J. R. 2003. Combinatorial crystallization of an RNA±protein complex. *Acta Crystallographica Section D,* D59**,** 466-473.
HOLLIEN, J. & WEISSMAN, J. S. 2006. Decay of endoplasmic reticulum-localized mRNAs during the unfolded protein response. *Science,* 313**,** 104-107.
HOUSELEY, J., LACAVA, J. & TOLLERVEY, D. 2006. RNA-quality control by the exosome. *Nature Reviews Molecular Cell Biology,* 7**,** 529-539.
HUANG, J. & LIANG, T. J. 1993. A novel hepatitis B virus (HBV) genetic element with Rev response element-like properties that is essential for

expression of HBV gene products. *Molecular and Cellular Biology,* 13**,** 7476-7486.
HUANG, Z. M. & S., Y. T. 1995. Role of the hepatitis B virus posttran- scriptional regulatory element in export of intronless transcripts. *Molecular and Cellular Biology,* 15**,** 3864-3869.
HUTIN, S., LEE, Y. & GLAUNSINGER, B. A. 2013. An RNA Element in Human Interleukin 6 Confers Escape from Degradation by the Gammaherpesvirus SOX Protein. *Journal of Virology,* 87**,** 4672–4682.
HWANG, J. & MAQUAT, L. E. 2011. Nonsense-mediated mRNA decay (NMD) in animal embryogenesis: to die or not to die, that is the question. *Current Opinion in Genetics & Development,* 21**,** 422-430.
JACKSON, B. R., BOYNE, J. R., NOERENBERG, M., TAYLOR, A., HAUTBERGUE, G. M., WALSH, M. J., WHEAT, R., BLACKBOURN, D. J., WILSON, S. A. & WHITEHOUSE, A. 2011. An interaction between KSHV ORF57 and UIF provides mRNA-adaptor redundancy in herpesvirus intron-less mRNA export. *Plos Pathogens,* 7.
JACKSON, B. R., NOERENBERG, M. & WHITEHOUSE, A. 2012. The Kaposi's sarcoma-associated herpesvirus ORF57 protein and its multiple roles in mRNA biogenesis. *Frontiers in Microbiology,* 3**,** 1-9.
JAMESON, D. M. & SAWYER, W. H. 1995. Fluorescence anisotropy applied to biomolecular interactions. *Methods Enzymology,* 246**,** 283-300.
JERABEK-WILLEMSEN, M., WIENKEN, C. J., BRAUN, D., BAASKE, P. & DUHR, S. 2011. *Molecular Interaction Studies Using Microscale Thermophoresis*.
JINEK, M., COYLE, S. M. & DOUDNA, J. A. 2011. Coupled 5′ Nucleotide Recognition and Processivity in Xrn1-Mediated mRNA Decay. *Molecular Cell,* 41**,** 600-608.
JONES, K. D., AOKI, Y., CHANG, Y., MOORE, P. S., YARCHOAN, R. & TOSATO, G. 1999. Involvement of interleukin-10 (IL-10) and viral IL-6 in the spontaneous growth of Kaposi's sarcoma herpesvirus-associated infected primary effusion lymphoma cells. *Blood,* 94**,** 2871-2879.
KABSCH, W. 2009. XDS. *Acta Crystallographica Section D: Biological Crystallography,* D66**,** 125-132.
KAN, Z., ROUCHKA, E. C., GISH, W. R. & STATES, D. J. 2001. Gene structure prediction and alternative splicing analysis using genomically aligned ESTs. *Genome Research,* 11**,** 889-900.
KAPP, L. D. & LORSCH, J. R. 2004. The molecular mechanics of eukaryotic translation. *Annual Review of Biochemistry,* 73**,** 657-704.
KASHIMA, I., JONAS, S., JAYACHANDRAN, U., BUCHWALD, G., CONTI, E., LUPAS, A. N. & IZAURRALDE, E. 2010. SMG6 interacts with the exon junction complex via two conserved EJC- binding motifs (EBMs) required for nonsense-mediated mRNA decay. *Genes & Development,* 24**,** 2440-2550.
KHABAR, K. S. 2005. The AU-rich transcriptome: more than interferons and cytokines, and its role in disease. *Journal of Interferon & Cytokine Research,* 25**,** 1-10.
KONARSKA, M. M., PADGETT, R. A. & SHARP, P. A. 1984. Recognition of cap structure in splicing in vitro of mRNA precursors. *Cell,* 38**,** 731-736.
KRONSTAD, L. M. & GLAUNSINGER, B. A. 2012. Diverse virus–host interactions influence RNA-based regulation during g-herpesvirus infection. *Current Opinion in Microbiology,* 15**,** 506–511.

LAING, C. & SCHLICK, T. 2011. Computational approaches to RNA structure prediction, analysis, and design. *Current Opinion in Structural Biology,* 21**,** 306–318.
LAKOWICZ, J. R. 2006. *Principles of Fluorescence Spectroscopy,* New York, USA., Springer.
LAL, A., MAZAN-MAMCZARZ, K., KAWAI, T., YANG, X., MARTINDALE, J. L. & GOROSPE, M. 2004. Concurrent versus individual binding of HuR and AUF1 to common labile target mRNAs. *EMBO Journal,* 23**,** 3092-3102.
LALLENA, M. J., CHALMERS, K. J., LLAMAZARES, S., LAMOND, A. I. & VALCARCEL, J. 2002. Splicing regulation at the second catalytic step by Sex-lethal involves 3 splice site recognition by SPF45. *Cell,* 109**,** 285-296.
LASKOWSKI, R. A., MACARTHUR, M. W., MOSS, D. S. & THORNTON, J. M. 1993. PROCHECK: a program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography,* 26**,** 283-291.
LE HIR, H., IZAURRALDE, E., MAQUAT, L. E. & MOORE, M. J. 2000. The spliceosome deposits multiple proteins 20-24 nucleotides upstream of mRNA exon-exon junctions. *EMBO Journal,* 19**,** 6860-6869.
LEE, Y. J. & GLAUNSINGER, B. A. 2009. Aberrant Herpesvirus-Induced Polyadenylation Correlates With Cellular Messenger RNA Destruction. *PLoS ONE,* 7**,** 1-16.
LEJEUNE, F., ISHIGAKI, Y., LI, X. & MAQUAT, L. E. 2002. The exon junction complex is detected on CBP80-bound but not eIF4E-bound mRNA in mammalian cells: Dynamics of mRNP remodeling. *EMBO Journal,* 21**,** 3536-3545.
LEJEUNE, F. & MAQUAT, L. E. 2005. Mechanistic links between nonsense-mediated mRNA decay and pre-mRNA splicing in mammalian cells. *Current Opinion in Cell Biology,* 17**,** 309-315.
LEVY, J. A. 1997. Three new human herpesviruses (HHV6, 7, and 8). *Lancet,* 349**,** 558-563.
LICATALOSI, D. D. & DARNELL, R. B. 2010. RNA processing and its regulation: global insights into biological networks. *Nature Reviews Genetics,* 11**,** 75-87.
LIM, L. P. & BURGE, C. B. 2001. A computational analysis of sequence features involved in recognition of short introns. *Proceedings of the National Academy of Sciences,* 98**,** 11193-11198.
LIU, H., RODGERS, N. D., JIAO, X. & KILEDJIAN, M. 2002. The scavenger mRNA decapping enzyme DcpS is a 43. member of the HIT family of pyrophosphatases. *EMBO Journal,* 21**,** 4699-4708.
LIU, Z. R. 2002. p68 RNA helicase is an essential human splicing factor that acts at the U1 snRNA–5 splice site duplex. *Molecular and Cellular Biology,* 22**,** 5443-5450.
LOHMAN, T. M. 1986. Kinetics of protein-nucleic acid interactions: use of salt effects to probe mechanisms of interaction. *CRC Critical Reviews in Biochemistry,* 19**,** 191-245.
LONG, J. C. & CACERES, J. F. 2009. The SR protein family of splicing factors: master regulators of gene expression. *Biochemical Journal,* 317**,** 15-27.

LUO, M. J. & REED, R. 1999. Splicing is required for rapid and effi- cient mRNA export in metazoans. *Proceedings of the National Academy of Sciences,* 96**,** 14937-14942.
LYKKE-ANDERSEN, J. & WAGNER, E. 2005. Recruitment and activa- tion of mRNA decay enzymes by two ARE-mediated decay activation domains in the proteins TTP and BRF-1. *Genes & Development,* 19**,** 351-361.
MA, P. & XUHUA XIA, X. 2011. Factors Affecting Splicing Strength of Yeast Genes. *Comparative and Functional Genomics* 1-13.
MALIK, P., BLACKBOURN, D. J. & CLEMENTS, J. B. 2004. The evolutionarily conserved Kaposi's sarcoma- associated herpesvirus ORF57 protein interacts with REF protein and acts as an RNA export factor. *Journal of Biological Chemistry,* 279**,** 33001-33011.
MALIK, P. & SCHIRMER, E. C. 2006. The Kaposi's sarcoma-associated herpesvirus ORF57 protein: A pleurotropic regulator of gene expression. *Biochemical Society Transactions* 34**,** 705–710.
MANGUS, D. A., EVANS, M. C. & JACOBSON, A. 2003. Poly(A)-binding proteins: multifunctional scaffolds for the post- transcriptional control of gene expression. *Genome Biology,* 4**,** 223.
MAQUAT, L. E. 2004. Nonsense-Mediated mRNA Decay: Splicing, Translation and mRNP Dynamics. *Nature Reviews Molecular Cell Biology,* 5**,** 89-99.
MASUDA, S., DAS, R., CHENG, H., HURT, E., DORMAN, N. & REED, R. 2005. Recruitment of the human TREX complex to mRNA during splicing. *Genes & Development,* 19**,** 1512-1517.
MATHEWS, D. H., MOSS, W. N. & TURNER, D. H. 2010. Folding and finding RNA secondary structure. *Cold Spring Harbor Perspectives in Biology,* 2**,** a003665.
MATHEWS, D. H., SABINA, J., ZUKER, M. & TURNER, D. H. 1999. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *Journal of Molecular Biology,* 288**,** 911-940.
MATHY, N., BÉNARD, L., PELLEGRINI, O., DAOU, R., WEN, T. & CONDON, C. 2007. 5'-to-3' exoribonuclease activity in bacteria: role of RNase J1 in rRNA maturation and 59 stability of mRNA. *Cell,* 129**,** 681-692.
MCCOY, A. J., GROSSE-KUNSTLEVE, R. W., ADAMS, P. D., WINN, M. D., STORONI, C., L. & READ, R. J. 2007. Phaser crystallographic software. *Journal of Applied Crystallography,* 40**,** 658–674.
MCILWAIN, D. R., PAN, Q., REILLY, P. T., ELIA, A. J., MCCRACKEN, S., WAKEHAM, A. C., ITIE-YOUTEN, A., BLENCOWE, B. J. & MAK, T. W. 2010. Smg1 is required for embryogenesis and regulates diverse genes via alternative splicing coupled to nonsense-mediated mRNA decay. *Proceedings of the National Academy of Sciences,* 107**,** 12186-12191.
MILES, S. A., REZAI, A. R., SALAZAR-GONZÁLEZ, J. F., VANDER MEYDEN, M., STEVENS, R. H., LOGAN, D. M., MITSUYASU, R. T., TAGA, T., HIRANO, T., KISHIMOTO, T. & MARTINEZ-MAZA, O. 1990. AIDS Kaposi sarcoma-derived cells produce and respond to interleukin 6. *Proceedings of the National Academy of Sciences,* 87**,** 4068–4072.
MITCHELL, P., PETFALSKI, E., SHEVCHENKO, A., MANN, M. & TOLLERVEY, D. 1997. The exosome: a conserved eukaryotic RNA processing complex containing multiple $3 \rightarrow 5'$ exoribonucleases. *Cell,* 91**,** 457-466.

MOONGA, B. S., ADEBANJO, O. A., WANG, H. J., LI, S., WU, X. B., TROEN, B., INZERILLO, A., ABE, E., MINKIN, C., HUANG, C. L. & ZAIDI, M. 2002. Differential effects of interleukin-6 receptor activation on intracellular signaling and bone resorption by isolated rat osteoclasts. *Journal of Endocrinology,* 173**,** 395-405.

MOORE, M. J. 2005. From birth to death: the complex lives of eukaryotic mRNAs. *Science,* 309**,** 1514-1518.

MORIN, B., COUTARD, B., LELKE, M., FERRON, F., KERBER, R., JAMAL, S., FRANGEUL, A., BARONTI, C., CHARREL, R., DE LAMBALLERIE, X., VONRHEIN, C., LESCAR, J., BRICOGNE, G., GUNTHER, S. & CANARD, B. 2010. The N-terminal domain of the arenavirus L protein is an RNA endonuclease essential in mRNA transcription. *Plos Pathogens,* 6**,** e1001038.

NOTREDAME, C. 2013. *T-Coffee* [Online]. Available: http://www.ebi.ac.uk/Tools/msa/tcoffee/ [Accessed 25 September 2013].

NOTREDAME, C., HIGGINS, D. G. & HERINGA, J. 2000. T-coffee: a novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology,* 302**,** 205 -217.

NOTT, A., LE HIR, H. & MOORE, M. J. 2004. Splicing enhances translation in mammalian cells: An additional function of the exon junction complex. *Genes & Development,* 18**,** 210-222.

OKSENHENDLER, E., CARCELAIN, G., AOKI, Y., BOULANGER, E., MAILLARD, A., CLAUVEL, J. P. & AGBALIKA, F. 2000. High levels of human herpesvirus 8 viral load, human interleukin-6, interleukin-10, and C reactive protein correlate with exacerbation of multicentric castleman disease in HIV-infected patients. *Blood,* 96**,** 2069-2073.

OTWINOWSKY, Z. 1991. Maximum likelihood refinement of heavy atom parameters. In Isomorphous Replacement and Anomalous Scattering. *Proceedings of the CCP4 Study Weekend***,** 80-86.

PARISIEN, M. & MAJOR, F. 2008. The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature,* 452**,** 51-55.

PARKER, R. & SHETH, U. 2007. P Bodies and the Control of mRNA Translation and Degradation. *Molecular Cell,* 25**,** 635-646.

PARKER, R. & SONG, H. 2004. The enzymes and control of eukaryotic mRNA turnover. *Nature Structural & Molecular Biology* 11**,** 121-127.

PERRIN, M. F. 1926. Polarization de la luniere de fluorescence. Vie moyenne de molecules dans l'etat excite", . *Journal de Physique et Le Radium,* 7**,** 390-401.

PERTEA, M., MOUNT, S. M. & SALZBERG, S. L. 2007. A computational survey of candidate exonic splicing enhancer motifs in the model plant Arabidopsis thaliana. *BMC Bioinformatics,* 8.

PETTERSEN, E. F., GODDARD, T. D., HUANG, C. C., COUCH, G. S., GREENBLATT, D. M., MENG, E. C. & FERRIN, T. E. 2004. UCSF Chimera--a visualization system for exploratory research and analysis. *Journal of Computational Chemistry,* 25**,** 1605-1612.

PETTITT, J., MÜLLER, B., STANSFIELD, I. & CONNOLLY, B. 2008. Spliced leader trans-splicing in the nematode Trichinella spiralis uses highly polymorphic, noncanonical spliced leaders. *RNA Biology,* 14**,** 760–770.

PHILLIPS, W. C., STEWART, A., STANTON, M., NADAY, I. & INGERSOLL, C. 2002. High Sensitivity CCD based detector. *Journal of Synchrotron radiation,* 9**,** 36-43.
POLLARD, T. D. 2010. A Guide to Simple and Informative Binding Assays. *Molecular Biology of the Cell,* 21**,** 4061-4067.
PRUITT, K. D., TATUSOVA, T. & MAGLOTT, D. R. 2013. *RefSeq: NCBI Reference Sequence Database* [Online]. Available: http://www.ncbi.nlm.nih.gov/nuccore [Accessed 07 August 2013].
RAMACHANDRAN, G. N. & SASISKHARAN, V. 1968. Conformation of polypeptides and proteins. *Advances in Protein Chemistry,* 23**,** 283–437.
REED, R. 1996. Initial splice-site recognition and pairing during pre-mRNA splicing. *Current Opinion in Genetics & Development,* 6**,** 215-220.
REINECK, P., WIENKEN, C. J. & BRAUN, D. 2010. Thermophoresis of single stranded DNA. *Electrophoresis,* 31**,** 279–286.
RENNE, R., LAGUNOFF, M., ZHONG, W. & GANEM, D. 1996. The size and conformation of Kaposi's sarcoma-associated herpesvirus (human herpesvirus 8) DNA in infected cells and virions. *Journal of Virology,* 70**,** 8151-8154.
REZAEE, S. A., CUNNINGHAM, C., DAVISON, A. J. & BLACKBOURN, D. J. 2010. Human herpesvirus 8, complete genome, NC_009333.1, RefSeq.
REZAEE, S. A. R., CUNNINGHAM, C., DAVISON, A. J. & BLACKBOURN, D. J. 2006. Kaposi's sarcoma-associated herpesvirus immune modulation: an overview. *Journal of General Virology,* 87**,** 1781–1804.
RHODES, G. 2006. *Crystallography Made Crystal Clear* Oxford, Academic Press.
RICHNER, J. M., CLYDE, K., PEZDA, A. C., CHENG, B. Y., WANG, T., KUMAR, G. R., COVARRUBIAS, S., COSCOY, L. & GLAUNSINGER, B. 2011. Global mRNA degradation during lytic gammaherpesvirus infection contributes to establishment of viral latency. *PLoS Pathog,* 7**,** e1002150.
RUPP, B. 2010. *Biomolecular Crystallography: Principles, Practice, and Application to Structural Biology,* New York, Garland Science.
RUVOLO, V., WANG, E., BOYLE, S. & SWAMINATHAN, S. 1998. The Epstein–Barr virus nuclear protein SM is both a post-transcriptional inhibitor and activator of gene expression. *Proceedings of the National Academy of Sciences,* 95**,** 8852-8857.
SAGUEZ, C., OLESEN, J. R. & JENSEN, T. H. 2005. Formation of export-competent mRNP: Escaping nuclear destruction. *Current Opinion in Cell Biology,* 17**,** 287-293.
SAKHARKAR, M. K., CHOW, V. T. & KANGUEANE, P. 2004. Distributions of exons and introns in the human genome. *In Silico Biology,* 4**,** 387-393.
SALAHUDDIN, S. Z., ABLASHI, D. V., MARKHAM, P. D., JOSEPHS, S. F., STURZENEGGER, S., KAPLAN, M., HALLIGAN, G., BIBERFELD, P., WONG-STAAL, F., KRAMARSKY, B. & GALLO, R. C. 1986. Isolation of a new virus, HBLV, in patients with lymphoproliferative disorders. *Science,* 234**,** 596-601.
SANFORD, J. R., WANG, X., MORT, M., VANDUYN, N., COOPER, D. N., MOONEY, S. D., EDENBERG, H. J. & LIU, Y. 2009. Splicing factor SFRS1 recognizes a functionally diverse landscape of RNA transcripts. *Genome Research,* 19**,** 381-394.

SAULIÈRE, J., MURIGNEUX, V., WANG, Z., MARQUENET, E., BARBOSA, I., LE TONQUÈZE, O., AUDIC, Y., PAILLARD, L., ROEST CROLLIUS, H. & LE HIR, H. 2012. CLIP-seq of eIF4AIII reveals transcriptome-wide mapping of the human exon junction complex. *Nature Structural & Molecular Biology,* 19**,** 1124-1131.

SAULIERE, J., SUREAU, A., EXPERT-BEZANCON, A. & MARIE, J. 2006. The Polypyrimidine Tract Binding Protein (PTB) Represses Splicing of Exon 6B from the beta-Tropomyosin Pre-mRNA by Directly Interfering with the Binding of the U2AF65 Subunit. *Molecular and Cellular Biology,* 26**,** 8755–8769.

SCHOENBERG, D. R. 2011. Mechanisms of endonuclease- mediated mRNA decay *WIREs RNA,* 2**,** 582-600.

SCHUMANN, S., JACKSON, B. R., BAQUERO-PEREZ, B. & WHITEHOUSE, A. 2013. Kaposi's Sarcoma-Associated Herpesvirus ORF57 Protein: Exploiting All Stages of Viral mRNA Processing. *Viruses,* 5**,** 1901-1923.

SEELIGER, D. & DE GROOT B. L. 2009. tCONCOORD-GUI: visually supported conformational sampling of bioactive molecules. *Journal of Computational Chemistry,* 30, 1160-1166.

SHATKIN, A. J. & MANLEY, J. L. 2000. The ends of the affair: Capping and polyadenylation. *Nature Structural & Molecular Biology,* 7**,** 838-842.

SHI, X. & HERSCHLAG, D. 2009. Fluorescence Polarization Anisotropy to Measure RNA Dynamics. *Methods in Enzymology,* 469**,** 287-302.

SHIBUYA, T., TANGE, T. O., SONENBERG, N. & MOORE, M. J. 2004. eIF4AIII binds spliced mRNA in the exon junction complex and is essential for nonsense mediated decay. *Nature Structural & Molecular Biology,* 11**,** 346-351.

SHYU, A. B., BELASCO, J. G. & GREENBERG, M. E. 1991. Two distinct destabilizing elements in the c-fos message trigger deadenylation as a first step in rapid mRNA decay. *Genes & Development,* 5**,** 221-231.

SILVA, P. A. G. C., PEREIRA, C. F., DALEBOUT, T. J., SPAAN, W. J. M. & BREDENBEEK, P. J. 2010. An RNA Pseudoknot Is Required for Production of Yellow Fever Virus Subgenomic RNA by the Host Nuclease XRN1. *J. Virol.,* 84**,** 11395-11406.

SINGH, G., KUCUKURAL, A., CENIK, C., LESZYK, J. D., SHAFFER, S. A., WENG, Z. & MOORE, M. J. 2012. The Cellular EJC Interactome Reveals Higher Order mRNP Structure and an EJC-SR Protein Nexus. *Cell,* 151**,** 750-764.

SINGH, K. K., RÜCKER, T., HANNE, A., PARWARESCH, R. & KRUPP, G. 2000. Fluorescence polarization for monitoring ribozyme reactions in real time. *Biotechniques* 29**,** 344-348, 350-351.

SMITH, C. W. & VALCARCEL, J. 2000. Alternative pre-mRNA splicing: The logic of combinatorial control. *Trends In Biochemical Science,* 25**,** 381–388.

SOLLER, M. 2006. Pre-messenger RNA processing and its regulation: a genomic perspective. *Cellular and Molecular Life Sciences,* 63**,** 796-819.

STEWART, M. 2007. Ratcheting mRNA out of the nucleus. *Molecular Cell,* 25**,** 327-330.

STRELOW, L. I. & LEIB, D. A. 1995. Role of the virion host shutoff (vhs) of herpes simplex virus type 1 in latency and pathogenesis. *Journal of Virology,* 69**,** 6779-6786.

STRELOW, L. I. & LEIB, D. A. 1996. Analysis of conserved domains of UL41of herpes simplex virus type 1 in virion host shutoff and pathogenesis. *Journal of Virology,* 70**,** 5665-5667.
STRELOW, L. I., SMITH, T. & LEIB, D. A. 1997. The virion host shutoff function of herpes simplex virus type 1 plays a role in corneal invasion and functions independently of the cell cycle. *Virology,* 231**,** 28–34.
STUTZ, F. & IZAURRALDE, E. 2003. The interplay of nuclear mRNP assembly, mRNA surveillance and export. *Trends In Biochemical Science,* 13**,** 319–327.
TACKE, R. & MANLEY, J. L. 1995. The human splicing factor ASF/SF2 and SC35 possess different, functionally significant RNA binding specificities. *EMBO Journal,* 14**,** 3540-3551.
TANGE, T. O., NOTT, A. & MOORE, M. J. 2004. The ever-increasing complexities of the exon junction complex. *Current Opinion in Cell Biology,* 16**,** 279-284.
THARUN, S. & PARKER, R. 2001. Targeting an mRNA for decapping: displacement of translation factors and association of the Lsm1p–7p complex on deadenylated yeast mRNAs. *Molecular Cell,* 8**,** 1075-1083.
THOMAS, J., PALUSA, S. G., PRASAD, K. V. S. K., ALI, G. S., SURABHI, G.-K., BEN-HUR, A., ABDEL-GHANY, S. E. & REDDY, A. S. N. 2012. Identification of an intronic splicing regulatory element involved in auto-regulation of alternative splicing of SCL33 pre-mRNA. *The Plant Journal,* 72**,** 935–946.
TINOCO JR, I. & BUSTAMANTE, C. 1999. How RNA Folds. *Journal of Molecular Biology,* 293**,** 271-281.
TUNNICLIFFE, R. B., HAUTBERGUE, G. M., KALRA, P., JACKSON, B. R., WHITEHOUSE, A., WILSON, S. A. & GOLO- VANOV, A. P. 2010. Structural basis for the recognition of cellu- lar mRNA export factor REF by herpes viral proteins HSV-1 ICP27 and HVS ORF57. *Plos Pathogens,* 7**,** e1001244.
UETZ, P., DONG, Y.-A., ZERETZKE, C., ATZLER, C., BAIKER, A., BERGER, B., RAJAGOPALA, S. V., ROUPELIEVA, M., ROSE, D., FOSSUM, E. & HAAS, J. 2006. Herpesviral Protein Networks and Their Interaction with the Human Proteome. *Science,* 311**,** 239-242.
VAGIN, A. & TEPLYAKOV, A. 1997. MOLREP: an Automated Program for Molecular Replacement. *Journal of Applied Crystallography,* 30**,** 1022-1025.
VALENCIA, P., DIAS, A. P. & REED, R. 2008. Splicing promotes rapid and efficient mRNA export in mammalian cells. *Proceedings of the National Academy of Sciences***,** 3386-3391.
VINCIGUERRA, P. & STUTZ, F. 2004. mRNA export: An assembly line from genes to nuclear pores. *Current Opinion in Cell Biology,* 16**,** 285-292.
WALSH, D. & MOHR, I. 2011. Viral subversion of the host protein synthesis machinery. 9**,** 860-875.
WALTER, N. G., WOODSON, S. A. & BATEY, R. T. 2009. *Non-Protein Coding RNAs,* Heidelberg, Springer.
WASHIETL, S., HOFACKER, I. L. & STADLER, P. F. 2005. Fast and reliable prediction of noncoding RNAs. *PNAS,* 102**,** 2454–2459.
WEBER, G. 1953. Rotational Brownian motion and polarization of the fluorescence of solutions. *Advances in Protein Chemistry,* 8**,** 415-459.

WEBER, G. & HERCULES, D. M. 1966. Fluorescence and Phosphorescence Analysis. *Principles and Applications.* New York: Interscience Publishers (J. Wiley & Sons).
WHITEHOUSE, A., COOPER, M. & MEREDITH, D. M. 1998. The immediate- early gene product encoded by open reading frame 57 of herpesvirus saimiri modulates gene expression at a posttranscriptional level. *Journal of Virology,* 72**,** 857-861.
WICKENS, M., ANDERSON, P. & JACKSON, R. J. 1997. Life and death in the cytoplasm: messages from the 3′ end. *Current Opinion in Genetics & Development,* 7**,** 220-232.
WIENKEN, C. J., BAASKE, P., ROTHBAUER, U., BRAUN, D. & DUHR, S. 2010. Protein Binding Assays in Biological Liquids using Microscale Thermophoresis. *Nature Communications,* 1**,** 1-7.
WILLIAMS, G. 2001. *EMBOSS Compseq* [Online]. Hinxton, Cambridge. Available: http://emboss.bioinformatics.nl/cgi-bin/emboss/compseq [Accessed 07 August 2013].
WILSON, T. & TREISMAN, R. 1988. Removal of poly(A) and con- sequent degradation of c-fos mRNA facilitated by 30 AU-rich sequences. *Nature,* 336**,** 396-399.
WILUSZ, C. J., WORMINGTON, M. & PELTZ, S. W. 2001. The cap-to-tail guide to mRNA turnover. *Nature Reviews Molecular Cell Biology,* 2**,** 237-246.
WINN, M. D., BALLARD, C. C., COWTAN, K. D., DODSON, E. J., EMSLEY, P., EVANS, P. R., KEEGAN, R. M., KRISSINEL, E. B., LESLIE, A. G., MCCOY, A., MCNICHOLAS, S. J., MURSHUDOV, G. N., PANNU, N. S., POTTERTON, E. A., POWELL, H. R., READ, R. J., VAGIN, A. & WILSON, K. S. 2011. Overview of the CCP4 suite and current developments. *Acta Crystallographica Section D: Biological Crystallography,* 67**,** 235-242.
WORKMAN, C. & KROGH, A. 1999. No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution. *Nucleic Acids Research,* 27**,** 4816-4822.
YAMASHITA, A., CHANG, T. C., YAMASHITA, Y., ZHU, W., ZHONG, Z., CHEN, C. Y. & SHYU, A. B. 2005. Concerted action of poly(A) nucleases and decapping enzyme in mammalian mRNA turnover. *Nature Structural & Molecular Biology,* 12**,** 1054-1063.
YANG, F., PENG, Y. & SCHOENBERG, D. R. 2004. Endonuclease- mediated mRNA decay requires tyrosine phosphorylation of polysomal ribonuclease 1 (PMR1) for the targeting and degradation of polyribosome- bound substrate mRNA. *Journal of Biological Chemistry,* 279**,** 48993-49002.
YANG, F. & SCHOENBERG, D. R. 2004. Endonuclease-mediated mRNA decay involves the selective targeting of PMR1 to polyribosome-bound substrate mRNA. *Molecular Cell,* 4**,** 435-445.
YANG, X. C., SULLIVAN, K. D., MARZLUFF, W. F. & DOMINSKI, Z. 2009. Studies of the 59 exonuclease and endonuclease activities of CPSF-73 in histone pre-mRNA processing. *Molecular and Cellular Biology,* 29:**,** 31-42.
YOSHIDA, H., MATSUI, T., YAMAMOTO, A., OKADA, T. & MORI, K. 2001. XBP1 mRNA is induced by ATF6 and spliced by IRE1 in response to ER stress to produce a highly active transcription factor. *Cell,* 107**,** 881-891.
YUAN, P., BARTLAM, M., LOU, Z., CHEN, S., ZHOU, J., HE, X., LV, Z., GE, R., LI, X., DENG, T., FODOR, E., RAO, Z. & LIU, Y. 2009. Crystal structure of an avian

influenza polymerase PA(N) reveals an endonuclease active site. *Nature,* 458**,** 909-913.
ZAVOLAN, M., KONDO, S., SCHONBACH, C., ADACHI, J., HUME, D. A., HAYASHIZAKI, Y. & GAASTERLAND, T. 2003. Impact of Alternative Initiation, Splicing, and Termination on the Diversity of the mRNA Transcripts Encoded by the Mouse Transcriptome. 13**,** 1290-1300.
ZEMORA, G. & WALDSICH, C. 2010. RNA folding in living cells. *RNA Biology,* 7**,** 634-641.
ZHENG, D., EZZEDDINE, N., CHEN, C.-Y. A., ZHU, W., HE, X. & SHYU, A.-B. 2008. Deadenylation is prerequisite for P-body formation and mRNA decay in mammalian cells. *Journal of Cell Biology,* 182**,** 89-101.
ZHOU, Z., LUO, M. J., STRAESSER, K., KATAHIRA, J., HURT, E. & REED, R. 2000. The protein Aly links pre-messenger-RNA splicing to nuclear export in metazoans. *Nature,* 407**,** 401-405.
ZILLNER, K., JERABEK-WILLEMSEN, M., DUHR, S., BRAUN, D., LÄNGST, G. & BAASKE, P. 2011. Microscale Thermophoresis as a Sensitive Method to Quantify Protein: Nucleic Acid Interactions in Solution. *Methods in Molecular Biology.* TOTOWA, NJ: Springer Protocols.
ZUKER, M. 2003a. *mFold* [Online]. Available: http://mfold.rna.albany.edu/?q=mfold/RNA-Folding-Form [Accessed 25 Mai 2013].
ZUKER, M. 2003b. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res,* 31**,** 3406-15.
ZUKER, M. & STIEGLER, P. 1981. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research,* 9**,** 133-148.

Appendix:

A)

>GFP - Expression vector pSYNV-MReGFP-DsRed-P - JN377893.1 GI:371926914

ATGGTGAGCAAGGGCGAGGAGCTGTTCACCGGGGTGGTGCCCATCCTGGTCGAGCT
GGACGGCGACGTAAACGGCCACAAGTTCAGCGTGTCCGGCGAGGGCGAGGGCGATG
CCACCTACGGCAAGCTGACCCTGAAGTTCATCTGCACCACCGGCAAGCTGCCCGTG
*CCCTGGCCCACCCTCGTGACCACCCTGACCTAC*GGCGTGCAGTGCTTCAGCCGCTA
CCCCGACCACATGAAGCAGCACGACTTCTTCAAGTCCGCCATGCCCGAAGGCTACG
TCCAGGAGCGCACCATCTTCTTCAAGGACGACGGCAACTACAAGACCCGCGCCGAG
GTGAAGTTCGAGGGCGACACCCTGGTGAACCGCATCGAGCTGAAGGGCATCGACTT
CAAGGAGGACGGCAACATCCTGGGGCACAAGCTGGAGTACAACTACAACAGCCACA
ACGTCTATATCATGGCCGACAAGCAGAAGAACGGCATCAAGGTGAACTTCAAGATC
CGCCACAACATCGAGGACGGCAGCGTGCAGCTCGCCGACCACTACCAGCAGAACAC
CCCCATCGGCGACGGCCCCGTGCYGCTGCCCGACAACCACTACCTGAGCACCCAGT
CCGCCCTGAGCAAAGACCCCAACGAGAAGCGCGATCACATGGTCCTGCTGGAGTTC
GTGACCGCCGCCGGGATCACTCTCGGCATGGACGAGCTGTACAAGTAATAAACTAC
AGCCACAACTCTACCTCCCCCACTATGAATAAACGACCTAACATATAATATAAGAA
AAACCAACAGAAATCATAATATTTTATTTGTCTGTTTGTATTATTTGTCTAG

>DsRed2 [Cloning vector pSAT6A-DsRed2-N1] AAY25372
DQ005468.1

ATGGCCTCCTCCGAGAACGTCATCACCGAGTTCATGCGCTTCAAGGTGCGCATGGA
GGGCACCGT
GAACGGCCACGAGTTCGAGATCGAGGGCGAGGGCGAGGGCCGCCCCTACGAGGGCC
ACAACACCGTGAAGCTGAAGGTGACCAAGGGCGGCCCCCTGCCCTTCGCCTGGGAC
ATCCTGTCCCCCCAGTTCCAGTACGGCTCCAAGGTGTACGTGAAGCACCCCGCCGA
CATCCCCGACTACAAGAAGCTGTCCTTCCCCGAGGGCTTCAAGTGGGAGCGCGTGA
TGAACTTCGAGGACGGCGGCGTGGCGACCGTGACCCAGGACTCCTCCCTGCAGGAC
GGCTGCTTCATCTACAAGGTGAAGTTCATCGGCGTGAACTTCCCCTCCGACGGCCC
CGTGATGCAGAAGAAGACCATGGGCTGGGAGGCCTCCACCGAGCGCCTGTACCCCC
GCGACGGCGTGCTGAAGGGCGAGACCCACAAGGCCCTGAAGCTGAAGGACGGCGGC
CACTACCTGGTGGAGTTCAAGTCCATCTACATGGCCAAGAAGCCCGTGCAGCTGCC
CGGCTACTACTACGTGGACGCCAAGCTGGACATCACCTCCCACAACGAGGACTACA
CCATCGTGGAGCAGTACGAGCGCACCGAGGGCCGCCACCACCTGTTCCTGCTGA

>Homo sapiens hemoglobin, beta (HBB) 626 bp mRNA
NM_000518.4 GI:28302128

ACATTTGCTTCTGACACAACTGTGTTCACTAGCAACCTCAAACAGACACCATGGTG
CATCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGT
GGATGAAGTTGGTGGTGAGGCCCTGGGCAGGCTGCTGGTGGTCTACCCTTGGACCC
AGAGGTTCTTTGAGTCCTTTGGGGATCTGTCCACTCCTGATGCTGTTATGGGCAAC
CCTAAGGTGAAGGCTCATGGCAAGAAAGTGCTCGGTGCCTTTAGTGATGGCCTGGC
TCACCTGGACAACCTCAAGGGCACCTTTGCCACACTGAGTGAGCTGCACTGTGACA

AGCTGCACGTGGATCCTGAGAACTTCAGGCTCCTGGGCAACGTGCTGGTCTGTGTG
CTGGCCCATCACTTTGGCAAAGAATTCACCCCACCAGTGCAGGCTGCCTATCAGAA
AGTGGTGGCTGGTGTGGCTAATGCCCTGGCCCACAAGTATCACTAAGCTCGCTTTC
TTGCTGTCCAATTTCTATTAAAGGTTCCTTTGTTCCCTAAGTCCAACTACTAAACT
GGGGGATATTATGAAGGGCCTTGAGCATCTGGATTCTGCCTAATAAAAAACATTTA
TTTTCATTGC

B)

pETM6T1 Vector

> KSHV ORF37 aka SOX (Codon optimised for E.coli / sent by Bahram Ibrahimi)
> Nucleotide sequence 1461nt G+C=51%
> Contains EcoRI (Ala/GCA-Ser/AGC) at 1178
> EcoRV (Pro/CCA-Trp/TGG) at 1064
> HindIII (xxx/XXX-xxx/XXX) at 105, 387, 935
> NdeI (Gln/CAG-Asp/GAT) at 918
> NheI (xxx/XXX-xxx/XXX) at 97
> Free of AseI/ BamHI/ BspHI/ NcoI/ XbaI/ XhoI

pETM6T1 vector Restriction Map

```
                                  T7 promoter       lac operator
241
TCGAGATCTCGATCCCGCGAAATTAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTAGAAATAA
320
241
AGCTCTAGAGCTAGGGCGCTTTAATTATGCTGAGTGATATCCCCTTAACACTCGCCTATTGTTAAGGGGAGATCTTTATT
320
           BglII              VspI                          BsrBI            XbaI


                                                        His-tag#1
                        RBS          M  G  S  S  H  H  H  H  H  H  S  S  M  N  K
E
321
TTTTGTTTAACTTTAAGAAGGAGATATACAT**ATG**GGCAGCAGCCATCACCACCACCACCATTCTAGTATGAACAAAGAAA
400
321
AAAACAAATTGAAATTCTTCCTCTATATGTATACCCGTCGTCGGTAGTGGTGGTGGTGGTAAGATCATACTTGTTTCTTT
400
                                    NdeI

              NusA                                 His-tag#2
       I  L  A    F  G  D  E  A  T  S  G  S  G  H  H  H  H  H  H  D  Y  D  I  P  T
T  E
401     TTTTGGCT
TTCGGTGACGAAGCGACTAGTGGTTCTGGTCATCATCATCATCATCACGATTACGATATCCCAACGACCGAA   480
401     AAAACCGA
AAGCCACTGCTTCGCTGATCACCAAGACCAGTAGTAGTAGTAGTAGTGCTAATGCTATAGGGTTGCTGGCTT   480
                                SpeI   DrdII                                EcoRV


          TEV site
        N  L  Y  F  Q* G  A  M  G  S
481
AACTTGTATTTCCAGGGCGCC**ATG**GGATCCGAATTCTGTACAGGCGCGCTTGCAGGACGTCGACGGTACCATCGATACGC
560
481
TTGAACATAAAGGTCCCGCGGTACCCTAGGCTTAAGACATGTCCGCGCGAACGTCCTGCAGCTGCCATGGTAGCTATGCG
560
```

```
                                    BamHI        BsrGI   BssHII       SalI
KpnI  MluI
                              NcoI       EcoRI                        AatII
ClaI


561
GTTCGAAGCTTGCGGCCGCACAGCTGTATACACGTGCAAGCCAGCCAGAACTCGTCCTGAAGACCCAGAGGATCTCGAGC
640
561
CAAGCTTCGAACGCCGGCGTGTCGACATATGTGCACGTTCGGTCGGTCTTGAGCAGGACTTCTGGGTCTCCTAGAGCTCG
640
           NspV       EagI       PvuII      PmlI
XhoI
             HindIII               BstZ17I


       His-tag#3 (optional)
641     ACCACCACCACCACCACTAATGTTAATTAAGTTGGGCGTTCCTAGGCTGATAAAA   695
641     TGGTGGTGGTGGTGGTGATTACAATTAATTCAACCCGCAAGGATCCGACTATTTT   695
                                              AvrII
```

Note: you can not use NdeI I site for clonning.

Vector is Km resistant.

NusA is not present in the map but it is located between His-tag#1 and His-tag#2.

The NusA amino acid missing between A and F are the following:

VVEAVSNEKALPREKIFEALESALATATKKKYEQEIDVRVQIDRKSGDFDTFRRWL
VVDEVTQPTKEITLEAARYEDESLNLGDYVEDQIESVTFDRITTQTAKQVIVQKVR
EAERAMVVDQFREHEGEIITGVVKKVNRDNISLDLGNNAEAVILREDMLPRENFRP
GDRVRGVLYSVRPEARGAQLFVTRSKPEMLIELFRIEVPEIGEEVIEIKAAARDPG
SRAKIAVKTNDKRIDPVGACVGMRGARVQAVSTELGGERIDIVLWDDNPAQFVINA
MAPADVASIVVDEDKHTMDIAVEAGNLAQAIGRNGQNVRLASQLSGWELNVMTVDD
LQAKHQAEAHAAIDTFTKYLDIDEDFATVLVEEGFSTLEELAYVPMKELLEIEGLD
EPTVEALRERAKNALATIAQAQEESLGDNKPADDLLNLEGVDRDLAFKLAARGVCT
LEDLAEQGIDDLADIEGLTDEKAGALIMAARNICW

C)

- SOX 244, Xrn1 & RNA 5 mg/mL

SOX 244, Xrn1& 51GFP	SOX 244, Xrn1 & 58HBB	SOX 244, Xrn1 & 51GFP
JCSG-PLUSTM HT-96 MD1-40		
D1	A2	A2(-A4A7-A10A12B2-B5)
D5	A3	A7
E6	A4	A10
E10	A7-A10 A12 B2-B4	B2
E11	B12	B12
E12	C1	C12
F3	C12	D1
F6	E6	D10

G1	E10	D12
G3	E11	E3
G9	E12	E5
H3	F3	E6
	F6	G1
	F7	G6
	G1	G8
	G3	G9-G10 H3-H6 H8
	G4-G7	H10
	G9-G10 H3	
	H8	
ProPlex HT-96 MD1-42		
A3-A6 A9-A10		
A11		
B2		
B9		
B11		
C3		A1
C5		A3-A6 A9-A10
C12 D1-D3	A1-A6 A9-A11	D6-D7 D9
D6	B4B6-B7 B9-	D11
	B11C2C3C5C7	
D12	D6-D7 D9-D12	D12

- SOX 244 & RNA 8 mg/mL

SOX 244 & 51GFP	SOX 244 & 58HBB	SOX 244 & 51GFP	SOX 244 & 58HBB-FAM	SOX WT & 23GFP
JCSG-PLUSTM HT-96 MD1-40				
G6	A2	A2	D12	A4
G7	A3	A8	F3	A5
G8	A7	A9	G3	A7
	A8	B11	G6	A8
	A9	C1	G8	A10
	A10	C3		A12
	A12	C4		B2
	B2	C8		B4
	B12	C12		B12
	C3	E10		C9
	C8	E12		D6
	C9	G1		E8
	C10	G6		E9
	CE3	G7		E10
	E10	H6		G6

	F6	H8		G7
	G6			H8
	G7			H11
	G8			H12
	H3			
	H8			
	H10			
ProPlex HT-96 MD1-42				
A6	A6	A1	D12	A3
	A9	A3		A6 A7
	A10	A5		A9 A10
	B3	A6		B3 –B6
	B4	A9		B10 B11-B12
	C5	A10		C1
	D6	B4		C4
	D11	B7		C6-C7
	D12	C5		D3-D7
		C8		D10
		D3		E4-E5
		D6		E7- E9 E10
		D7 D9 D10		E12
				F2-F3
				F5-F6
				G9
				H6
				H10

D)

The SOX proteins, wild type and mutants, used in the biochemical and biophysical experiments were all recombinantly produced and purified to a high level of purity. From figure D-1 A below, it can be seen that the lysate supernatant contained a variety of proteins of different molecular weights. The majority of these did not bind to the HisTrap column. The major elution product from this column was of the same molecular weight as the NusA-SOX construct (114 kDa), which predominantly eluted between fractions 1 to 11, however significant impurities were nonetheless evident. These fractions were therefore pooled and loaded onto a Q-sepharose column to eliminate most small molecular weight impurities (Figure D-1 B). Some degradation products of NusA and/or SOX, however, were evident. After overnight dialysis and cleavage with TEV protease, a second anion exchange step was undertaken to separate NusA (60 kDa) from SOX

(55 kDa), which are close in molecular weight (Figure D-1 C). The elution product from this column that was of the correct molecular weight for SOX was collected and loaded onto a gel filtration column to finally eliminate smaller molecular weight impurities (Figure D-1 D).

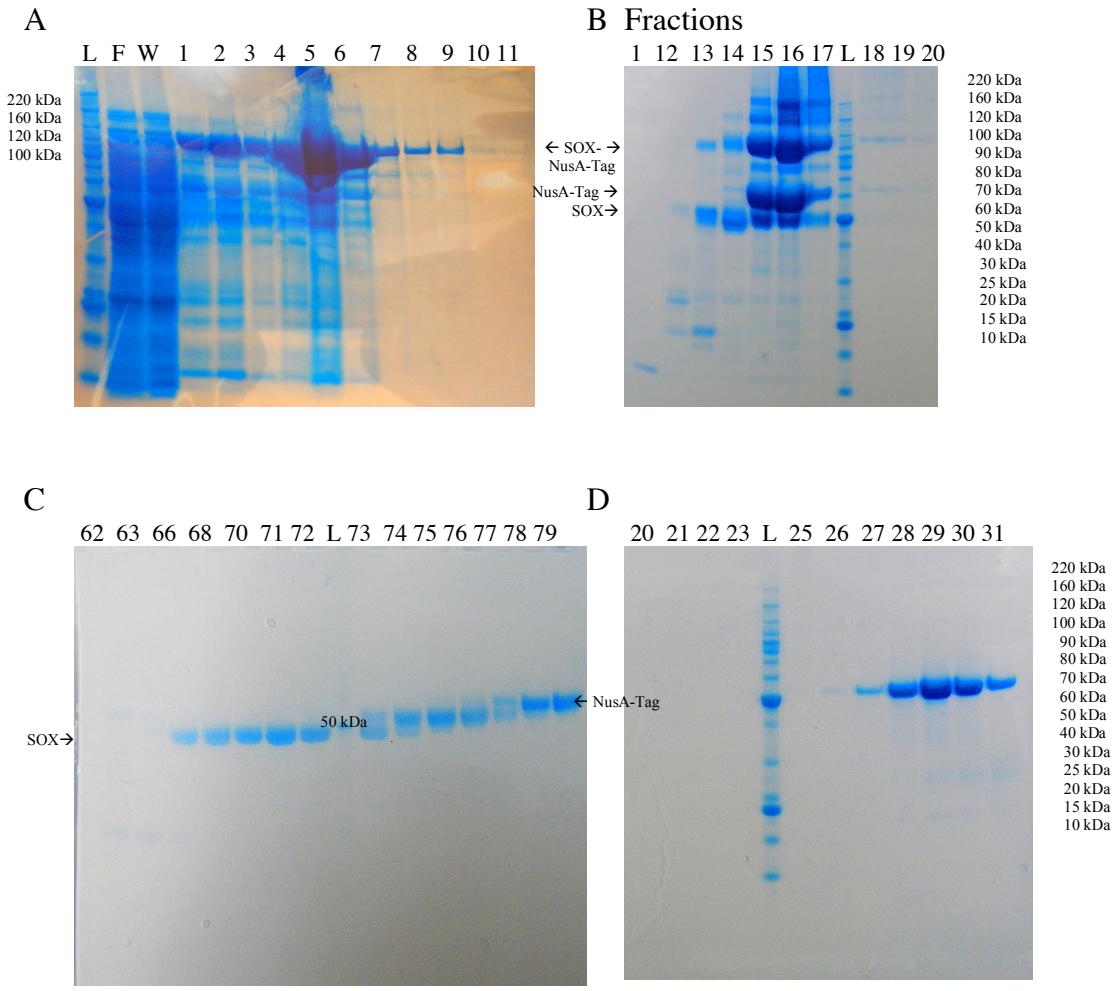


Figure D-1: Purification of Recombinant SOX

A) The SDS-PAGE gel of the HisTrap step has the major band of the NusA-SOX construct (114 kDa) in fractions 1 to 11. B) The SDS-PAGE gel of the 1st Q-sepharose step shows the major bands of the NusA-SOX construct (114 kDa) and likely degradation products in fractions 15-17 . C) The SDS-PAGE gel of the 2nd Q-sepharose step shows the two major bands for SOX (55 kDa) and NusA (60 kDa) separating in fractions 62-79. D) SDS-PAGE gel of the gel filtration step, where the bands in fractions 25-32 are SOX (55 kDa). L = Molecular weight ladder; F = Flow through; W = Wash.

E)

The Xrn1 protein used in the biochemical and biophysical experiments were all recombinant produced and purified. From figure E-1 A it can be seen that the lysate supernatant contained a variety of proteins of different molecular weights. The majority of these did not bind to the HisTrap column. From the HisTrap step, high overexpression of Xrn1 was not obvious, but a band spreading from fraction 27-37 was consistent with the molecular weight of the Xrn1 construct (144 kDa) (Figure E-1 A). These fractions were therefore pooled and loaded onto a gel filtration column to eliminate other molecular weight impurities (Figure E-1 B). Following this step, a clear band for Xrn1 was visible, though many impurities were still present. These were eliminated in the SOX-Xrn1 pull down experiment (See in section 5.1.1, Figure 5-1).

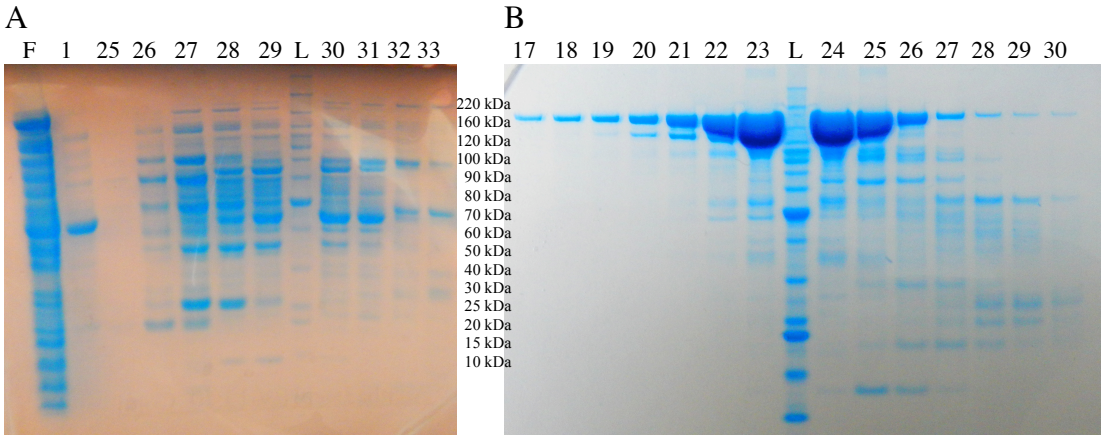


Figure E-1: Purification of Recombinant Xrn1 (SDS-Page)

A) SDS-PAGE gel of the HisTrap step, where the Xrn1 (144 kDa) band is seen in fraction 27-37, along with other impurities. B) In the SDS-PAGE gel of the gel filtration step Xrn1 (144 kDa) and other impurities can be seen in fractions 17-30 contain. L = Molecular weight ladder; F = Flow through.

F)

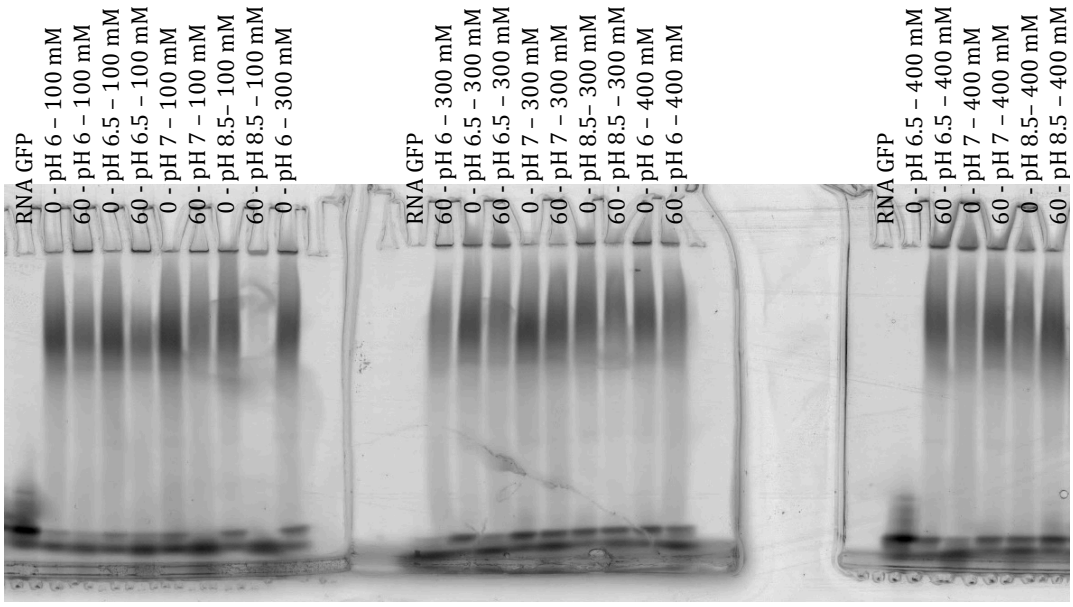


Figure F-1: Optimal Binding Conditions for SOX and 51 Nucleotides GFP RNA

The salt concentration was varied from 100 to 400 mM NaCl in 100mM increments and the pH was varied from pH 6-8.5 in pH units of 0.5. The binding experiments were run 0 minutes of incubation at 37 °C and 60 minutes of incubation at 37 °C. The most discreet band was observed for 100 mM NaCl and a pH of 7; n=3.

Additional Appendix Chapter 7

In Silico Assessment of Potential Druggable Pockets on the Surface of α_1-Antitrypsin Conformers

Anathe Olivia Maria
Patschull Lafitte-Laplace

Thesis submitted for the degree of Doctor of Philosophy

Institute of Structural and Molecular Biology

Department of Biological Sciences
Birkbeck College
University of London

October 2013

Declaration

I, Anathe Olivia Maria Patschull Lafitte-Laplace, hereby declare that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated. I declare that the method "Provar", described in section e' in the methods (2') and the results (3 f') section represented in the figure 3-8' were obtained by Dr. Paul Ashford in collaborative work. Lakshmi Segu set up the crystallization of the crystal that lead to the structure 3ne4.

Anathe Olivia Maria Patschull Lafitte-Laplace
October 2013

Abstract

The search for druggable pockets on the surface of a protein is often performed on a single conformer, treated as a rigid body. Transient druggable pockets may be missed in this approach. A systematic *in silico* analysis of surface clefts across multiple conformers of the metastable protein α_1-antitrypsin (A1AT) was performed. Pathological mutations disturb the conformational landscape of A1AT, triggering polymerisation that leads to emphysema and hepatic cirrhosis. Computational screens for small molecule inhibitors of polymerisation have generally focused on one major druggable site visible in all crystal structures of native A1AT. Solving the highest resolution structure of native A1AT to date allowed highly detailed comparison of this cavity's flexibility demonstrated by 3 structures at ≤2.0 Å resolution. An alternative approach scanned all surface clefts observed in multiple crystal structures of A1AT and in 100 computationally produced conformers, mimicking the native solution ensemble. The persistence, variability and druggability of these pockets were assessed. Molecular docking of publicly available libraries of small molecules was then used to explore scaffold preferences for each site. This approach identified a number of novel target sites for drug design. In particular one transient site showed favourable characteristics for druggability due to high enclosure and hydrophobicity. Hits against this and other druggable sites achieved docking scores corresponding to a K_d in the μM-nM range, comparing favourably with a recently identified promising lead. Preliminary ThermoFluor studies support the docking predictions. In conclusion, this strategy shows considerable promise compared with the conventional single pocket/single conformer approach to *in silico* screening. The best-scoring ligands warrant further experimental investigation.

Contents

Abbreviations

#		
	3D	**Three-Dimensional**
A		
	A1AT	**α_1-Antitrypsin**
C		
	cDNA	**Complementary DNA**
	CD	**Circular Dichroism**
D		
	Da	**Dalton**
	DMSO	**Dimethyl Sulfoxide**
I		
	IFD	**Induced Fit Docking**
M		
	MMT	**Dl-Malic Acid:MES:Tris Base**
P		
	PDB	**Protein Data Bank**
K		
	K	**Kelvin**
	K_d	**Dissociation Constant**
P		
	PAGE	**Polyacrylamide Gel Electrophoresis**
	PEG	**Polyethylene Glycol**
S		
	S	**Stressed**
	SDS	**Sodium Dodecyl Sulphate**
R		
	R	**Relaxed**
	RCL	**Reactive Centre Loop**
	RMSD	**Root-Mean-Square Deviation**

List of Figures

List of Tables

Acknowledgments

I would like to thank Dr. Irilenia Nobeli and Dr. Bibek Gooptu for their guidance and good work atmosphere, Dr. Claire Naylor, Dr. Tracey Barrett and Dr. Nora Cronin for their assistance with the structural refinement, Dr. Louise Briggs for her technical assistance with the ThermoFluor studies, and Dr. Mun Peak Nyon for her assistance and patience teaching protein purification and crystallisation.

1'. **Introduction**

The desire to modulate protein function with small molecules that can be administered as drugs has led to a plethora of studies attempting to define and calculate the "druggability" of sites on a protein (An et al., 2005, Halgren, 2009, Hajduk et al., 2005, Huang and Jacobson, 2010, Sheridan et al., 2010). Most studies have relied on experience from inhibiting enzymes acting on small molecule substrates. Here the target sites are well-formed surface pockets, characterized by high curvature and low solvent accessibility. Recently "harder" targets have been addressed. These include protein-protein interactions and proteins belonging to large homologous superfamilies e.g. kinases. In the former, the interfaces are larger and flatter (Wells and McClendon, 2007). In the latter, inhibiting the common active site risks serious cross-class side effects. Both these issues may be addressed by targeting clefts that are not necessarily associated directly with the protein's biochemical function. The idea is that binding of small molecules to such clefts may be more favourable and could still allosterically modulate protein function, e.g. via preferential stabilization of a particular state within the conformational landscape of the protein in solution.

The search for suitable allosteric clefts requires consideration of functional relevance and druggability. Functional relevance is usually less obvious from structural snapshots for an allosteric site than an active site. It may be deduced experimentally by mutagenesis, or through observation of the binding site of known ligands. Druggability has traditionally been indirectly assessed by computational studies (docking) or *in vitro* screening. More recently, quantitative predictors of cleft druggability have been devised (Halgren, 2009, Sheridan et al., 2010, Nayal and Honig, 2006, Cheng et al., 2007, Schmidtke and Barril, 2010, Hajduk et al., 2005). These commonly assess the size, shape, buriedness and hydrophobic character of a site. However, a major limitation is currently not addressed routinely: the transient character of some clefts that may otherwise be of interest in drug design. Druggable pockets on a protein's surface are most commonly assessed using a single 3D structure. This is unsatisfactory because proteins undergo dynamic changes in solution, sampling multiple conformations, each with potentially different surface pockets. The existence of multiple conformers is

especially relevant to ligand recognition. Ligand binding inherently tends to conformational selection (Ma et al., 2002, Kumar et al., 2000), a process by which protein-ligand interactions lower the free energy of a conformer, increasing the stability and population of a state that may otherwise rarely be observed. In recent years some notable efforts have been made to identify transient sites. In the approach pioneered by Eyrisch and Helms (Eyrisch and Helms, 2007), trajectory snapshots from molecular dynamics simulations revealed transient pockets on the surfaces involved in protein-protein interactions. Transient pockets can be revealed using more computationally efficient methods, albeit usually at the cost of reduced pocket diversity (Eyrisch and Helms, 2009). Moreover pocket tracking across multiple structures with the program fpocket can highlight important changes to a pocket, arising from both dynamics of a single protein as well as evolutionary time in a family of homologues (Schmidtke et al., 2010). Recently, the importance of employing multiple protein conformers in virtual screening has been highlighted (Bottegoni et al., 2011, Nichols et al., 2012, Ivetac and McCammon, 2012) and there is a growing trend for incorporating notions of flexibility in the docking process. Despite these pioneering studies, most current *in silico* screening starts with the selection of a single pocket from a single conformer. Two strategies can build upon crystallographic data to improve this situation. Firstly the solution of multiple high-resolution crystal structures in different conditions allows interpretation of subtle variations in conformation around druggable pockets. Alternatively multiple plausible conformer variants may be generated *in silico* to mimic the native conformational ensemble in solution. The second approach is necessarily less rooted in experimental data, but on the other hand it is more likely to identify potential, transiently populated druggable pockets missed in single conformer approaches.

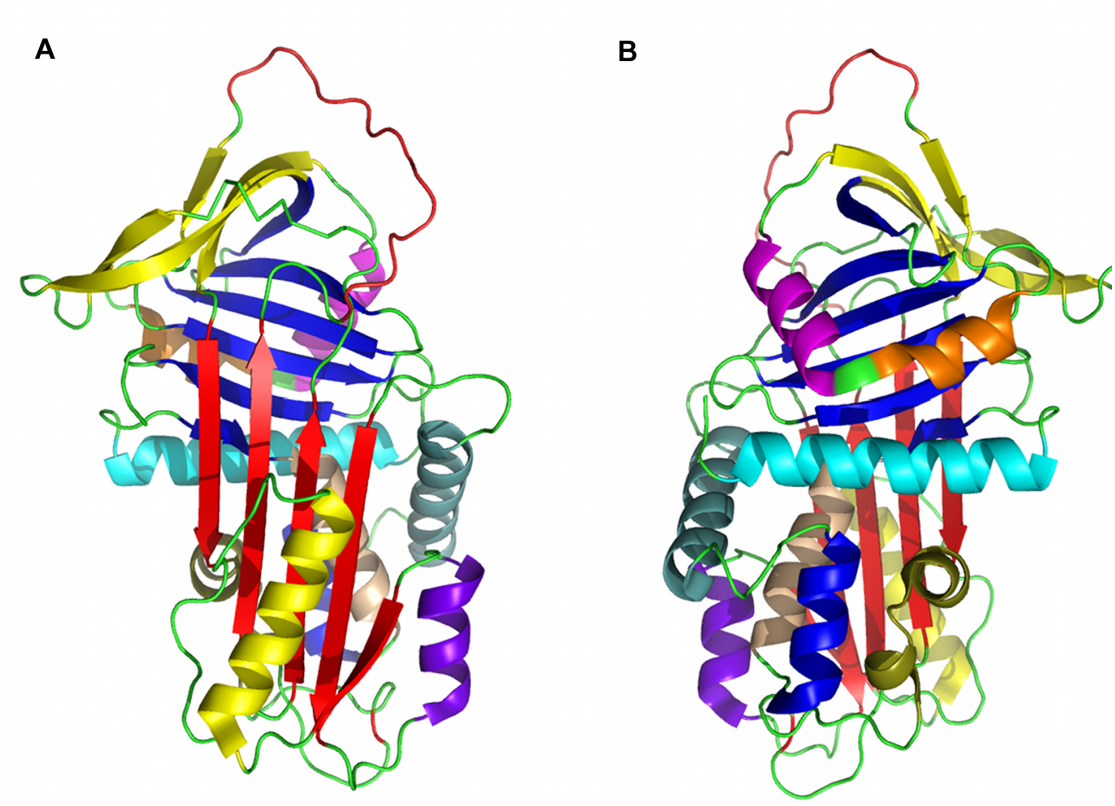


Figure 1-1': The Structure of the Wild Type α_1-Antitrypsin

Front (A) and back (B) views of the structure of A1AT in cartoon representation (PDB entry: 1qlp). The secondary elements are coloured as follows. ß-sheets: A (red), B (blue), and C (yellow); helices: A (cyan), B (apricot), C (blue), D (grey-green), E (purple), F (yellow), G (orange), H (pink), I (olive); loops: reactive centre loop (RCL, red), all other loops (green).

Both approaches were used to study druggable sites in α_1-antitrypsin (A1AT), the archetypal member of the serpin (serine protease inhibitor) superfamily (Silverman et al., 2001). Its characteristic native fold (Figure 1-1') is metastable and this is key to its antiprotease function (Huntington et al., 2000). It is an excellent candidate system in which to assess these strategies. Firstly, A1AT is a medically important target. Its metastability is subverted by pathogenic mutations that cause A1AT to polymerise. This causes diseases of the liver (neonatal hepatitis, cirrhosis and hepatocellular carcinoma) and lung (early-onset emphysema) through loss- and gain-of-function mechanisms (Gooptu and Lomas, 2008). Secondly, the biological function and dysfunction of serpins is coupled to marked conformational changes involving large rearrangements of their structure (Gooptu and Lomas, 2009).

Moreover, extensive mutagenesis experiments demonstrate that mutations around surface clefts can significantly alter the stability of native A1AT (Seo et al., 2000). Metastability is therefore related to pocket vacancy, indicating that ligand binding in a range of allosteric sites may modulate stability, and hence, pathological conformational change. Lastly, a range of high-resolution crystallographic datasets are available for wild type and mutant A1AT species in native, metastable (or stressed 'S') and relaxed ('R'), hyperstable states, allowing comparison of computationally derived conformers with structural data. Following docking studies, the most promising findings have been assessed experimentally, identifying small molecule ligands with potential for development as novel therapeutics.

2'. Materials and Methods

a'. α_1-Antitrypsin Purification, Crystallography and Assessment

pQE31 plasmid containing cDNA encoding hexahistidine-tagged recombinant wild-type A1AT was transfected into XL1 Blue *Escherichia coli* cells (Stratagene). The proteins were expressed and purified as described previously (Parfrey et al., 2003). They were characterized using SDS–PAGE, nondenaturing PAGE and transverse urea gradient (TUG) PAGE, circular-dichroism (CD) spectroscopy and enzyme-inhibitory activity and kinetics assays (Stone and Hofsteenge, 1986, Dafforn et al., 2004).

Crystals of A1AT were grown in 0.1 M MMT buffer (1:2:2 dl-malic acid:MES:Tris base) pH 6.0, 20%(w/v) PEG 1500, 330 mM N-{4-hydroxy-3-methyl-5-[(1H-1,2,4,5-tetrazol-3-yl)sulfanyl]phenyl}-4-methylbenzenesulfonamide by hanging-drop vapour diffusion at 293 K. These crystals were then loop-mounted and cryocooled in cryoprotectant buffer [0.1 M MMT pH 6.0, 20%(w/v) PEG 1500, 20%(v/v) glycerol]. Synchrotron diffraction data were collected on beamline 23.1 at the ESRF, Grenoble, France. Processing of the X-ray diffraction data was performed using *iMOSFLM* (Battye et al., 2011) and *SCALA* (Evans, 2006). The structure of A1AT was solved by molecular replacement with *Phaser* (McCoy et al., 2007) to a resolution of 1.8 Å using the coordinates of the native A1AT crystal

structure (PDB entry 1qlp; (Elliott et al., 2000)) as a search model. An initial model was constructed using *Coot* (Emsley and Cowtan, 2004) and the structure was refined using *REFMAC5* (Murshudov et al., 1997). Iterative cycles of model building and refinement were carried out until the R factors stabilized. The stereochemistry of the final model (PDB entry 3ne4; (Patschull et al., 2011)) was checked using *PROCHECK* (Laskowski et al., 1994).

The cavity flanking α-sheet A in the new structure was assessed and compared with those observed in the two structures of nearest resolution (PDB entries 1qlp and 2qug; (Pearce et al., 2008)) using the program *SiteMap* 2.5 and other programs from the Schrödinger suite (Schrödinger LLP, New York, USA). The crystal structures were prepared using the Protein Preparation Wizard protocol in the *Maestro* program. Ligands, waters and other cocrystallized agents were deleted and H atoms were added. The protassign script was used to optimize intramolecular contacts. The impref script was used to perform restrained minimization of the protein (default settings in *Maestro* v.9.2). All structures were superposed using the structalign utility from Schrödinger. A site was defined as an enclosed region comprising at least 15 site points (default settings in *SiteMap* v.2.5). *SiteMap* uses an algorithm to identify and characterize favourable sites in a protein structure for drug binding. Probe-based and energy based methods are used to estimate the interaction energy between probe and protein along a three-dimensional grid that samples the space around the structural model. These values are combined with geometry terms to give a druggability scoring function that is a function of volume and site enclosure (solvent exclusion). A penalty factor is calculated for hydrophilicity. Other parameters that are calculated for each site are volume, solvent exposure, contacts, hydrophobicity and hydrogen-donor/acceptor sites.

b'. Selection of α_1-Antitrypsin Crystal Structures

A1AT structures were retrieved from the PDB using the SAS tool available in PDBsum (Laskowski, 2009). The amino acid sequence of the structure with PDB id 1qlp was used to search the PDB, and all sequences with percentage sequence

identity higher than 97% were kept (this very high cut-off was used to retain only A1AT structures). Among identical sequences representing identical states, the highest resolution available was kept. Structures with cleaved chains, where the break in the chain was not in the RCL were removed. The final dataset (summarised in Table 2-1') comprised structures that sampled different features, such as the stressed and relaxed forms, point mutations, and ligands that induce stability. More specifically, there are six native stressed and two relaxed A1AT structures, all with resolution better than or equal to 2.6 Å. The six native stressed structures can be separated into two groups. The first group comprises PDB entries 1qlp (Elliott et al., 2000), 2qug (Pearce et al., 2008)and 3cwm (Pearce et al., 2008), which have no mutations and share nearly 100% sequence similarity (except for minor variations in the length of the C- and/or N-terminus). A partially stabilising ligand, citrate, is present in 3cwm. The second group comprises 1hp7 (Kim et al., 2001), 1oph (Dementiev et al., 2003) and 3drm (Gooptu et al., 2009), all representing the native stressed fold but with partially stabilising mutations in the sequence. Finally, of the two relaxed structures, one is an uncleaved kinetic trap of A1AT (1iz2 (Im et al., 2002)) with ten mutations, and the other is a cleaved form, with no mutations in its sequence, and co-crystallised with the substrate (1ezx (Huntington et al., 2000)).

Table 2-1': The Dataset of Selected Crystal Structures of A1AT used in this Study

Description	PDB id	Resolution (Å)	Mutations
Stressed – Native wild type	1qlp (Elliott et al., 2000)	2.00	None
Stressed – Native wild type	2qug (Pearce et al., 2008)	2.00	None
Stressed – Native with citrate bound	3cwm (Pearce et al., 2008)	2.51	None
Stressed – Native mutant	1hp7 (Kim et al., 2001)	2.10	Ala70Gly
Stressed – Native mutant	3drm (Gooptu et al., 2009)	2.20	Thr114Phe
Stressed – Native mutant	1oph (Dementiev et al., 2003)	2.30	Ph351Leu, Thr59Ala, Thr68Ala, Ala70Gly, Cys232Ser, Met358Arg, Met374Ile, Ser381Ala, Lys387Arg
Relaxed – Uncleaved RCL (Latent - kinetic trap)	1iz2 (Im et al., 2002)	2.20	Phe51Leu, Thr59Ala, Thr68Ala, Ala70Gly, Arg101His, Val364Ala, Met374Ile, Glu376Asp, Ser381Ala, Lys387Arg
Relaxed – Cleaved reactive loop	1ezx (Huntington et al., 2000)	2.60	None

c'. Identification of Surface Pockets and Calculation of their Properties

One protein chain from each crystal structure in the dataset was prepared using the Protein Preparation Wizard protocol available in the Schrödinger suite (*Maestro* package version 9.0 from Schrödinger, LLC). Ligands, waters and other co-crystallised agents were deleted and hydrogen atoms were added. The protassign script was used to optimise intramolecular contacts. The impref script was used to perform a restrained minimisation of the protein, with a maximum root mean square deviation (RMSD) of 0.30 Å.

All structures were superimposed on the native wild type protein (1qlp) using the structalign utility from Schrödinger. The site recognition software *SiteMap* 2.3 (*Maestro* package version 9.0 from Schrödinger, LLC) was run on all 8 crystal structures to identify the top 10 ranked potential ligand-binding sites. *SiteMap* uses an algorithm analogous to the Goodford's GRID algorithm (Weber et al., 1991), which uses interaction energies between the protein and grid probes to locate energetically favourable sites. Sites were kept if they comprised at least 15 site points. A restrictive hydrophobicity definition, a standard grid (1.0 Å) and the OPLS2005 force field were used (default settings in *SiteMap* 2.3).

The following physicochemical properties of the sites were calculated by the *SiteMap* program: size, volume, degree of enclosure/exposure, degree of contact, hydrophobic/-philic character, hydrophobic/-philic balance and hydrogen-bonding possibilities (acceptors/donors). In addition, *SiteMap* calculates two scores for each site; the *SiteScore* (Equation 2-1') and *Dscore* (Equation 2-2').

Equation 2-1': The SiteScore is defined as

$$SiteScore = 0.0733\sqrt{n} + 0.6688e - 0.20p$$

where ,

n = the number of site points (capped at 100),

e = enclosure,

p = hydrophilicity of the site (capped at 1.0).

Equation 2-2': The druggability score, Dscore, is defined as:

$$Dscore = 0.094\sqrt{n} + 0.60e - 0.324p$$

where,

n = the number of site points (capped at 100),

e = enclosure,

p = hydrophilicity of the site (uncapped).

The developers of *SiteMap* suggest that a cut-off in the SiteScore of 0.80 can be used to differentiate between drug-binding and non-drug-binding sites, with scores higher than 1.0 being indicative of highly promising sites (Wang et al., 2005). The Dscore can help to distinguish between undruggable and druggable sites, by penalising highly hydrophilic sites, as ligands binding to such sites would be very polar, and would be quickly eliminated by the organism. This does not mean that the site cannot bind any ligands, but that it would be difficult to find high affinity drug-like ligands for such a site (Halgren, 2009).

Nine sites were identified in the dataset of crystal structures of A1AT. These sites were labelled A to I. The geometric centre of each site as seen in the native wild type protein (1qlp), or, in the case of the I site, as seen in the structures bearing the Ala70 to Gly mutation (1oph, 1iz2 and 1hp7), was calculated and it was used to identify sites in all other crystal structures and computationally produced conformers. This was done as follows: If the geometric centre of a site k was within 3.75 Å of the geometric centre of any site s (where $s \in \{A,B,C,D,E,F,G,H,I\}$) then site k was assigned the letter of the site s (i.e. the two sites were thought to coincide). This cut-off is strict and it was chosen after manual inspection of several cases where sites were very close to each other, but where it was still possible to discriminate between them. Sites C and E overlap in many conformers and in these cases they were assigned the label "C_E". If the calculated distances for a new site were between 3.75 and 10 Å, the sites were inspected and assigned manually. If all distances were above 10 Å, the site was categorised as being new. Inspection of all "new" sites found in conformers of 1qlp led to approximately half of these sites being reassigned to one of the original nine sites (A to I). The remaining unassigned sites included mostly low-scoring sites, which were ignored in the present analysis.

d'. Generation of Protein Conformers using *CONCOORD*

CONCOORD 2.0 (de Groot et al., 1997) was used to produce alternative conformations for the native wild type proteins (1qlp and 2qug). The input structures were prepared with Schrödinger's Protein PreparationWizard, as detailed above. *CONCOORD* builds a library of distance constraints based on the observed interatomic distances in the original structure. Interactions deemed to be stronger are given tighter constraints. The program then produces randomly a large number of potential conformations, and attempts to correct structures with atom-pair distances falling outside the allowed regions. 1000 iterations were applied of the correction algorithm per structure, and the structures were rejected whose interatomic distances violated the original distances by more than 3 nm in total. *CONCOORD* was set to an output of 100 novel conformations for the native wild type proteins (1qlp and 2qug), which fulfilled the distance constraints. The maximum RMSD from the original structure was 2.96 Å. *CONCOORD* was also run to produce 5000 conformers based on the 1qlp structure. These were only used for comparison to the more limited 100 runs. All computationally produced conformers were superimposed on the native wild type (1qlp) using the structalign program.

The quality of the *CONCOORD* conformers was evaluated using the *PROSESS* server (Laskowski et al., 1994) available at http://prosess.ca. Table 2-2' contains a summary of these results.

Table 2-2': Overall Quality Results for Crystal Structures and *in Silico* Conformers of A1AT Selected for Docking Assessed by the *PROSESS* Server (http://prossess.ca)

PDB ID/ Conformer ID	Overall quality score	Covalent bond quality	Non-covalent/ packing quality	Torsion angle quality
1qlp	9.5	7.5	7.5	8.5
2qug	1.5	7.5	7.5	5.5
3cwm	7.5	7.5	7.5	7.5
3drm	9.5	7.5	8.5	7.5
1oph	9.5	7.5	8.5	7.5
1iz2	9.5	7.5	7.5	7.5
1ezx	9.5	7.5	8.5	7.5
Conf_77 (used for sites A, C)	6.5	7.5	7.5	7.5
Conf_85 (used for site B)	7.5	7.5	7.5	7.5
Conf_95 (used for site D)	6.5	7.5	6.5	7.5
Conf_53 (used for sites E, F)	6.5	7.5	7.5	7.5
Conf_20 (used for site G)	7.5	7.5	7.5	7.5
Conf_87 (used for site H)	7.5	7.5	7.5	7.5
Conf_57 (used for site I)	7.5	7.5	6.5	8.5

e'. Automated Assessment and Visualisation of Surface Pocket Variability

For each of the 100 *CONCOORD* conformers potential druggable sites were identified using *SiteMap* 2.3, as detailed previously. The *SiteMap* output files were then merged into single PDB files containing all predicted site sphere coordinates

and were used as input to the in-house pocket variability visualisation method Provar. The Provar method is explained briefly here: For each conformation, residues within 3.75Å of any *SiteMap* sphere were considered as being pocket-lining and assigned a score of 1, all other residues were assigned a score of 0. These scores were summed across all conformations and divided by the number of conformers (100) to assign each residue a probability value representing the likelihood that it borders a predicted site. These values were written to the B-factor column of the PDB file (1qlp), and results were displayed using Chimera. Residue atoms and ribbons were rendered on a continuous colour scale from white (low probability set to the value of the first quartile of the distribution) to red (high probability set to the value of the third quartile).

f'. Docking

Each of the nine sites (A to I) was used as a target for docking small molecules. For each site, the *CONCOORD* conformer that was selected to dock to was the one with the highest volume among the ones with the top five SiteScores as predicted by *SiteMap*. This selection was justified on the grounds that the highest SiteScore was not always associated with the largest cavity, but in rigid receptor docking a larger cavity, which allows more room for ligands to bind can potentially make up for the lack of side-chain flexibility during docking. Receptor grids were calculated with *Glide* (*Maestro* package version 9.0 from Schrödinger, LLC), keeping default settings. The grid box was centred on the calculated geometric mean of the particular site. The box side lengths were set to the maximum value of 14 Å.

All ligand libraries used in this study were prepared using *LigPrep* (*Maestro* package version 9.0 from Schrödinger, LLC). The preparation involved the generation of up to 32 stereoisomers (where these were not defined), tautomers, and protonation states corresponding to a pH of 7± 2 (using epik), as well as an energy minimisation of the 3D structure using the OPLS2005 force field. The DrugBank 3.0 (Wishart et al., 2008, Chang and Woolsey, 2006) library comprised 5897 entries after filtering to remove entries larger than 500 Daltons. Following *LigPrep*

preparation, this library consisted of 12115 small molecules. The library referred to as "ZINC fragments" is a representative library of fragments, based on the 3632 ZINC "clean fragments" subset clustered at the 60% Tanimoto similarity (downloaded from ZINC on the 05/06/2011). These clean fragments dataset obeyed the following criteria: xlogP<= 2.5, molecular weight <= 250 Daltons and number of rotatable bonds <= 5. Following *LigPrep* preparation, this library contained 5324 small molecules. Finally, a small subset of the PubChem library (1326 ligands related to thymol and extracted from PubChem, using the "Similar Compounds Search" on the web entry for thymol) was also prepared using *LigPrep*.

Glide with standard precision (SP) scoring was used for docking. Epik 2.0 state penalties were used in the final scoring. The highest scoring pose per ligand was kept and post-docking minimisation was switched on.

g'. Induced Fit Docking

A small number of molecules were selected for induced fit docking (IFD). All these molecules had shown promising *Glide* SP scores in preliminary docking trials, but some had poor scores following the inclusion of a protein preparation step. This suggested that IFD might be able to restore or even improve on the original scores, as it allows the protein side chains to optimise their position in the presence of the ligand. The IFD protocol (*Maestro* package version 9.0 from Schrödinger, LLC) available within the Schrödinger suite was employed (Sherman et al., 2006). Briefly, this protocol involves docking the ligand using a softened potential, and refining selected docked poses using Prime side-chain prediction and minimisation (Sherman et al., 2006). The refined protein conformations are then used for the final *Glide* docking step, where ligands are redocked, keeping the protein rigid. Default values were used for all *Glide* and Prime parameters. As the protein was prepared in advance no additional refinement was performed at this stage. For the initial *Glide* docking both the receptor and ligand van der Waals scaling were set to 0.50. Up to 20 poses were kept. The Prime induced fit step refined residues within 5.0 Å of the ligand poses by optimising their side chains. In

the final step, the ligand poses were redocked using *Glide* SP into structures within 30.0 kcal.mol^{-1} of the top 20 structures.

The IFD protocol was applied to a small selection of ligands docked in the I and C sites. This procedure was also applied to dock the CG compound (Mallya et al., 2007) to the A site in the native wild type A1AT (1qlp).

h'. ThermoFluor Studies

The ThermoFluor (fluorescent dye-based thermal shift) assay was performed using the iQ5 Real Time detection System (Bio-Rad – PCR Machine). Protein unfolding was monitored by measuring the fluorescence of the solvatochromic fluorescent dye SYPRO Orange, signalling unfolding of the protein. The compounds to be screened were dissolved in 100 % dimethyl sulfoxide (DMSO) to give a stock solution of 20 mM. The assay was performed in 96- well plates, each well totalling a volume of 25 μL. Every assay had a final concentration of 1 mg.mL^{-1} of A1AT, 1 mM of compound giving a final DMSO concentration of 5 %, to which 1 μL SYPRO Orange (1:200 dilution) was added. Furthermore, the influence of DMSO and A1AT concentration on the thermal shift were analysed. The DMSO concentration was varied to 5 %, 10 % and 15 % and the concentration of A1AT from 1 mg.mL^{-1} to 5 mg.mL^{-1}. Each trial was repeated 6 times (except the 5 mg.mL^{-1} concentration of A1AT, n=2). The starting temperature for each run was 10 °C increasing to 95 °C in 0.5 °C steps.

3'. Results

a'. α1-Antitrypsin High Resolution Structure

Lakshmi Segu (Gooptu group, ISMB/Birkbeck) had previously obtained a crystal, that was shot and its data collected at the Diamond Synchrotron (Didcot, UK). This data was then used to determine the highest resolution (1.8 Å) crystallographic structure of native A1AT to date (PDB entry 3ne4; (Patschull et al., 2011); Figure 3-1'a). Refinement statistics are listed in Table 3-1'. As expected, its overall fold and the positioning of secondary-structure elements were highly similar to the previous 2.0 Å resolution structure ((Elliott et al., 2000); PDB entry 1qlp; Cα

RMSD 0.3 Å ; Figure 3-1'b). However, the higher resolution was associated with a reduction in B factors overall (Figure 3-1'c) and improved confidence in details such as the positioning of side-chain atoms and water molecules.

Table 3-1': X-ray Data Collection and Processing Statistics for Native Wildtype α_1-Antitrypsin Crystal Structure 3ne4

Space group	**C2**
Cell dimensions (Å, °)	a=114.4, b=38.9, c=88.8, β = 104.3
Resolution (Å)	42.11 - 1.81 (1.91 – 1.81)
No. of reflections	92961 total 34169 unique
R_{merge}	0.07 (0.274)
Completeness (%)	98.5 (99.1)
Multiplicity	2.7 (2.6)
($I/\sigma(I)$)	10.0 (3.4)
$R_{\text{-cryst}}$ (%)	18.7
R_{free} (%)	23.3
B_{ave} (Å^2)	
Mainchain	23.9
Sidechain	28.8
No. of Water Molecules	217
RamachandranPlot, Residues:	
- Preferred region	96.5 %
- Allowed region	3.3 %
- Disallowed region	0.3 %
RMSD from ideal	
Bond lengths (Å)	0.015
Bond angles (°)	1.5

b'. Variability in the Solvation α1-Antitrypsin

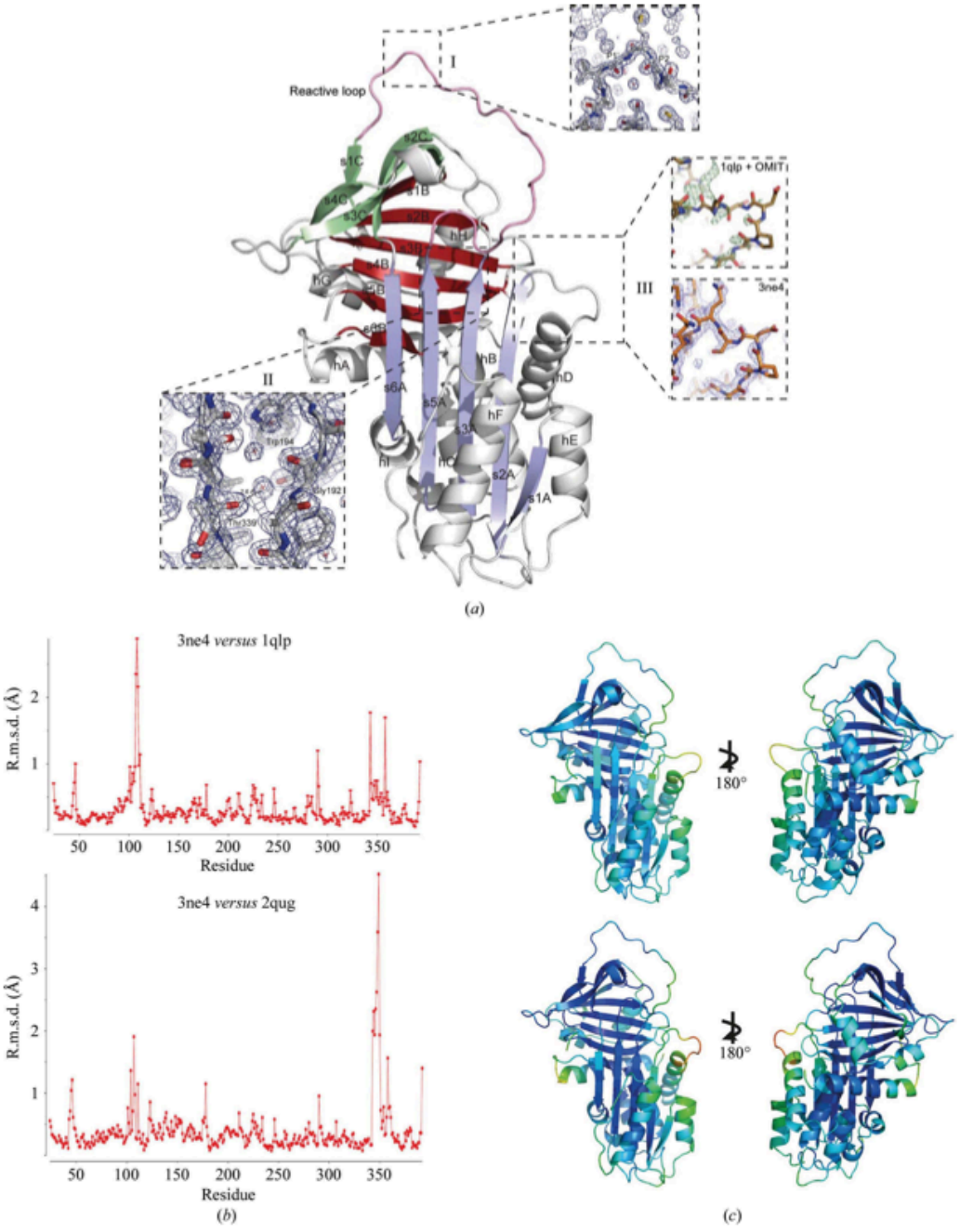


Figure 3-1': 1.8 Å Resolution Crystal Structure of α1-Antitrypsin

(a) 1.8 Å resolution crystal structure of A1AT (PDB entry 3ne4) with α-helices and β -strands labelled (e.g. helix A, hA; strand 1 of β-sheet A, s1A). Strands within a β-sheet are colour-coded together (A, blue; B, bronze; C, green). Detail is shown for the following. Box I, the reactive centre of the molecule in the canonical conformation. Box II, the 'breach' position that is the site of initial intramolecular loop insertion during monomeric conformational transitions. Box III, the fit of the

hD–s2A turn. The upper panel shows the rigid fit of 1qlp (gold) together with the initial OMIT map (Fo - Fc at 3σ density when residues 105–110 are omitted; positive difference density in green, negative in red). The lower panel shows the final fit of 3ne4 (orange) to the final map (blue, 2Fo - Fc at 1σ density). (b) RMSD for observed A1AT residues in 3ne4 compared with 1qlp (upper panel) and 2qug (lower panel) calculated using the SUPERPOSE program from the CCP4 suite (Winn et al., 2011). (c) Comparison of B factors in 1qlp (above) and 3ne4 (below). Low/high values are indicated by rainbow-spectrum colouring by PyMOL using a preset scale (blue for low to red for high). Whilst overall B factors are lower in 3ne4 (range 9.60–83.99 Å^2, mean 23.9 Å^2) than 1qlp (range 13.82–96.92 Å^2, mean 38.4 Å^2), the hD–s2A turn is associated with increased values in both relative to the global values. Other regions that show relative increases in B factor are the C-terminal end of helix A and the upper turn of helix F, which is believed to be dynamic in solution and to remodel during formation of the intermediate.

Occupancy of alternative rotameric orientations for Val216 and Ile340 became apparent during refinement. These are found on β-strand s4C and in the hinge region between β-strand s5A and the reactive loop. Coordination of the canonical inhibitory conformation at the reactive centre of the molecule by a water molecule between the side chain of Ser283 and the main-chain carbonyls of the P2 and P10 residues is confirmed in the new structure (Figure 3-1'a, box I). This water is observed in only one (PDB entry 1qlp) of the 2.0 Å resolution crystal structures of native A1AT solved previously. In the other case (PDB entry 2qug; (Pearce et al., 2008)), similarly to the 2.1 Å resolution structure (PDB entry 1hp7; (Kim et al., 2001)), this water is not seen and the canonical conformation of these residues is distorted. Accordingly, the RMSD is greater for 3ne4 and 2qug across the reactive-loop residues (340–362) than between 3ne4 and 1qlp (Figure 3-1'b). Moreover, the solvation environment between Trp194 and the plane of β-sheet A is clearly seen to involve three water molecules (numbered 7, 49 and 54) whose centroids lie within 3–5 Å of the Trp side chain (Figure 3-1'a, box II). They are coordinated by nearby main-chain carbonyl O atoms. This is of interest since changes around Trp194 are reported by changes in intrinsic fluorescence spectrometry of A1AT. It is therefore commonly used as a reporter residue for conformational change involving

rearrangements around its position underlying the top of β-sheet A (Dafforn et al., 1999, Tew and Bottomley, 2001). High-resolution structures of latent (Im et al., 2002) and cleaved (Yamasaki et al., 2010) A1AT clearly show the exclusion of solvent in this region. Previous structures of native wild-type A1AT (Pearce et al., 2008, Elliott et al., 2000) have indicated the presence of zero, one or two waters in this region. In 3ne4 one of the three water molecules hydrogen bonds to the carbonyl O atom of Thr339. This interaction prevents the formation of a typical interstrand hydrogen bond between the carbonyl of residue Thr339 at the top of s5A and Gly192 at the top of s3A. It therefore facilitates the opening of the upper s4A insertion site, which necessitates separation of these residues at the top of s3A and s5A. The upper s4A site superficially appears to be less accessible to reactive-loop or peptide annealing in A1AT compared with other native serpins. However, this finding shows how initial insertion of a residue in this 'P14 position' (Schechter and Berger notation) does not come at the cost of breaking an interstrand hydrogen bond in native A1AT.

c'. Conformational Variability in the A site

The most significant difference in main-chain conformation between the new structure and the search model 1qlp occurs at the hD–s2A turn (residues 105–110, Figure 3-1'b) that forms the upper boundary of the hydrophobic pocket targeted for allosteric polymerization blockade (Figure 1-1' and 3-3'; (Mallya et al., 2007)). This region is typically less well ordered than the overall fold in crystal structures of many native serpins, including A1AT. The improved resolution obtained in the current study aided confident fitting into observed density through use of an OMIT map (Figure 3-1'a, box III). The hD–s2A region is associated with relatively low B factors in latent (Im et al., 2002) and cleaved (Yamasaki et al., 2010) species in which the cavity is filled, but high B factors relative to other regions in crystal structures of native A1AT.

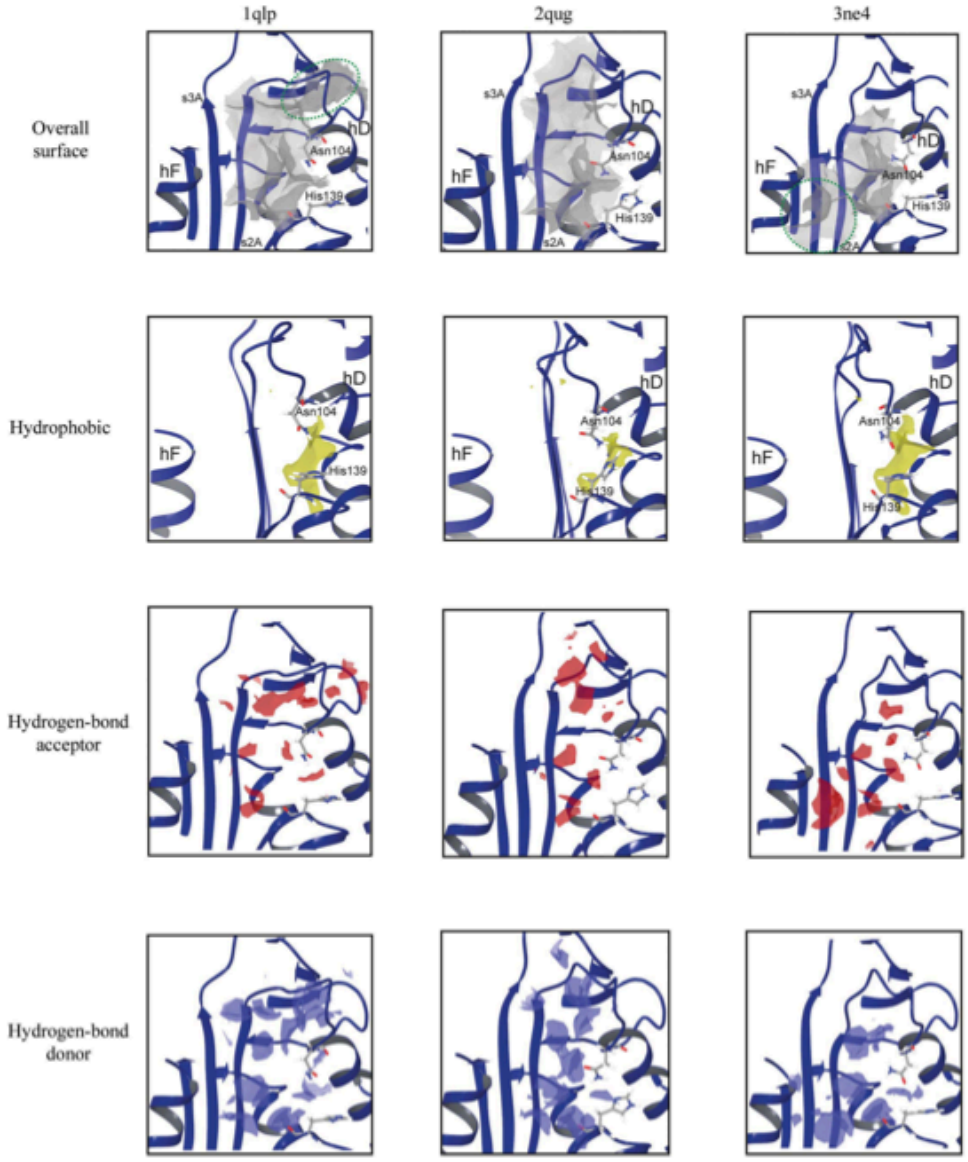


Figure 3-2': *SiteMap* Analysis of the A site in 1qlp, 2qug and 3ne4

The A sites in the 3 crystal structures at ≤2.0 Å of native A1AT (PDB 1qlp, 2qug and 3ne4) are shown as identified by SiteMap in a surface representation (top). The total surface (grey) is shown above the component parts that can participate in ligand binding through hydrophobic interactions (yellow), via hydrogen-bond acceptance (red) or via hydrogen-bond donation (blue). Green dashed lines demarcate dynamic channel topologies implied by the three structures.

Despite this and the differences observed here between the 1qlp and 3ne4 structures, they are both based upon data to high resolution and have good enough bond geometries to be reasonably confident of the accuracy of model building in

each case. Moreover, the turn is not near lattice contacts in either structure. The differences between 1qlp and 3ne4 are therefore likely to reflect alternative conformations of this region that are involved in conformational exchange.

Table 3-2': Cavity Characteristics as Calculated by the Program *SiteMap* for the A site

Structure	Size (No. of site points)	Vol. (Å^3)	SiteScore	Dscore	Exp.	Enc.
1qlp	123	252	1.009	1.029	0.606	0.709
2qug	78	183	0.937	0.945	0.639	0.703
3ne4	90	162	1.019	1.052	0.583	0.726
'Tight binders'	N/C	N/C	≥0.8	Sub-mM Kd correlates with ≥1.01	≤0.49	≥0.78

Structure	Contact	Phob.	Phil.	Bal.	Don/Acc
1qlp	0.908	0.713	1.033	0.691	0.714
2qug	0.905	0.557	0.943	0.591	0.967
3ne4	0.861	1.028	0.843	1.219	1.056
'Tight binders'	Mean 1.0	Mean 1.0	Mean 1.0	Mean 1.6	N/C

SiteMap output values are given for volume (Vol.), exposure (Exp.), van der Waals contacts (Contact), hydrophobicity (Phob.), hydrophilicity (Phil.) and the weighted balance of these characteristics (Bal.) and also for hydrogen-bond donor/acceptor ratio (Don/Acc). Overall scores and those for the general ligand (SiteScore) and drug-like compound (Dscore) binding characteristics are also shown. 'Tight binders' refers to values derived from observed correlation with or deliberate calibration against database of binding sites and ligand interactions characterized in vitro (Halgren, 2009). N/C, not calibrated by these studies.

Variability in the allosteric cavity (A site) was assessed by *SiteMap* analysis of high-resolution structures: 1qlp (which was first used to define it; (Elliott et al., 2000)), 2qug (Pearce et al., 2008), another 2.0 Å resolution crystal structure of native A1AT, and 3ne4 (Figure 3-2'). In addition to the variability in the hD–s2A turn region, the major differences observed between the A site in 1qlp and 3ne4 are

the topology at its upper and lower poles and, where topology is conserved, in the hydrogen-bond acceptor characteristics (Figure 3-2', red). The hD–s2A turn movement observed in 3ne4 relative to 1qlp abolishes an upper recess within the A site in the latter structure (top left panel, green ellipse). In contrast, in 3ne4 a groove at the lower pole of the A site entrance becomes continuous with it (top right panel, green ellipse). An innermost hydrophobic (Figure 3-2', yellow) chamber shows similar topology between 1qlp and 3ne4, as do the hydrogen-bond donor (blue) characteristics in the conserved core region.

The hD–s2A turn in 2qug more closely resembles that seen in 3ne4 than the same region in 1qlp (Figure 3-1'b). However, this feature alone does not appear to entirely dictate the overall cavity characteristics assessed by *SiteMap*. Thus, while the allosteric A site in 2qug displays a truncated upper channel relative to 1qlp, it does not become continuous with a channel at its lower pole. Moreover, the central region of the hydrophobic chamber seen in the other structures is lost in 2qug, dividing it. 2qug maintains similar hydrogen-bond acceptor characteristics of those cavity regions that are shared with the other two structures. However, the hydrogen-bond donor characteristics in the 2qug cavity are more concentrated within a narrower distribution than that in either 1qlp or 3ne4. *SiteMap* also scores sites for a number of parameters that have been correlated with tight ligand binding and druggability (i.e. tight binding of drug-like molecules; (Halgren, 2009)). These outputs are listed for the A sites assessed in the three different structures, together with cutoffs and mean values correlated with observed behaviour (Table 3-2'). The overall scores for ligand-binding propensities (SiteScore) and drug-like molecule binding propensities (Dscore) are also listed. These data are consistent with the topological observations in Figure 3-2' in quantifying variability around favourable characteristics for drug binding.

d'. Identification of Surface Pockets Present in Crystal Structures of α_1-Antitrypsin

The eight top-ranking surface clefts identified by *SiteMap* (Halgren, 2009) were denoted A-H and are shown on the structure of native A1AT (PDB entry: 1qlp) in Figure 3-4'. Sites A, D and G are each clearly distinct from other cavities, whereas sites C, B and E, as well as F and H are very close in space. The B, D and

G sites are all defined by loop regions. In the case of site B, the loop involved is the reactive centre loop (RCL). It is also interesting to note that sites C and E are proximal to the glycosylation site Asn247, whereas D is proximal to glycosylation site Asn46. The largest predicted site on the native wild type A1AT (1qlp) is site A, adjacent to strand 2 of β-sheet A. This site scores the highest for tight binding of drug-like ligands with *SiteMap* scores (SiteScore 1.03, Dscore 1.03) highly consistent with those observed in sites binding drugs with a submicromolar K_d (mean 1.01) (Halgren, 2009).

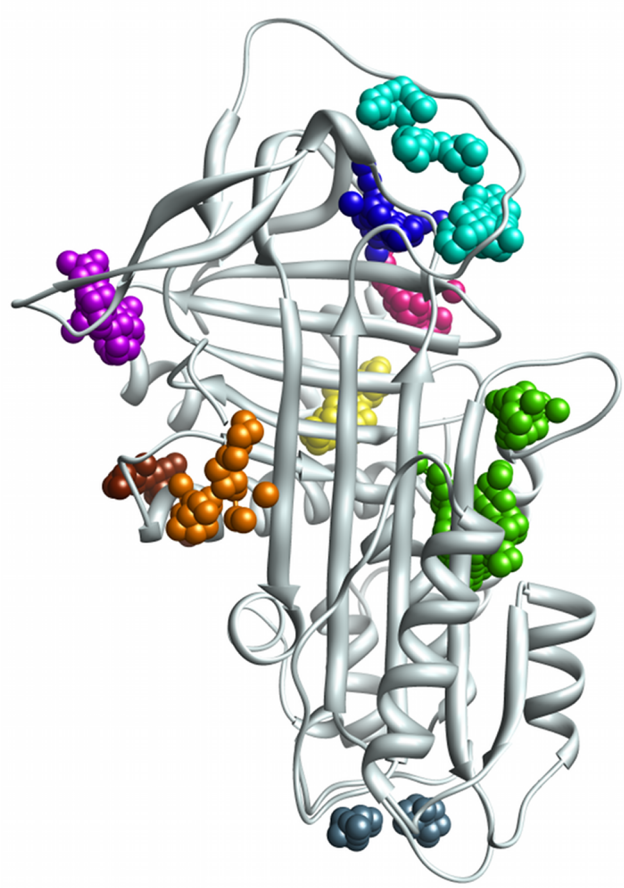


Figure 3-4': The Nine Top-Ranking Surface Pockets Identified by *SiteMap* on α_1-Antitrypsin

Coloured spheres represent the SiteMap predictions for eight top-ranking surface clefts on the wild type A1AT (PDB entry 1qlp, in grey cartoon representation): site A: green, B: cyan, C: blue, D: purple, E: fuchsia, F: orange, G: slate blue, H: brown. The yellow spheres correspond to the ninth site, I, a cleft identified on crystal structures of A1AT containing the Ala70Gly mutation.

Having identified potentially interesting sites on a single crystal structure, the persistence of these sites was assessed across the dataset of different crystal structures of A1AT (Table 2-1' and Figure 3-5'A). In structures containing the stabilizing mutation Ala70Gly (PDB entries: 1hp7, 1oph and 1iz2), an additional site was identified (here referred to as site "I"), located between the H-helix, the s4-s6 of the B ß-sheet and the A-helix. This site is small (45 $Å^3$), and very hydrophobic (the ratio of hydrophobic to hydrophilic character measured by *SiteMap*'s "balance" property is 5.1, with 1.6 being the average balance for tight-binding sites (Halgren, 2009)). Despite the small size of this site, the corresponding Site- and Dscores (0.92 and 0.92 respectively) calculated by *SiteMap* indicate a promising pocket for targeting with small molecule drug-like ligands. Although site I is present as a cavity in the remaining non-mutated structures, it is not solvent-accessible, and so is not identified by *SiteMap*. Interestingly, in PDB entry 1hp7, sites E and I are combined by *SiteMap* into one site, indicating that a ligand could possibly straddle both. Of the remaining eight sites, four are present in both the stressed and relaxed forms of A1AT (C, D, E, F), and four are only found in the stressed form (A, B, G and H). Results for the properties of each site are summarised in Figure 3-5'.

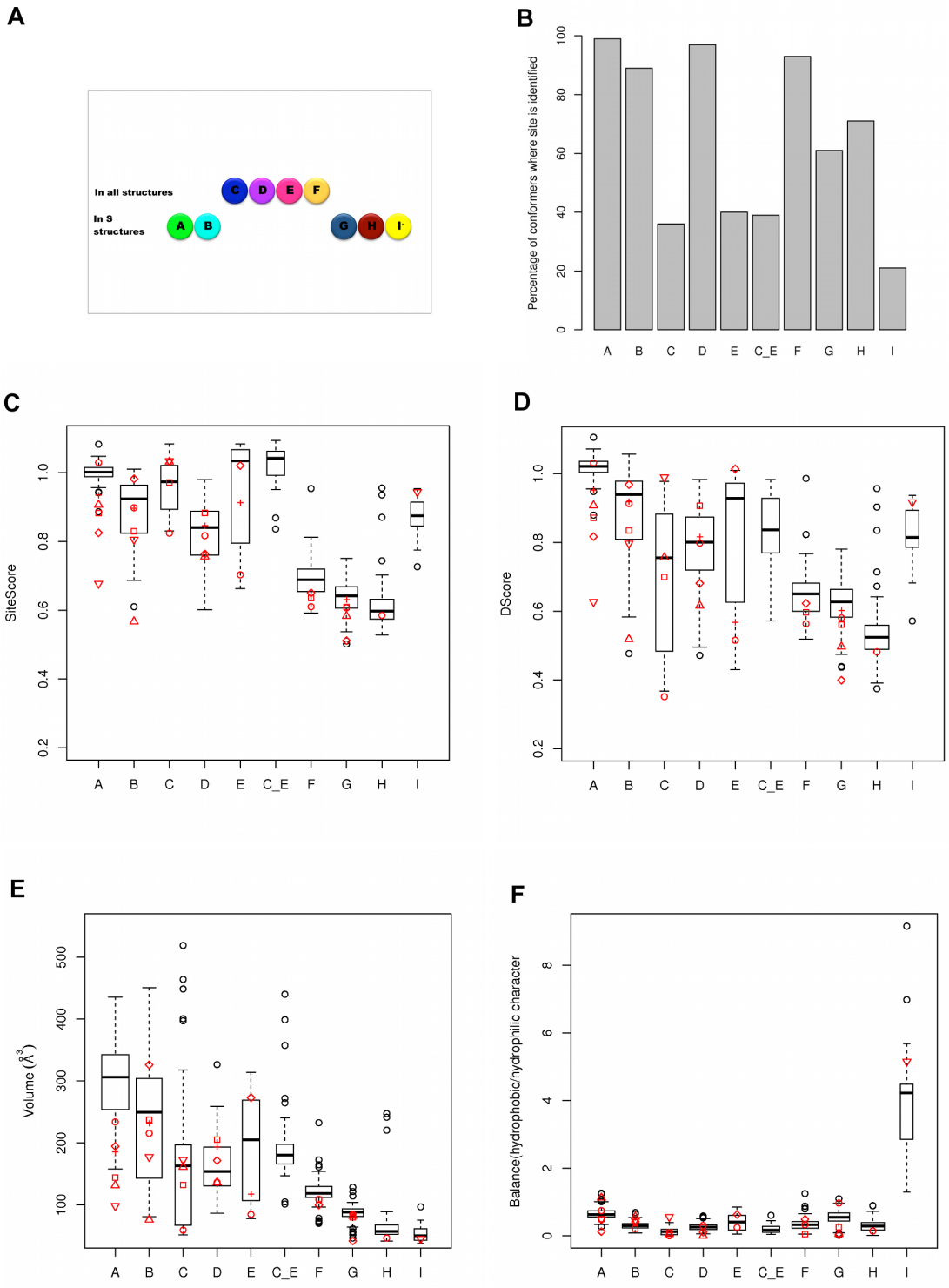


Figure 3-5': Properties of Surface Pockets in Crystal Structures and *in Silico* Conformers of α_1-Antitrypsin

Persistence of clefts A-I among A1AT crystal structures (A) and computationally produced conformers (B). Where the sites C and E overlapped, the data are

presented under the label "C_E". The distribution of SiteMap calculated properties for the 100 in silico conformers are shown as boxplots: SiteScore (C), DScore (D), site volume (E) and hydrophobic vs. hydrophilic character balance (F). The corresponding data for crystal structures are shown as red symbols superimposed on the boxplots; 1qlp (circle), 2qug (plus sign), 3cwm (square), 1hp7 (diamond), 3drm (triangle point up), 1oph (triangle point down). Data are shown only for sites identified within PDB entries for native (stressed, 'S') forms of A1AT, as these are likely to be the appropriate target states for the design of polymerization inhibitors.

A large variation is observed in the volumes of all the larger sites among the eight crystal structures studied reflecting significant conformational changes across this dataset. However, even the largest of these sites (1qlp, site A: 234 $Å^3$) is small compared with the average volume of drug-binding sites (reported as 600 to 900 $Å^3$, depending on the method used to measure them (Perot et al., 2010)). Nevertheless, six of the sites (A, B, C, D, E and I) have a median SiteScore higher than 0.8, the recommended value for distinguishing drug-binding from non-drug binding sites (Halgren, 2009). Sites A, C, and E demonstrate SiteScores >1.01, consistent with submicromolar drug-binding, in at least one crystal structure (Halgren, 2009).

e'. Incidence and Variability of Surface Pockets within a Computationally-Generated Conformer Ensemble

An ensemble of 100 A1AT conformations was generated from the native wild type structure 1qlp using the distance constraints-based method within *CONCOORD* (de Groot et al., 1997) (Figure 3-6'). *SiteMap* was then used to assess pockets A-I across the entire computationally generated native-like ensemble. The frequency of occurrence of each site across all conformers is summarised in Figure 3-5'B. The boxplots in Figures 3-5'C-F summarise selected *SiteMap* property results for these sites. Similar trends for the volumes and site scores are observed for conformations produced using more extensive sampling, or a different structure of native wild type A1AT (2qug) as the starting point for the *CONCOORD* simulation (data not shown). The majority of the values for the Site- and DScores (Figures 3-5'C and 3-5'D respectively), volume (Figure 3-5'E), and hydrophobic/philic balance (Figure 3-5'F) for pockets in the crystal structures are

within the boxplot limits. Thus the A1AT cavity characteristics explored by the computational conformers are supported by crystallographic observations. In addition, more detailed assessment of the computational conformers demonstrating the maximum Dscore for each cavity using the *PROSESS* server (Berjanskii et al., 2010) indicated they were of comparable quality to experimental structures in the dataset in terms of geometry and packing (Table 2-2').

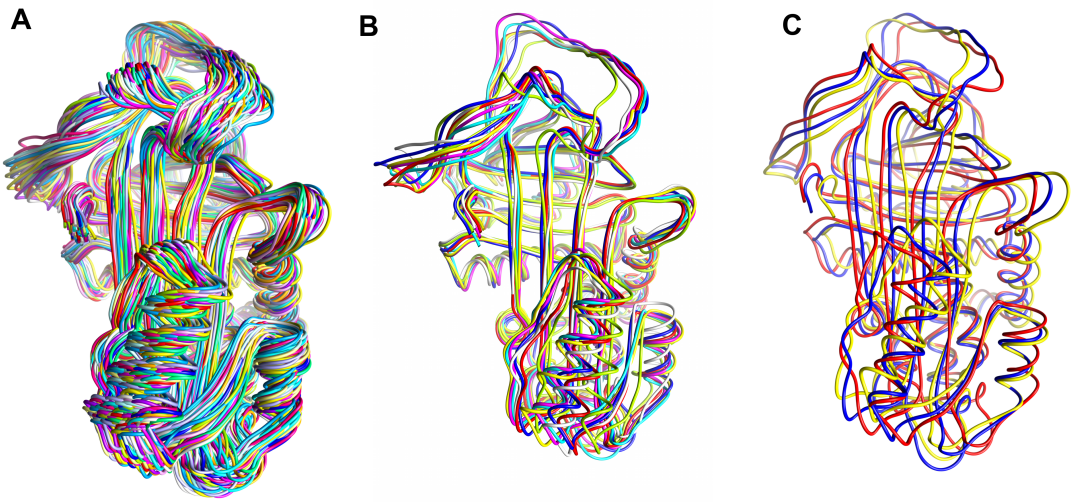


Figure 3-6': Exploration of Conformational Space of A1AT using *CONCOORD*

CONCOORD-generated conformers from a native wild type A1AT structure ((PDB: 1qlp). (A) All 100 conformers used to analyse druggability of sites and their occurrence. (B) The 7 structures used for docking to sites A-I; colours for conformers are: white (site G), magenta (sites E and F), cyan (site I), yellow (sites A and C), red (site B), blue (site H), green (site D). (C) Three selected conformers depicting the extent to which structural variation was simulated.

The behaviour of the A site across the computationally generated conformeric ensemble demonstrates the conservative nature of the conformational lability simulated by the program *CONCOORD*. Within the dataset of crystal structures of native A1AT the A site is largest and most druggable in 1qlp, the starting template for this *CONCOORD* simulation. The site is retained in 96% of the generated ensemble (Figure 3-5'B) and displays higher volumes (Figure 3-5'E) and druggability scores (Figures 3-5'C & D) across these conformers than observed across the crystallographic structures. Despite this conservative approach, the

ensemble generated by *CONCOORD* demonstrates that even these small fluctuations can have major consequences for surface clefts in A1AT, simulating pocket "breathing" in solution. Thus pocket volumes varied ≤ 3-fold for many sites (Figure 3-5'E) and druggability scores showed up to 2-fold variation (Figure 3-5'D). For many sites a source of high variability was their merging with other sites via formation of a channel of interconnected subsites. In particular, a channel ran from the RCL to the H-helix incorporating sites B, C, E and I in various combinations across several conformers (Figure 3-7').

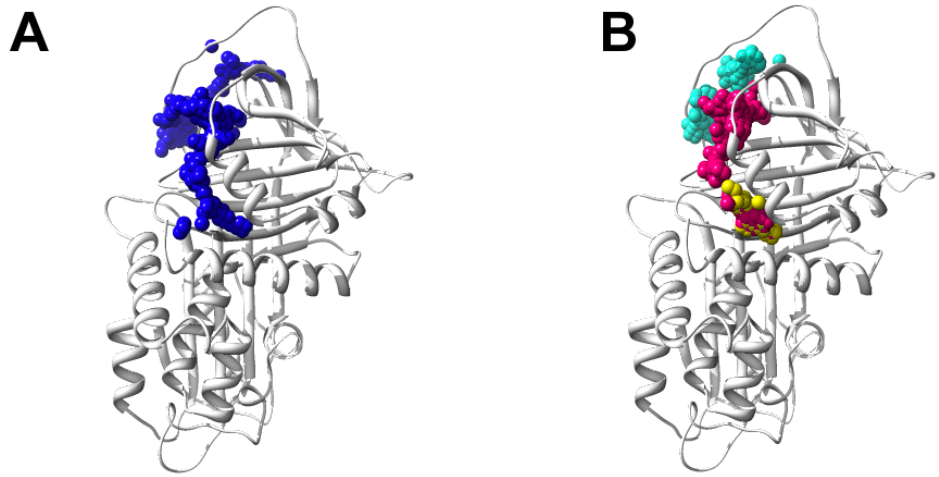


Figure 3-7': A Channel of Interconnecting Pockets on the Surface of A1AT

(A) A channel of interconnecting surface pockets (blue spheres) defined by the RCL at the top and the H-helix at the bottom can be seen in several in silico produced A1AT conformers. (B) This channel is split up into separate sites in most conformers: B (cyan), E (fuchsia), I (yellow). These subsites themselves occasionally overlap as in the case shown here, e.g. site E can "spill into" the spaces usually occupied by sites I and B.

A number of other sites have the potential to achieve druggability scores comparable to site A within the ensemble. However, the spread of scores across the conformational ensemble (Figure 3-5'C) indicates that the ligand-favouring properties of these sites are subject to greater fluctuation than the A site. Only three sites (F, G and H) have median SiteScores below the 0.8 recommended cut-off for promising drug targets. In general, the SiteScore for a pocket correlates with the volume of that pocket, but it is interesting that site I, although relatively small,

scores very highly (its median druggability score is highest after site A, among sites not defined by the RCL). This is probably due to its strongly hydrophobic environment (see Figure 3-5'F), which has highly favourable drug binding characteristics.

f'. Surface Cleft Variability Assessed by Provar

The variability of each predicted site in terms of the residues that line the site was assessed using Provar (Ashford et al., 2011), a method recently developed in the Nobeli group for the calculation and depiction of surface cleft variability. Provar uses an ensemble of conformers and their predicted pockets as input, calculates the propensity of each residue to line a pocket, and aids visualization by mapping the results on a single conformer structure. Provar results for the 100 *CONCOORD* conformers of A1AT are summarized in Figure 3-8'. The Provar analysis is consistent with *SiteMap* analysis data (Figure 3-5'B), and provides additional information about which residues are consistently part of a pocket and which are only occasionally so. For example, the majority of the residues lining the A pocket appear to be persistently part of a cleft across the *CONCOORD* conformer ensemble (Figure 3-8'C). By contrast, of the residues surrounding the I pocket, only three are consistently pocket-lining: Leu276, Ile375 and Lys380 (Figure 3-8'D). As the I pocket is only identified in about a quarter of all conformers, these residues must be often part of a different pocket that incorporates part of the I site. Moreover, Provar offers an insight into how conformational changes affect a pocket: pockets that have many of their residues coloured red (e.g. site A, Figure 3-8'C) are likely to be changing in volume (as evidenced also in Figure 3-5') by "breathing"-style motions that inflate and deflate the site without having much effect on which residues are pocket-lining. Sites that have many residues surrounding them coloured pink (e.g. site I) are either transiently observed, or change shape and volume by burying and exposing different parts of the site in different conformers. Such sites are consequently more likely to be missed by software that identifies pockets, if only one conformation or poor sampling of conformers is used.

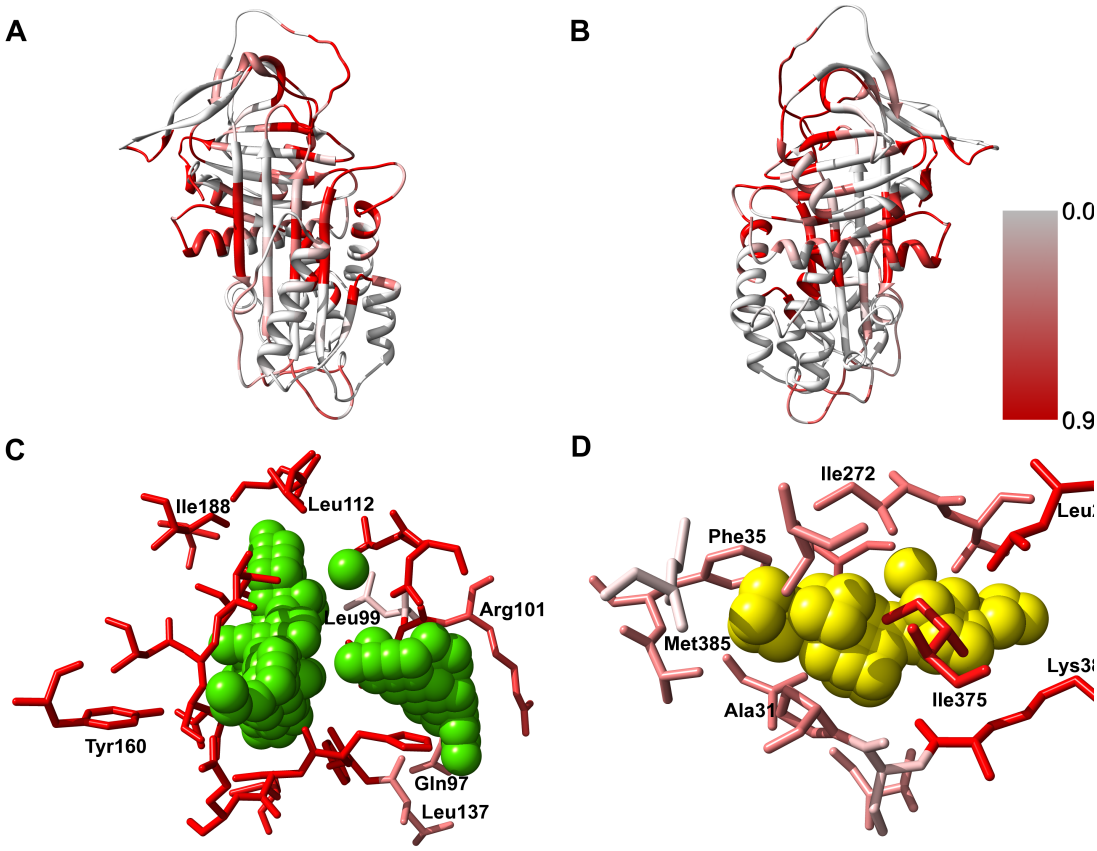


Figure 3-8'. The Pocket-Lining Propensity of the Residues of α_1-Antitrypsin Calculated with Provar

Ribbon representation of A1AT (front, A and back, B) coloured by the residue-based Provar probabilities. Provar colours each protein residue according to its probability of being pocket-lining in an ensemble of conformers (here, 100 CONCOORD-produced conformations of A1AT). The first (0.05) and third quartile (0.92) of the probability distribution are used as the white and red limits of the spectrum respectively. Hence, residues appearing red belong to the top quartile distribution, i.e., in this case, they are pocket-lining in more than 92% of the conformers. (C) and (D): The SiteMap predictions for two pockets (A and I respectively) are shown as solid spheres, and every residue with an atom within 3.75 Å of any sphere is shown in stick representation coloured by its Provar value. Depth-cueing has been switched off in these figures to preserve the variation in the colouring of the residues.

g'. Global Fragment and DrugBank Library Docking Studies

To further characterize the A-I pockets a set of representative fragments from the ZINC database and compounds from the DrugBank library were docked to each of the sites on A1AT using *Glide*. The *in silico* fragment screen identified high-scoring fragments for each site, highlighting chemotypes that may be used as starting points for future *in vitro* exploration. The ZINC identification codes for the 5 top-scoring fragments against each site are provided in Table 3-3'.

Table 3-3': Results for Top-Ranking Fragments against each of the Sites A-I on A1AT

Site	Rank 1 (score)	Rank 2 (score)	Rank 3 (score)	Rank 4 (score)	Rank 5 (score)
A	ZINC01581130 (-8.0)	ZINC00347000 (-7.8)	ZINC13728763 (-7.7)	ZINC08746511 (-7.6)	ZINC57218770 (-7.5)
B/C/E	ZINC49587279 (-7.72)	ZINC04521093 (-7.68)	ZINC02028426 (-7.35)	ZINC13283774 (-7.32)	ZINC01645671 (-7.29)
D	ZINC02293661 (-7.3)	ZINC00339659 (-7.1)	ZINC17377281 (-7.0)	ZINC01559484 (-7.0)	ZINC16037356 (-7.0)
F/H	ZINC13217456 (-7.1)	ZINC01678957 (-6.9)	ZINC05545529 (-6.9)	ZINC05286128 (-6.7)	ZINC25949941 (-6.7)
G	ZINC08627928 (-7.9)	ZINC04629171 (-7.1)	ZINC14983615 (-7.0)	ZINC38701009 (-6.9)	ZINC12403009 (-6.9)

The ZINC molecule identification codes and Glide SP docking score (within brackets, in kcal/mol) for each of the five top-ranking fragments docked to sites on A1AT are listed. Results for sites B, C, E and F, H are merged.

The docked poses of these fragments can be used to define pharmacophores for each site. Encouragingly, the top scoring fragments for the A site clustered in the area identified in a previous proof-of-principle study as a target for pharmacophores capable of blocking polymerization of A1AT while preserving inhibitory function (Figure 3-9'). The area of the pharmacophore is defined by Asn104, Thr114 and His139, and several of the fragment poses favour hydrogen bonds to the threonine and histidine residues.

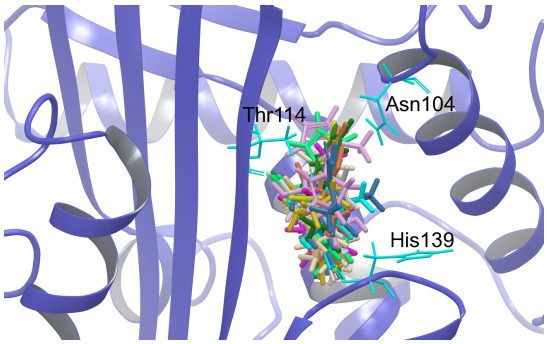


Figure 3-9': Fragment Docking to the A Site Targets the Pharmacophore Defined by Asn104, Thr114, and His139

Best poses of the top-scoring 20 fragments (coloured sticks) from the ZINC dataset docked in the A site of A1AT (cartoon, blue). The majority of these fragments fill the pocket defined by Thr114 and Asn104 at the top, and His139 at the bottom (thin sticks, cyan), identified in a previous study as a potential allosteric site for targeting A1AT polymerization. Some of the fragments take advantage of hydrogen bonding opportunities presented by His139 and Thr114.

The protein-fragment interactions within the other, less well characterized, sites provide great insight into the ligand-binding capabilities of these pockets. For example, the top 10 fragments in the I site have at least one hydrogen bond to one of three residues: Thr273 (side-chain oxygen OG1 acts as an acceptor to 5 ligands), Lys380 (backbone oxygen O acts as donor to 7 ligands) and His269 (ND1 acts as donor to 4 ligands). Moreover two areas within the site are often occupied by hydrophobic rings. These findings can be used to build a pharmacophore template for further searches of additional ligand databases.

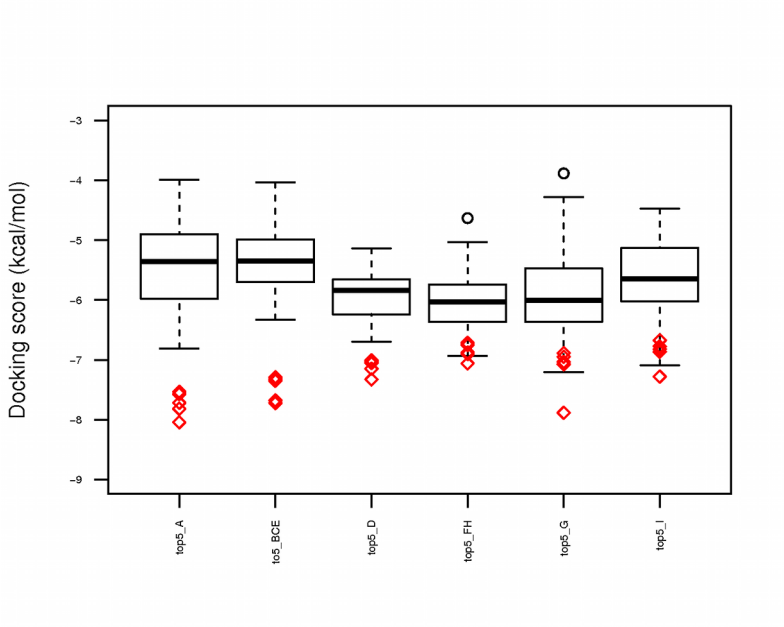


Figure 3-10': Site Specificity of High-Scoring Fragment Molecules

Red diamonds represent the docking scores for the top 5 scoring fragments for each of the sites A, BCE, D, FH, G, and I. The boxplots summarise the corresponding (merged) distributions of docking scores for the same five fragments docked to all other sites.

Overall the sites identified by *SiteMap* analysis demonstrated specificity even when probed with small fragment compounds, that are intrinsically more likely than larger compounds to bind promiscuously (Chen and Shoichet, 2009). Top-scoring fragments for each site typically scored better for binding at that site than against any other site (Figure 3-10').

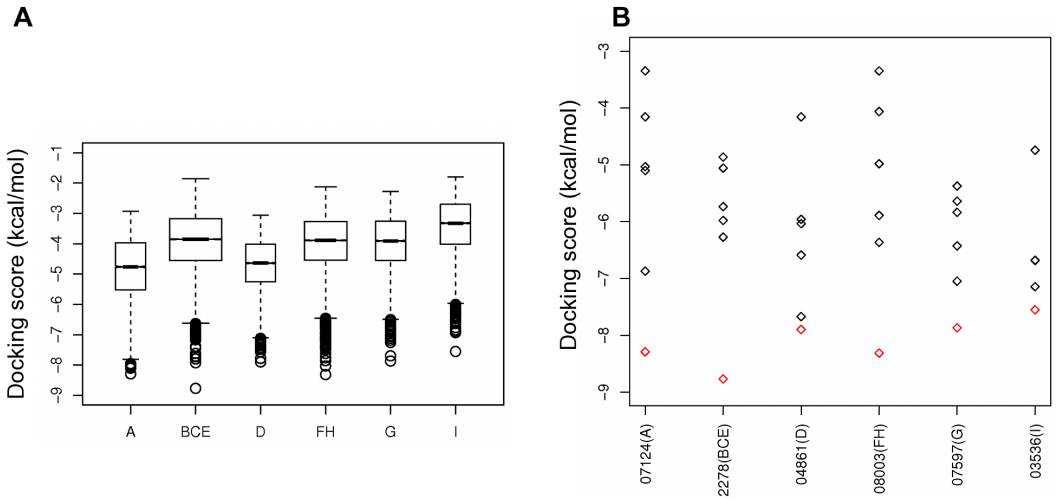


Figure 3-11': Results from Docking the DrugBank Collection against Nine Pockets on α_1-Antitrypsin

(A) Boxplot distributions of docking scores for DrugBank molecules docked to each of the nine sites A to I. Only the top-ranking pose is included for each ligand and only ligands of molecular weight less than 500 Daltons are included in this plot. (B) The best-scoring ligand for each site is assigned a worse score when docked against each of the other sites. The red diamonds represent the best docking score for each ligand depicted in Table 3-4', when docked to the site where it is ranked top. The black diamonds correspond to the scores for each of these ligands when docked to all other sites. The x-axis labels correspond to the DrugBank ID of the ligand and, in brackets, the site against which it is selected as "best-scoring", e.g. 07124(A) refers to DrugBank entry DB07124 which achieves its best score against site A.

In the second docking experiment scanned all pockets with the DrugBank collection of small molecules in an effort to identify any high-ranking ligands that are already used, or being tested as drugs for different targets. 12,115 small molecule ligand structures based on 5,897 molecules from the DrugBank library were docked using *Glide* (see Methods for details) to each of the nine surface clefts A-I. Docking scores for each ligand successfully docked to each site are summarized in Figure 3-11'. In these plots the distribution of docking scores for sites B, C and E (labelled as site BCE) and those for sites F and H (FH) are combined, as the necessity of using a reasonable-size receptor grid in docking

permits ligands to dock within neighbouring sites to that upon which the grid is centred. As is usual for docking calculations, the majority of the ligands interacted *in silico* with relatively poor predicted binding energies (-3 to -5 kcal/mol), indicating poor potential for drug development. However, promisingly, low energy outliers in these distributions achieve scores in the range of -7.5 to -8.8 kcal/mol for each site (Table 3-4' and Figure 3-12'). These scores are comparable to the score of compound "CG", a molecule identified in a previous study as an inhibitor of A1AT polymerization (CG achieves a score of -8.7 kcal/mol against its target site (A) after induced fit docking using *Glide*). Moreover, the best-scoring ligand for each site appeared highly selective for that site (Figure 3-11'B). The best overall scores were achieved for sites BCE and FH. The highest-scoring ligand interaction was for 7,8-dihydro-7,7-dimethyl-6-hydroxypterin (DrugBank ID DB02278). Despite the relatively small size (209 Da) of this ligand, it achieved a score of -8.8 kcal/mol against the BCE site. However in the simulations, this molecule bound the RCL with likely adverse effects on the enzyme inhibitory function of A1AT.

Table 3-4': The Best-Scoring and "Best-Efficient" Small Molecules from DrugBank Docked against each of the Sites A-I on A1AT

	Best overall docking score			**Best scoring within ten most efficient**		
Site	DrugBank ID	Glide SP score (kcal/mol)	Molecular Weight (Daltons)	DrugBank ID	Glide SP score (kcal/mol)	Molecular Weight (Daltons)
A	DB07124	-8.3	384.4	DB00610	-7.9	167.2
B/C/E	DB02278	-8.8	209.2	Same as best overall		
D	DB04861	-7.9	405.4	DB02377	-7.2	150.1
F/H	DB08003	-8.3	486.5	DB00529	-6.8	126.0
G	DB07597	-7.9	163.2	Same as best overall		
I	DB03536	-7.5	379.4	DB03329	-6.2	111.2

Diagrams, IUPAC names and PubChem CIDs for all DrugBank entries in this table can be found in Figure 3-12'.

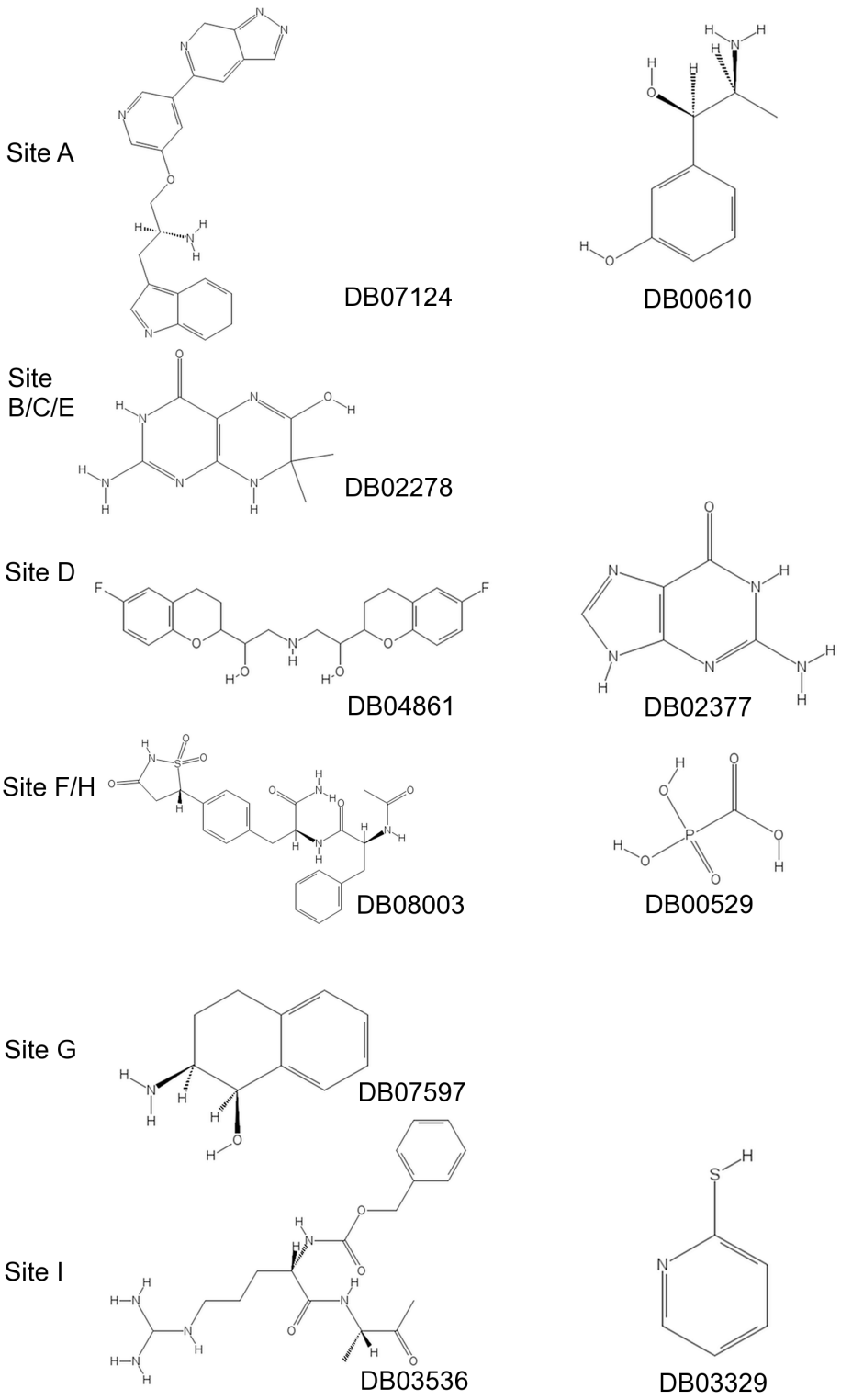


Figure 3-12': Top-Scoring DrugBank Molecules against the α_1-Antitrypsin Sites

IUPAC names and PubChem CIDs for the DrugBank IDs in Figure 3-11' and Table 3-4' are:

DB07124: 3-[(2S)-2-amino-3-[(5-{7H-pyrazolo[3,4-c]pyridin-5-yl}pyridin-3-yl)oxy]propyl]-6H-indole, PubChem CID: 46937052;
DB00610: 3-[(1R,2S)-2-amino-1-hydroxypropyl]phenol, PubChem CID: 5906
DB02278: 2-amino-6-hydroxy-7,7-dimethyl-3,4,7,8-tetrahydropteridin-4-one, PubChem CID: 3340355;
DB04861: 1-(6-fluoro-3,4-dihydro-2H-1-benzopyran-2-yl)-2-{[2-(6-fluoro-3,4-dihydro-2H-1-benzopyran-2-yl)-2-hydroxyethyl]amino}ethan-1-ol, PubChem CID: 71301;
DB02377: 2-aminopurin-6-one, PubChem CID: 764;
DB08003: (2S)-2-acetamido-N-[(2S)-1-amino-1-oxo-3-[4-[(5S)-1,1,3-trioxo-1,2-thiazolidin-5-yl]phenyl]propan-2-yl]-3-phenylpropanamide, PubChem CID: 9547915;
DB00529: phosphonoformic acid, PubChem CID: 3415;
DB07597: (1R,2S)-2-amino-1,2,3,4-tetrahydronaphthalen-1-ol, PubChem CID: 6420129;
DB03536: benzyl N-[(1S)-4-[(diaminomethyl)amino]-1-{[(2S)-3-oxobutan-2-yl]carbamoyl}butyl]carbamate, PubChem CID: 6398520;
DB03329: pyridine-2-thiol, PubChem CID: 2723698.

Since larger compounds (>350 Da) are considered unfavourable as leads for drug design, ligand efficiency was used to identify the 10 best performing ligands for each site. Ligand efficiency is traditionally defined as the docking score divided by the number of heavy atoms, but the natural logarithm of the ligand efficiency, is proposed to give a better fit to experimental data (Halgren, 2009, Sheridan et al., 2010) and so this measure was used instead. Within these best ligand efficiency sets the ligand with the best overall docking score was selected to avoid overcompensating for size at the expense of docking score. Some of these ('best-efficient') ligands conserved interactions that are important in the binding of the highest scoring ligand overall ('best-overall'). Thus, within the I site, hydrogen bonding of a charged amine group to the backbone of Ser140 was seen with both the best-efficient (DrugBank ID: DB00610) and best-overall (DB07124) ligands. Similarly the aromatic ring of the most efficient ligand for the I site (DB03329) overlaps with the positions of all other aromatic rings in the top 10 scoring ligands.

h'. Induced Fit Screening for Promising I Site Ligands

For a flexible protein, like A1AT, rigid receptor docking is likely to miss many ligands that require small structural rearrangements in order to fit some of the smaller sites. In this case, docking calculations that allow for induced fit are recommended. The induced fit protocol was applied to the I site, as this is the smallest of all and more likely to benefit from such a protocol. Hereby ligands are docked into sites in a soft mode (repulsive forces are very much reduced), then the protein and the ligand are allowed to relax. Finally the ligand is redocked to the relaxed conformer of the receptor. The induced fit docking protocol dramatically changes the results for some ligands.

Two natural compounds that gave promising results were menthol and thymol. Menthol (DB000825) is a natural compound of mint oils that scores reasonably well (-6.6 kcal/mol) in the original docking trial (with the receptor kept rigid) and, more importantly, ranks eighth out of the 10,000 reported ligand poses. Following induced fit docking, this score improves dramatically to -8.5 kcal/mol, aided by a small rearrangement of His269, which results in an additional hydrogen bond to the ligand. Thymol is another interesting hit against site I. In preliminary docking experiments (without prior protein refinement in *Glide*) thymol was the fourth best scoring molecule against this site. Thymol is a natural product of thyme and a known protein binder (Vincent et al., 2000) that is used as a stabilizer in pharmaceuticals as well as an antiseptic, vermifuge, antibiotic and fungicide, so it may be an interesting ligand to explore. Unlike many of the larger ligands that were found bound mostly on the outside of the cavity, thymol docked inside and showed a good complementarity to the site. Following protein refinement (a recommended procedure in *Glide*), thymol could not be docked inside the I site, resulting in a very poor docking score (Figure 3-13'A). However, after induced fit docking thymol could enter the cavity and achieved a *Glide* score of -8.3 kcal/mol (Figure 3-13'B). Finally, a series of molecules comprising the thymol scaffold resulted in several good hits, the top-scoring one being 5-ethyl-2-(4-ethyl-2-hydroxy-phenyl)phenol (PubChem CID: 19850961), which binds the I site with an impressive score of -10 kcal/mol. This score is equivalent to a K_d prediction in the nanomolar range (Figure 3-13'C).

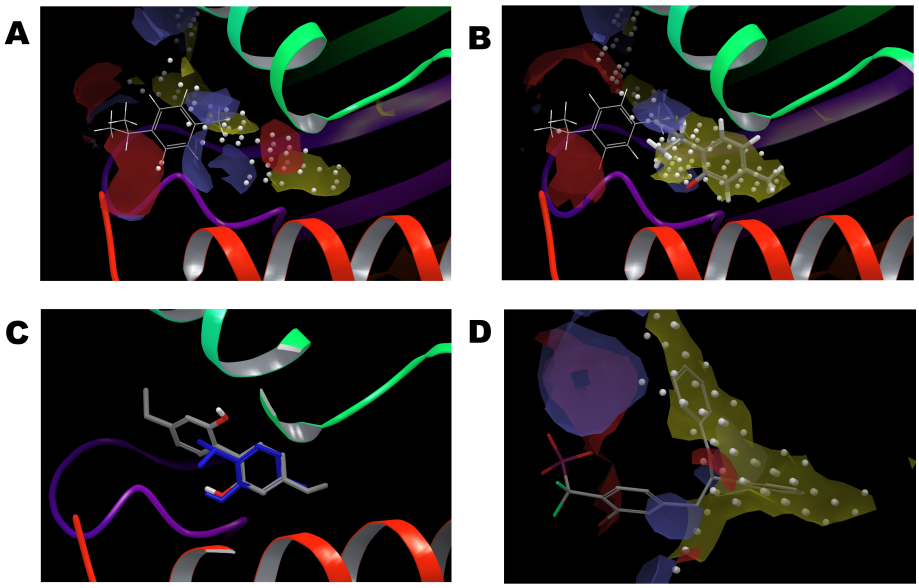


Figure 3-13':. Induced Fit Docking allows the Discovery of High Affinity Hits for Site I

(A) Thymol (DrugBank ID DB02513, in wire representation) docks on the outside of the main cavity of the I site (small white spheres) and does not reach the hydrophobic pocket within the cavity (yellow surface), resulting in a poor docking score (-3.2 kcal/mol).

(B) After induced fit docking, thymol (in stick representation) enters the site, which now comprises a larger hydrophobic cavity; the docking score is consequently greatly improved to -7.8 kcal/mol. The initial docked pose of thymol before the application of IFD is shown superimposed in wire format.

(C) A derivative of thymol, 5-ethyl-2-(4-ethyl-2-hydroxyphenyl)phenol, (PubChem CID 19850961, sticks coloured by element) achieves an impressive score of -10 kcal/mol after induced fit docking, whilst retaining the original thymol pose (in blue) for the substructure that is common to both molecules.

(D) Best-ranking pose for DrugBank ID DB07263 ([{2-bromo-4-[(2R)-3-oxo-2,3-diphenylpropyl]phenyl}(difluoro)methyl]phosphonic acid, in stick representation) following induced fit docking. In this protein conformer, the channel connecting sites I and C has been opened creating two hydrophobic subpockets (predicted by SiteMap and depicted here in yellow semi-transparent surface). Two of the aromatic rings of this ligand are placed in these subpockets. This ligand achieves a very good docking score (-9.5 kcal/mol), despite the fact that several hydrogen

bonding opportunities (depicted by the blue and red surfaces, corresponding to H-bond donor and acceptor, respectively) are not satisfied in the case of this ligand.

Interestingly a further hydrophobic pocket may transiently form next to the originally identified I site and in some conformers, is continuous with it. This can allow larger ligands with two rings connected by a flexible linker to dock in a way that takes advantage of both hydrophobic patches. For example docking DrugBank entry DB07263 using the induced fit protocol, gives the pose depicted in Figure 3-13'D where two of the aromatic rings are placed in the two hydrophobic subpockets making up the site in this conformer (yellow surfaces in Figure 3-13'D). This pose achieves a very respectable *Glide* score of -9.5 kcal/mol. As this particular ligand does not take full advantage of the hydrogen bonding opportunities clearly depicted in the *SiteMap* surfaces of the site (Figure 3-13'D surfaces in blue and red). It is therefore reasonable to hypothesise that the affinity could be further improved by adding suitable functional groups that could interact with polar residues on the receptor.

i'. ThermoFluor Experiments Validate Interactions Predicted in Silico

A small number of hits from the docking studies were assayed using thermal shift experiments (ThermoFluor). All compounds selected for testing had shown promising docking scores either using the induced fit protocol, or in preliminary docking studies using rigid receptor docking. Table 3-5' summarises the results for the three of the eight ligands tested that demonstrated significant thermal shifts (4-nitrocatechol, 2,6-diisopropylphenol and thymol), compared with the control substance (DMSO). The corresponding ThermoFluor graphs for these three ligands can be found in Figure 3-14'; they show the difference between the assayed melting temperature of the incubation with compound and the average melting temperature of an appropriate A1AT control incubated under the same conditions (in this case, DMSO concentration). One of the eight compounds assayed (4-nitrocatechol, predicted to bind at the I site) demonstrated an average thermal shift exceeding 1 °C. Interestingly, two of the compounds representing hits to the I site (thymol, and 2,6-diisopropylphenol) appeared to destabilise the protein, causing a negative shift

in the melting temperature, which was particularly pronounced for thymol (average of -1.66 °C). Negative shifts may also be due to the hydrophobic nature of the compounds, which under the assay conditions may induce non-specific destabilization of the folded state (Cimmperman et al., 2008). Whether stabilizing or destabilizing, the observed shifts in the melting temperature of A1AT support the docking results and indicate that the compounds assayed are most likely interacting with A1AT.

Table 3-5': Shifts in Melting Temperature of A1AT in the Presence of Selected Small Molecule Ligands (ThermoFluor Assay)

Molecule Name	DrugBank ID	Average Thermal Shift in °C (*p*-value)	Predicted site of binding	Best Glide SP score after IFD (kcal/mol)
Thymol	DB02513	-1.66 (0.0089)	I	-7.8
4-Nitrocatechol	DB03407	1.92 (0.0002)	I	-6.9
2,6-Diisopropylphenol	DB00818	-1.21 (0.0021)	I	-8.3

The quoted p-values are the result of a Welch two-sample t-test (performed using the R statistical software) testing the null hypothesis that the difference in the mean values of the distribution of the thermal shift values for DMSO and the distribution of the thermal shift values observed for each ligand is zero. The null hypothesis was rejected for p-values < 0.01.

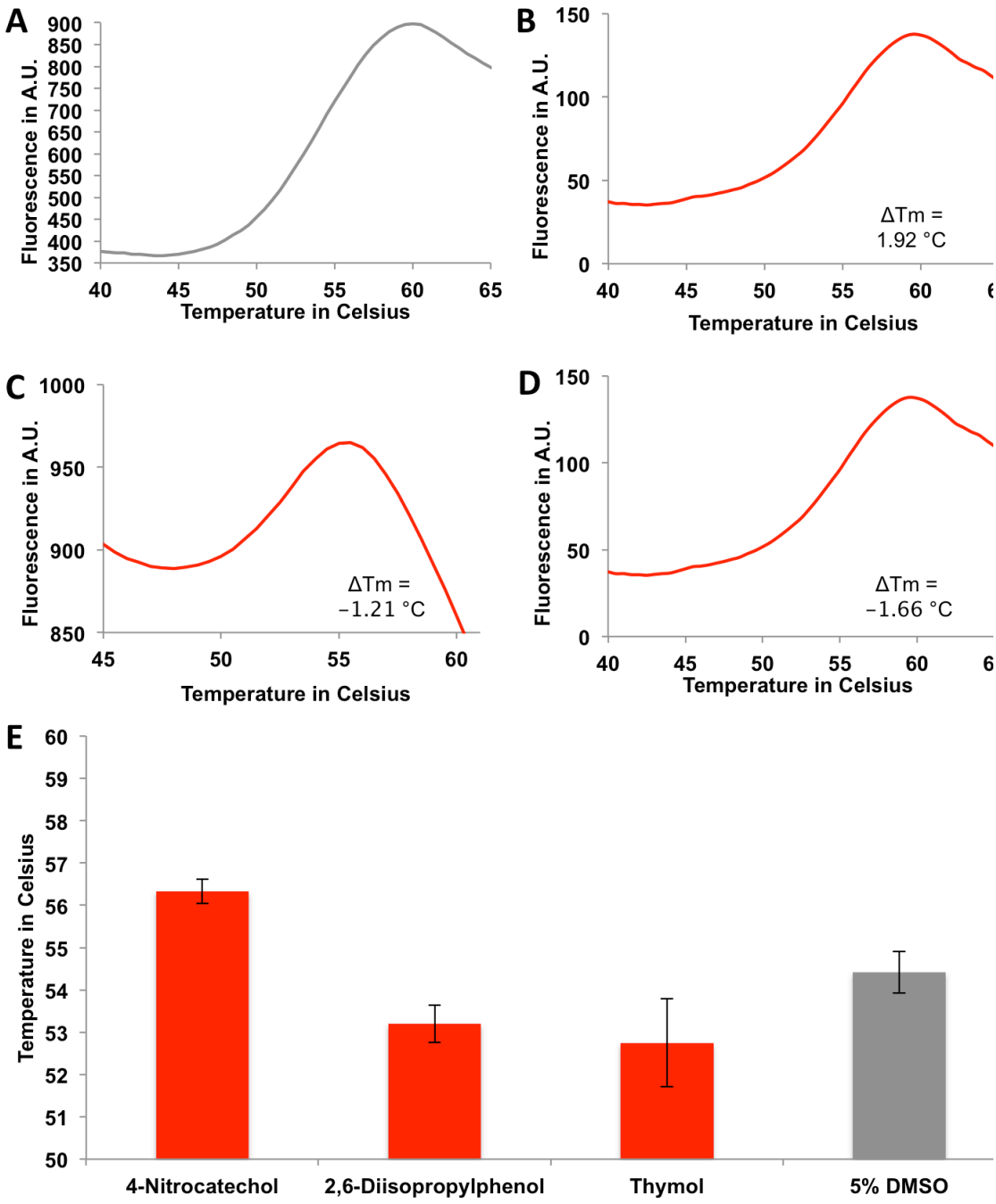


Figure 3-14': Thermal Shift and Melting Temperature Assays for A1AT Incubated with Selected Ligands

Fluorescence-based (Thermofluor) thermal shift assay curves for A1AT incubated with small molecule ligands. Only ligands with significant thermal shifts are shown. Representative curves obtained in the presence of these ligands (solubilised in DMSO, final concentration 5% (v/v)) are shown in plots A to D (control with 5% DMSO in grey, data from incubation with ligands in red). The mean ΔT_m is shown for A1AT incubated with each ligand: (A) 5% DMSO control, (B) 4-nitrocatechol,

(C) 2,6-diisopropylphenol, (D) thymol. (E) Mean melting temperatures and standard deviations for A1AT incubated with these three ligands (red) or 5% DMSO control (grey).

4'. **Discussion**

The use of relatively low-resolution crystal structures as a supplement to high-resolution structures has been proposed as a promising strategy for sampling the conformational space explored by drug targets and thus aiding drug design (Furnham et al., 2006). In the case of the site A, comparison between the 1qlp, 2qug and 3ne4 structures of A1AT provides the benefit of this outcome without the potential inaccuracies of model building inherent at lower resolutions.

Transiently druggable pockets on the surface of proteins can be missed by *in silico* screens to identify the most promising target site on a protein, commonly based upon a single structural snapshot. Such pockets are of particular interest in cases where the protein target undergoes large conformational variations, as in the archetypal serpin A1AT. The alternative methodology presented here characterizes more pockets, and simulates their solution behaviour in greater detail than a single conformer/single pocket approach.

In this study, efforts were focused upon identifying druggable pockets on the surface of native A1AT that could be the targets of inhibitors blocking polymerization. Previous *in silico* attempts to identify small molecules that can act as inhibitors of polymerisation have concentrated on one prominent allosteric site (defined here as the A site), a large cavity between the ß-sheet A and the D-helix (Elliott et al., 2000). This site was seen as a good drug target, as the space filling Thr114Phe mutation situated in the A site reduces polymerisation and preserves inhibitory function of native wild type A1AT *in vitro*, and increases secretion in a mammalian cell model of disease (Parfrey et al., 2003, Gooptu et al., 2009). Drug design studies based on the Thr114Phe mutant and *in silico* research focusing on this site have led to ligands that blocked polymerisation of A1AT *in vitro* (Chang et al., 2011). However, they did so irreversibly and with the undesirable side effect of blocking the inhibitory action of A1AT (Gooptu et al., 2009, Mallya et al., 2007).

To improve targeting to this site, the new 3ne4 structure that was solved provides direct observational data defining breathing motions of the A-site.

However, there is both scope and need for targeting alternative sites on A1AT. A recent attempt at identifying such sites across a range of serpins has revealed at least one site where selected sugars and amino acid derivatives may bind, acting as chemical chaperones that reduce polymerization (Singh et al., 2011). Therefore in parallel potentially druggable sites on A1AT were identified, that have not yet been targeted in *in silico* screens.

To validate alternative potentially druggable sites identified by *in silico* exploration of multiple conformeric variants derived from high resolution structures of native A1AT their presence in the larger set of lower resolution structures was investigated. Indeed, crystal structures of A1AT allow us a glimpse of the variety of conformations sampled by this protein. This inherent flexibility, intimately linked to function, is dispersed across the whole protein (Im et al., 1999, Im et al., 2002, Seo et al., 2000, Seo et al., 2002, Ryu et al., 1996) and thus potentially reflected in the properties of pockets on the surface. Analysis of available crystal structures revealed considerable variability in the surface clefts between different conformers, and suggested that this variability should not be ignored in structure-based drug design. The work showed that the variability of potential druggable pockets could be extensively probed using a relatively cheap, constraints-based computer simulation that efficiently explores part of the protein conformational space. Additionally, this approach identified both novel (transient) sites, and also pre-existing pockets deemed non-druggable in a single crystal structure that could attain druggable characteristics in the solution ensemble. Identification of such sites is the first step towards a structure-based drug design strategy that would seek to stabilize conformations where these sites are present and druggable. Such an approach may be particularly fruitful in proteins like A1AT, where the design of small molecule modulators has to strike a delicate balance between stabilizing the stressed state in order to reduce the protein's tendency to polymerise, and preserving the protein's antiprotease function.

Provar analysis of the variability of each pocket provides insight into the basis of this variability. This is useful for *in silico* induced-fit type screening within high throughput studies, where it is necessary to keep the number of residues that are allowed flexibility as low as possible. The combination of druggability and variability predictions may be relevant for many proteins that are deemed difficult to target due to their flexibility. Therefore, automating this process is now an ongoing goal of the Nobeli group.

The conformers in which each pocket achieved its highest druggability score were selected for docking studies, employing the publicly available database of marketed and experimental drugs DrugBank. These docking experiments highlighted several low molecular weight ligands that scored well on individual sites and were specific for these sites. Promisingly, several of the docking scores of the best-scoring ligands at the novel targets are comparable to the docking score of compound "CG", a molecule previously identified as an inhibitor of A1AT polymerization *in vitro* and in mammalian cells (Mallya et al., 2007).

This approach has revealed sites with potential for future *in vitro* studies. A small but very hydrophobic site (site I) that is present in about one fifth of the *in silico*-produced conformers was initially identified by *SiteMap* in three crystal structures, carrying the Ala70Gly mutation. This mutation is known to increase the stability of the stressed state, oppose the propensity to polymerisation and retain the functionality of the protein, while inducing widespread changes in cavity sizes within A1AT. Further analysis showed that this site is present in all other crystallographic structures but it is not solvent accessible. Crucially it became solvent accessible in about one fifth of the conformers generated *in silico* from the wild type native structure 1qlp, indicating that transient solvent accessibility may be feasible in solution in the absence of mutations. Site I is therefore a potential ligand target site with some characteristics suggesting that ligand binding might induce local, stabilising conformational change. Support for this idea comes from mutagenesis studies that showed 13 mutations in the region of the I site (e.g. the space-filling mutation His269Tyr) increased stability, while preserving inhibitory function (Seo et al., 2000). Thymol and menthol are both small, hydrophobic natural products that showed high complementarity to the I site, and are considered

safe for use in the pharmaceutical industry. Following induced fit docking they achieved scores comparable to the score for the compound "CG", previously identified as an inhibitor of A1AT polymerization *in vitro* and in mammalian cells (Mallya et al., 2007). Some thymol derivatives achieved even better results, although the effect of their binding could be destabilizing, as suggested by preliminary ThermoFluor experiments.

The channel of interconnecting sites B, C and E is also potentially interesting as the druggability scores for these pockets are persistently high, and some of the best docking scores are results of docking ligands to these sites. However, the obvious caveat of docking to this site is that many of the ligands will interact with the RCL loop, thus potentially interfering with A1AT's antiprotease function. Indeed, mutation experiments have shown that the sequence between Arg196 to Glu279 can carry 9 mutations that increase the stability of the A1AT, but in several cases also decrease functional activity (Seo et al., 2002). Some of the other sites explored in the study may be more promising in terms of their position on the surface and lower likelihood of affecting inhibitory function.

There are obvious caveats in the approach presented here. The conformers generated using *CONCOORD* are artificially produced. Nevertheless they appear realistic when assessed for geometry and packing by the structure validation server *PROSESS* (Berjanskii et al., 2010). Moreover, the range of cavity characteristics observed was consistent with the variation observed between crystal structures. Although the conformational space of the protein is unlikely to be fully explored using *CONCOORD*, this technique did identify interesting pocket variations. The more recent program t*CONCOORD*, may further improve exploration of larger variations in molecular structure in future work (Eyrisch and Helms, 2009). The definition of pockets on the protein surface can vary significantly between programs, thus results presented here are specific to *SiteMap* predictions. Similarly the calculation of pocket volume and other properties are very much dependent on the definition of pocket boundaries, which varies widely across different software. Calculations of druggability have an empirical basis and are derived from previous correlations of scoring function predictions with *in vitro* observations of drug-like ligand binding. They do not guarantee *in vitro* binding affinity in a new system but

provide a reasonable starting point for docking studies *in silico* and *in vitro*. Finally, the docking calculations are subject to many approximations. They should therefore be considered as a screening tool, based upon goodness of fit of certain ligands against each site, to enrich true positive hits among the ligand rankings.

In summary, this promising strategy utilizes multiple protein conformer structures to identify both persistent and transiently druggable surface pockets. This approach was applied to A1AT, whose conformational flexibility suggests that the usual one conformer/one pocket approach to screening is likely to be inadequate. Pockets identified on the surface of A1AT show considerable variability across conformers. Moreover, a novel, transient pocket with druggability potential was identified (Patschull et al., 2012). Hits to this and other sites identified by this work compare favourably with a previously identified promising lead. An unusually high proportion of the limited set of *in silico* hits targeted at the I site and assayed by the ThermoFluor method alter the melting temperature of A1AT. These data are consistent with an *in vitro* interaction and warrant further experiments to pursue these ligands and I site targeting.

5'. **References**

AN, J., TOTROV, M. & ABAGYAN, R. 2005. Pocketome via comprehensive identification and classification of ligand binding envelopes. *Molecular and Cellular Proteomics,* 4**,** 752-761.

ASHFORD, P., MOSS, D., ALEX, A., YEAP, K., POVIA, A., NOBELI, I. & WILLIAMS, M. 2011. Visualisation of variable binding pockets on protein surfaces by probabilistic analysis of related structure sets.

BATTYE, T. G. G., KONTOGIANNIS, L., JOHNSON, O., POWELL, H. R. & LESLIE, A. G. W. 2011. iMOSFLM: a new graphical interface for diffraction-image processing with MOSFLM. *Acta Crystallography,* D67**,** 271–281.

BERJANSKII, M., LIANG, Y., ZHOU, J., TANG, P., STOTHARD, P., ZHOU, Y., CRUZ, J., MACDONELL, C., LIN, G., LU, P. & WISHART, D. S. 2010. PROSESS: a protein structure evaluation suite and server. *Nucleic Acids Research,* 38**,** W633-W640.

BOTTEGONI, G., ROCCHIA, W., RUEDA, M., ABAGYAN, R. & CAVALLI, A. 2011. Systematic Exploitation of Multiple Receptor Conformations for Virtual Ligand Screening. *PLoS ONE,* 6**,** e18845.

CHANG, Y.-P., MAHADEVA, R., PATSCHULL, A., NOBELI, I., EKEOWA, U., MCKAY, A., THALASSINOS, K., IRVING, J., HAQ, I., NYON, M., CHRISTODOULOU, J., ORDONEZ, A., MIRANDA, E. & GOOPTU, B. 2011. Targeting Serpins in High-Throughput and Structure-Based Drug Design. *Methods in Enzymology,* 501**,** IN PRESS.

CHANG, Z. & WOOLSEY, J. 2006. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Research,* 34**,** 668-672.

CHEN, Y. & SHOICHET, B. K. 2009. Molecular docking and ligand specificity in fragment-based inhibitor discovery. *Nature Chemical Biology,* 5**,** 358-364.

CHENG, A. C., COLEMAN, R. G., SMYTH, K. T., CAO, Q., SOULARD, P., CAFFREY, D. R., SALZBERG, A. C. & HUANG, E. S. 2007. Structure-based maximal affinity model predicts small-molecule druggability. *Nature Biotechnology,* 25**,** 71-75.

CIMMPERMAN, P., BARANAUSKIENE, L., JACHIMOVICIŪTE, S., JACHNO, J., TORRESAN, J., MICHAILOVIENE, V., MATULIENE, J., SEREIKAITE, J., BUMELIS, V. & MATULIS, D. 2008. A Quantitative Model of Thermal Stabilization and Destabilization of Proteins by Ligands. *Biophysical Journal,* 95**,** 3222-3231.

DAFFORN, T. R., MAHADEVA, R., ELLIOTT, P. R., SIVASOTHY, P. & LOMAS, D. A. 1999. A kinetic mechanism for the polymerisation of α1-antitrypsin. *Journal of Biological Chemistry,* 274**,** 9548–9555.

DAFFORN, T. R., PIKE, R. N. & BOTTOMLEY, S. P. 2004. Physical characterization of serpin conformations. *Methods,* 32**,** 150–158.

DE GROOT, B. L., VAN AALTEN, D. M. F., SCHEEK, R. M., AMADEI, A., VRIEND, G. & BERENDSEN, H. J. C. 1997. Prediction of protein conformational freedom from distance constraints. *Proteins: Structure, Function, and Genetics,* 29**,** 240-251.

DEMENTIEV, A., SIMONOVIC, M., VOLZ, K. & GETTINS, P. G. W. 2003. Canonical Inhibitor-Like Interactions Explain Reactivity of A1-Proteinase Inhibitor Pittsburgh and Antithrombin with Proteinases. *The Journal of Biological Chemistry,* 278**,** 37881-37887.

ELLIOTT, P. R., PEI, X. Y., DAFFORN, T. R. & LOMAS, D. A. 2000. Topography of a 2.0Å structure of a1-antitrypsin reveals targets for rational drug design to prevent conformational disease. *Protein Science,* 9**,** 1274-1281.

EMSLEY, P. & COWTAN, K. 2004. Coot: model-building tools for molecular graphic. *Acta Crystallography,* D60**,** 2126-2132.

EVANS, P. 2006. Scaling and assessment of data quality. *Acta Crystallography,* D62**,** 72–82.

EYRISCH, S. & HELMS, V. 2007. Transient Pockets on Protein Surfaces Involved in Protein–Protein Interaction. *Journal of Medicinal Chemistry,* 50**,** 3457-3464.

EYRISCH, S. & HELMS, V. 2009. What induces pocket openings on protein surface patches involved in protein-protein interactions? *Journal of Computeraided Molecular Design,* 23**,** 73-86.

FURNHAM, N., BLUNDELL, T. L., DEPRISTO, M. A. & TERWILLIGER, T. C. 2006. Is one solution good enough? *Nature Structural and Molecular Biology,* 13**,** 184-185.

GOOPTU, B. & LOMAS, D. A. 2008. Polymers and inflammation: disease mechanisms of the serpinopathies. *The Journal of Experimental Medicine,* 205**,** 1529-1534.

GOOPTU, B. & LOMAS, D. A. 2009. Conformational pathology of the serpins: themes, variations and therapeutic strategies. *Annual Review of Biochemistry,* 78**,** 147-176.

GOOPTU, B., MIRANDA, E., NOBELI, I., MALLYA, M., PURKISS, A., LEIGH BROWN, S. C., SUMMERS, C., PHILLIPS, R. L., LOMAS, D. A. & BARRETT, T. E. 2009. Crystallographic and cellular characterisation of two mechanisms stablising the native fold of a_1-antitrypsin : implications for disease and drug design. *The Journal of Biological Chemistry,* 387**,** 857-868.

HAJDUK, P. J., HUTH, J. R. & FESIK, S. W. 2005. Druggability Indices for Protein Targets Derived from NMR-Based Screening Data. *Journal of Medicinal Chemistry,* 48**,** 2518-2535.

HALGREN, T. A. 2009. Identifying and characterizing binding sites and assessing druggability. *Journal of Chemical Information and Modeling,* 49**,** 377-389.

HUANG, C. & JACOBSON, K. 2010. Detection of protein-protein interactions using nonimmune IgG and BirA-mediated biotinylation. *Biotechniques,* 49**,** 881-886.

HUNTINGTON, J. A., READ, R. J. & CARRELL, R. W. 2000. Structure of a serpin-protease complex shows inhibition by deformation. *Nature,* 407**,** 923-926.

IM, H., SEO, E. J. & YU, M.-H. 1999. Metastability in the Inhibitory Mechanism of Human A1-Antitrypsin. *The Journal of Biological Chemistry,* 274**,** 11072-11077.

IM, H., WOO, M.-S., HWANG, K. Y. & YU, M.-H. 2002. Interactions causing the kinetic trap in serpin protein folding. *The Journal of Biological Chemistry,* 277**,** 46347-46354.

IVETAC, A. & MCCAMMON, J. A. 2012. A molecular dynamics ensemble-based approach for the mapping of druggable binding sites. *Methods in Molecular Biology,* 819**,** 3-12.

KIM, S.-J., WOO, J.-R., SEO, E. J., YU, M.-H. & RYU, S.-E. 2001. A 2.1Å resolution structure of an uncleaved a_1-antitrypsin shows variability of the reactive centre and other loops. *Journal of Molecular Biology,* 306**,** 109-119.

KUMAR, S., MA, B., TSAI, C. J., SINHA, N. & NUSSINOV, R. 2000. Folding and binding cascades: dynamic landscapes and population shifts. *Protein Science,* 9**,** 10-19.

LASKOWSKI, R. A. 2009. PDBsum new things. *Nucleic Acids Research,* 37**,** D355-359.

LASKOWSKI, R. A., MACARTHUR, M. W., MOSS, D. S. & THORNTON, J. M. 1994. PROCHECK: a program to check the stereo chemical quality of protein structures. *Journal of Applied Crystallography,* 26**,** 283-291.

MA, B., SHATSKY, M., WOLFSON, H. J. & NUSSINOV, R. 2002. Multiple diverse ligands binding at a single protein site: a matter of pre-existing populations. *Protein Science,* 11**,** 184-197.

MALLYA, M., PHILLIPS, R., L., SALDANHA, S. A., GOOPTU, B., LEIGH BROWN, S. C., TERMINE, D. J., SHIRVANI, A. M., WU, Y., SIFERS, R. N., ABAGYAN, R. & LOMAS, D. A. 2007. Small molecules block the polymerization of Z alpha1-antitrypsin and increase the clearance of intracellular aggregates. *Journal of Medicinal Chemistry,* 50**,** 5357-5363.

MCCOY, A. J., GROSSE-KUNSTLEVE, R. W., ADAMS, P. D., WINN, M. D., STORONI, C., L. & READ, R. J. 2007. Phaser crystallographic software. *Journal of Applied Crystallography,* 40**,** 658–674.

MURSHUDOV, G. N., VAGIN, A. A. & DODSON, E. J. 1997. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallography,* D53**,** 240-255.

NAYAL, M. & HONIG, B. 2006. On the nature of cavities on protein surfaces: application to the identification of drug-binding sites. *Proteins: Structure, Function, and Bioinformatics,* 63**,** 892-906.

NICHOLS, S. E., BARON, R. & MCCAMMON, J. A. 2012. On the use of molecular dynamics receptor conformations for virtual screening. *Methods in Molecular Biology,* 819**,** 93-103.

PARFREY, H., MAHADEVA, R., RAVENHILL, N., ZHOU, A., DAFFORN, T. R., FOREMAN, R. C. & LOMAS, D. A. 2003. Targeting a surface cavity of a_1-antitrypsin to prevent conformational disease. *The Journal of Biological Chemistry,* 278**,** 33060-33066.

PATSCHULL, A. O. M., GOOPTU, B., ASHFORD, P., DAVITER, T. & NOBELI, I. 2012. In Silico Assessment of Potential Druggable Pockets on the Surface of a1-Antitrypsin Conformers. *PLoS ONE,* 7**,** 1-15.

PATSCHULL, A. O. M., SEGU, L., NYON, M. P., LOMAS, D. A., NOBELI, I., BARRETT, T. E. & GOOPTU, B. 2011. Therapeutic target-site variability in α1-antitrypsin characterized at high resolution. *Acta Crystallographica Section F: Structural Biology and Crystallization Communications,* D67**,** 1492-1497.

PEARCE, M. C., MORTON, C. J., FEIL, S. C., HANSEN, G., ADAMS, J. J., PARKER, M. W. & BOTTOMLEY, S. P. 2008. Preventing serpin aggregation: the molecular mechanism of citrate action upon antitrypsin unfolding. *Protein Science,* 17**,** 2127-2133.

PEROT, S., SPERANDIO, O., MITEVAL, M. A., CAMPROUX, A.-C. & VILLOUTREIX, B. O. 2010. Druggable pockets and binding site centric

chemical space: a paradigm shift in drug discovery. *Drug Discovery Today,* 15**,** 656-667.

RYU, S.-E., CHOI, H.-J., KWON, K.-S., LEE, K. N. & YU, M.-H. 1996. The Native Strains in the Hydrophobic Core and Flexible Reactive Loop of a Serine Protease Inhibitor: Crystal structure of an Uncleaved A1-Antitrypsin at 2.7Å. *Structure***,** 1181-1192.

SCHMIDTKE, P. & BARRIL, X. 2010. Understanding and predicting druggability. A high-throughput method for detection of drug binding sites. *Journal of Medicinal Chemistry,* 53**,** 5858-5867.

SCHMIDTKE, P., LE GUILLOUX, V., MAUPETIT, J. & TUFFERY, P. 2010. fpocket: online tools for protein ensemble pocket detection and tracking. *Nucleic Acids Research,* 38**,** W582-W589.

SEO, E. J., IM, H., MAENG, J. S., KIM, K. E. & YU, M. H. 2000. Distribution of the Native Strain in Human a1-Antitrypsin and Its Association with Protease Inhibitor Function. *The Journal of Biological Chemistry,* 275**,** 16904-16909.

SEO, E. J., LEE, C. & YU, M.-H. 2002. Concerted Regulation of Inhibitory Activity of a1-Antitrypsin by the Native Strain Distributed throughout the Molecule. *The Journal of Biological Chemistry,* 277**,** 14216-14220.

SHERIDAN, R. P., MAIOROV, V. N., HOLLOWAY, M. K., CORNELL, W. D. & GAO, Y. D. 2010. Drug-like density: a method of quantifying the "bindability" of a protein target based on a very large set of pockets and drug-like ligands from the Protein Data Bank. *Journal of Chemical Information and Modeling,* 50**,** 2029-2040.

SHERMAN, W., DAY, T., JACOBSON, M. P., FRIESNER, R. A. & FARID, R. 2006. Novel Procedure for Modeling Ligand/Receptor Induced Fit Effects. *Journal of Medicinal Chemistry,* 49**,** 534-554.

SILVERMAN, G. A., BIRD, P. I., CARRELL, R. W., CHURCH, F. C., COUGHLIN, P. B., GETTINS, P., IRVING, J., LOMAS, D. A., MOYER, R. W., PEMBERTON, P., REMOLD O'DONNELL, E., G., S., TRAVIS, J. & WHISSTOCK, J. 2001. The serpins are an expanding superfamily of structurally similar but functionally diverse proteins. Evolution, novel functions, mechanism of inhibition and a revised nomenclature. *The Journal of Biological Chemistry,* 276**,** 33293-33296.

SINGH, P., KHAN, M. S., NASEEM, A. & JAIRAJPURI, M. A. 2011. Analysis of surface cavity in serpin family reveals potential binding sites for chemical chaperone to reduce polymerization. *Journal of Molecular Modeling***,** IN PRESS.

STONE, S. R. & HOFSTEENGE, J. 1986. Kinetics of the inhibition of thrombin by hirudin. *Biochemistry,* 25**,** 4622–4628.

TEW, D. J. & BOTTOMLEY, S. P. 2001. Probing the equilibrium denaturation of the serpin a1- antitrypsin with single tryptophan mutants; evidence for structure in the urea unfolded state. *Journal of Molecular Biology,* 313**,** 1161-1169.

VINCENT, F., SPINELLI, S., RAMONI, R., GROLLI, S., PELOSI, P., CAMBILLAU, C. & TEGONI, M. 2000. Complexes of porcine odorant-binding protein with odorant molecules belonging to different chemical classes. *Journal of Molecular Biology,* 300**,** 127-139.

WANG, R., FANG, X., LU, Y., YANG, C.-Y. & WANG, S. 2005. The PDBbind Database: Methodologies and Updates. *Journal of Medicinal Chemistry,* 48**,** 4111-4119.

WEBER, A., HALGREN, T. A., DOYLE, J. J., LYNCH, R. J., SIEGL, P. K. S., PARSONS, W. H., GREENLEE, W. J. & PATCHETT, A. A. 1991. Design and Synthesis of P2-P1'-Linked Macrocyclic Human Renin Inhibitors. *Journal of Medicinal Chemistry,* 34**,** 2692-2701.

WELLS, J. A. & MCCLENDON, C. L. 2007. Reaching for high-hanging fruit in drug discovery at protein-protein interfaces. *Nature,* 450**,** 1001-1009.

WISHART, D. S., KNOX, C., GUO, A. C., CHENG, D., SHRIVASTAVA, S., TZUR, D., GAUTAM, B. & M., H. 2008. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Research,* 36**,** D901-906.

YAMASAKI, M., SENDALL, T. J., HARRIS, L. E., LEWIS, G. M. & HUNTINGTON, J. A. 2010. Loop-sheet mechanism of serpin polymerization tested by reactive centre loop mutations. *Journal of Biological Chemistry,* 285**,** 30752-30758.

6'. **'Appendix**

A'. – ConCoord 2.0 Script

```
#!/bin/tcsh -f
#$ -cwd
#$ -j y
#$ -S /bin/tcsh
#$ -V
setenv DATA /d/d640/u/ubcgw9ap/1qlp_concoord_run/pdb
setenv PDB 1qlpHydSuperposed
setenv DSSP_EXE /d/d640/u/ubcgw9ap/dsspcmbi

unlimit
limit coredumpsize 0

source /d/d610/s/concoord/concoord.csh
cd /d/d640/u/ubcgw9ap/1qlp_concoord_run/

if (-e ATOMS.DAT) then
  rm -f ATOMS.DAT
endif
if (-e BONDS.DAT) then
  rm -f BONDS.DAT
endif
if (-e MARGINS.DAT) then
  rm -f MARGINS.DAT
endif

/bin/rm disco*
/bin/rm dist*

dist -dssp $DSSP_EXE -r -p ${DATA}/${PDB}.pdb <</eoi
2
2
/eoi

echo "DIST DONE"

disco -ox ./disco.${PDB}.xtc -op ./disco.${PDB} -n 100 -i 1000 -viol 3. -bump -t
1000.0
```

B' & C'.- C-Shell and Perl Script

C-Shell Script executing the Perl Script:

```
#!/bin/csh
```

```
echo "site_Alpha\t Protein_Name\t site_Number\t Balance\t Contact\t Don_Acc\t
Dscore\t Enclosure\t Entry_Name\t Exposure\t Hydrophilic\t Hydrophobic\t
site_score\t Size\t Volume\t Coordinate_X\t Coordinate_Y\t Coordinate_Z\t
Distance_A\t Distance_B\t Distance_C\t Distance_D\t Distance_E\t Distance_F\t
Distance_G\t Distance_H \t Distance_LL \n"> 1qlpConCoo_SM_score.txt

foreach file (rot*1qlp*.mae)

./SiteMapData1qlp.pl $file >> 1qlpConCoo_SM_score.txt

end
```

Executed Perl Script:

```
#!/usr/bin/perl -w

use strict;

die "Usage: script.pl maestro_file_with_sites\n" unless (1<=scalar(@ARGV));
my $maefile = $ARGV[0];
open(MAEFILE, $maefile) || die "Could not open file $maefile\n";

my @array;

$array[0] = "Entry_name";
$array[1] = "sitescore";
$array[2] = "Size";
$array[3] = "Dscore";
$array[4] = "Volume";
$array[5] = "Exposure";
$array[6] = "Enclosure";
$array[7] = "Contact";
$array[8] = "Phobic";
$array[9] = "Philic";
$array[10] = "Balance";
$array[11] = "DonAcc";

my $protein;
my $site;
my $x;
my $y;
my $z;
my $sumx;
my $sumy;
my $sumz;
my $Cx;
my $Cy;
my $Cz;

#my $sNa;
```

```
#my $sNb;
#my $sNc;
#my $sNd;
#my $sNe;
#my $sNf;
#my $sNg;
#my $sNh;


while (my $line = <MAEFILE>) {
  if ($line =~ /^\s+(.+?)_site_(\d+)/){
    $protein = $1;
    $site = $2;
    $Cx=0.0;
    $Cy=0.0;
    $Cz=0.0;
    $sumx=0.0;
    $sumy=0.0;
    $sumz=0.0;

    my %hash=();
    for (my $i = 0; $i < 12; $i++) {
      $line = <MAEFILE>;
      chomp($line);
      $line =~ /\s+(\d+)/g;
      $hash{ $array[$i] } = $line;
    }

    do {
      $line = <MAEFILE>;
    } until ($line =~ /\s+\:\:\:/);


    for (my $i = 0; $i < $hash{ $array[2] }; $i++) {
        $line = <MAEFILE>;
        chomp($line);

      $line =~ /^\s+\d+\s+\d+\s+(\S+)\s+(\S+)\s+(\S+)/;
          $x = $1;
      $y = $2;
          $z = $3;


          $sumx += $x;
          $sumy += $y;
      $sumz += $z;

     }
     $Cx = $sumx/$hash{ $array[2] };
     $Cy = $sumy/$hash{ $array[2] };
```

```
$Cz = $sumz/$hash{ $array[2] };


my $da= sqrt(((23.6945454545455-$Cx)*(23.6945454545455-
$Cx))+((7.47181818181818-$Cy)*(7.47181818181818-
$Cy))+((22.8463636363637-$Cz)*(22.8463636363637-$Cz)));
my $db= sqrt(((6.10666666666666-$Cx)*(6.10666666666666-
$Cx))+((13.9974074074074-$Cy)*(13.9974074074074-
$Cy))+((41.3433333333334-$Cz)*(41.3433333333334-$Cz)));
my $dd= sqrt(((-5.39870967741936-$Cx)*(-5.39870967741936-$Cx))+((-
12.1035483870968-$Cy)*(-12.1035483870968-$Cy))+((29.3809677419355-
$Cz)*(29.3809677419355-$Cz)));
my $de= sqrt(((3.29294117647059-$Cx)*(3.29294117647059-
$Cx))+((13.825294117647-$Cy)*(13.825294117647-$Cy))+((23.9805882352941-
$Cz)*(23.9805882352941-$Cz)));
my $dh= sqrt(((3.54344827586207-$Cx)*(3.54344827586207-$Cx))+((-
12.7893103448276-$Cy)*(-12.7893103448276-$Cy))+((16.1479310344828-
$Cz)*(16.1479310344828-$Cz)));
my $df= sqrt(((13.2971428571429-$Cx)*(13.2971428571429-$Cx))+((-
12.5171428571429-$Cy)*(-12.5171428571429-$Cy))+((25.8314285714286-
$Cz)*(25.8314285714286-$Cz)));
my $dc= sqrt(((1.26758620689655-$Cx)*(1.26758620689655-
$Cx))+((11.8313793103448-$Cy)*(11.8313793103448-
$Cy))+((33.6651724137931-$Cz)*(33.6651724137931-$Cz)));
my $dg= sqrt(((36.8685714285714-$Cx)*(36.8685714285714-$Cx))+((-
9.73142857142857-$Cy)*(-9.73142857142857-$Cy))+((4.97428571428572-
$Cz)*(4.97428571428572-$Cz)));
my $dll= sqrt(((6.01853997-$Cx)*(6.01853997-$Cx))+((5.30592851-
$Cy)*(5.30592851-$Cy))+((16.3633573-$Cz)*(16.3633573-$Cz)));

my $sN= "NA";
if ($da<=3.75) {
  $sN= "A";
}

if ($db<=3.75) {
  $sN="B";
}

if ($dc<=3.75) {
  $sN="C";
}

if ($dd<=3.75) {
  $sN="D";
}

if ($de<=3.75) {
  $sN="E";
}
```

```
    if ($df<=3.75) {
      $sN="F";
    }

    if ($dg<=3.75) {
      $sN="G";
    }

    if ($dh<=3.75) {
      $sN="H";
    }

    if ($dll<=3.75) {
      $sN="LL";
    }

    if ($da>=10 and $db>=10 and $dc>=10 and $dd>=10 and $de>=10 and
$df>=10 and $dg>=10 and $dh>=10 and $dll>=10) {
      $sN="NEW";
    }

     if ($da>=20 and $db>=20 and $dc>=20 and $dd>=20 and $de>=20 and
$df>=20 and $dg>=20 and $dh>=20 and $dll>=20) {
      $sN="!NEW!";
    }

    print "$sN\t$protein\t$site\t";
    foreach my $key (sort keys %hash) {
      print "$hash{$key}\t";
    }
    print "$Cx\t$Cy\t$Cz\t$da\t$db\t$dc\t$dd\t$de\t$df\t$dg\t$dh\t$dll\n";
  }
}
```

D'. Statistics on Site Frequency in eight PDB Crystal Structures of α_1-Antitrypsin

Table: Description of Sites and Frequency of Occurrence in Eight α_1-Antitrypsin Structures

Site	Number of Sites[§]	PDB Code where Sites are present	Max. Site Dscore	Max. Site-Score	Min. Site-Score	Mean Site-Score (±SD)	Max. Site Volume (in $Å^3$)	Min. Site Volume (in $Å^3$)	Mean Site Volume (±SD) (in $Å^3$)
A[(S)]	6	1qlp, 2qug,	1.031[a]	1.029[a]	0.677[c]	0.876 (±0.12	234[a]	98[c]	164.583

		1oph, 1hp7, 3drm,3cwm.				6)			(±41.55)
B(S)	6	1qlp, 2qug, 1oph, 1hp7, 3drm,3cwm.	0.968 d	0.982 d	0.567 e	0.830 (±0.130)	326 d	76 e	210.659 (±75.089)
C	5	1qlp, 1oph, 3drm,3cwm, 1ezx.	0.989 c	1.034 c	0.824 a	0.972 (±0.078)	173 c	59 a	122.177 (±43.47)
D	6	1qlp, 2qug, 1hp7, 3drm, 3cwm, 1ezx.	0.907 f	0.883 f	0.727 h	0.795 (±0.054)	205 f	136 a,e	174.518 (±26.179)
E	4	1qlp, 2qug, 1oph, 1ezx.	1.014 d	1.020 d	0.687 h	0.831 (±0.141)	273 d	60 h	133.599 (±83.055)
F	5	1qlp, 1hp7, 3cwm, 1iz2, 1ezx.	1.005 g	0.984 g	0.586 h	0.693 (±0.147)	249 g	64 h	125.812 (±63.561)
G(S)	5	1qlp, 2qug, 1hp7, 3drm, 3cwm.	0.602 b	0.632 b	0.511 d	0.588 (±0.042)	83 b	42 d	74.088 (±16.355)
H(S)	1	1qlp.	0.482 a	0.585 a	-	-	46 a	-	-
I	2	1oph, 1iz2 , (1hp7)*.	0.917 c	0.944 c	0.848 g	0.896 (±0.048)	90 g	45 c	67.228 (±22.295)

Key:

** this site was cropped to another site and hence not included in the count, but the cavity is solvent accessible in this structure.*

[§] number of structures out of 8, in which sites were identified by SiteMap2.3. (S) is present in the stressed form, but absent in the relaxed form.

[a-h] score obtained from the following structure: [a] 1qlp, [b] 2qug, [c] 1oph, [d] 1hp7 ,[e] 3drm, [f] 3cwm, [g] 1iz2 or [h] 1ezx.

E' – Statistics on Site Frequency in 1qlp ConCoord Conformers of α_1-Antitrypsin

Table: Description of Sites and Frequency of Occurrence in New native WT (1qlp) Conformations

Site	Number of Sites[§]	Max Dscore	Max. Site-Score	Min. Site-Score	Mean Site-Score (±SD)	Max. Site Volume (in $Å^3$)	Min. Site Volume (in $Å^3$)	Mean site Volume (±SD) (in $Å^3$)
A	96	1.072	1.082	0.9425	1.002 (±0.022)	435.61	173.21	298.038 (±67.478)
B	59	1.057	1.010	0.610	0.876 (±0.096)	450.702	80.605	215.648 (±83.215)
C	33	0.977	1.083	0.824	0.956 (±0.081)	518.959	51.45	151.003 (±117.222)
D	92	0.983	0.980	0.601	0.820 (±0.089)	228.438	86.436	159.394 (±34.637)
E	40	1.009	1.083	0.663	0.931 (±0.145)	313.845	77.861	189.173 (±81.772)
F	91	0.767	0.800	0.592	0.683 (±0.047)	164.64	69.972	119.270 (±17.086)
G	58	0.781	0.751	0.501	0.636 (±0.050)	128.282	54.88	87.879 (±13.083)

H	69	0.714	0.744	0.528	0.602 (±0.044)	89.18	41.503	59.076 (±9.878)
I	21	0.937	0.953	0.726	0.871 (±0.059)	96.726	37.387	52.428 (±11.026)

Key:

[§] *number of conformers out of 100, in which sites were identified by SiteMap2.3.*

F'.– Statistics on Site Frequency in 2qug ConCoord Conformers of α_1-Antitrypsin

Table: 2qug Description of Sites and Frequency of Occurrence in New Conformations

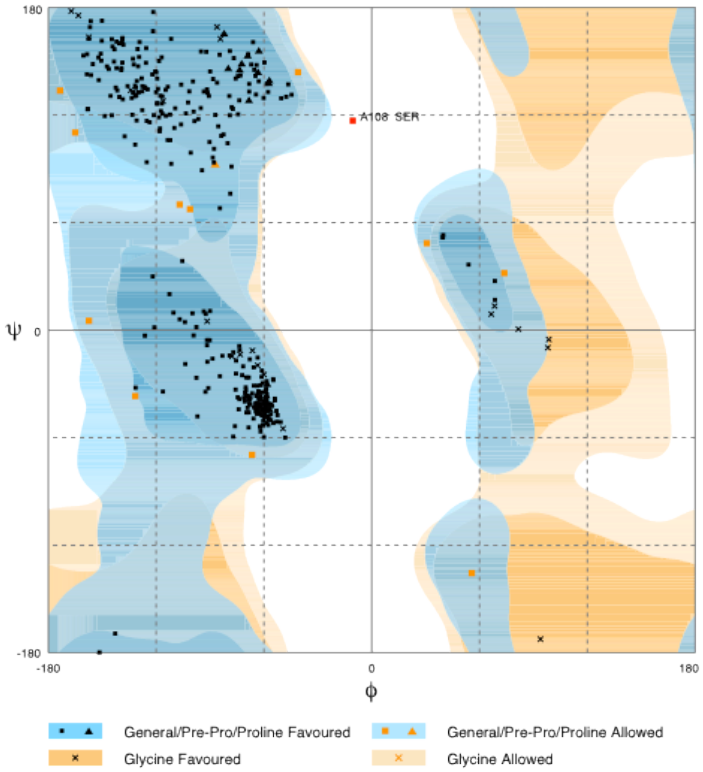
Site	**Number of Sites[§]**	**Max D-score**	**Max. SiteScore**	**Min. SiteScore**	**Mean SiteScore (±SD)**	**Max. Site Volume (in Å^3)**	**Min. Site Volume (in Å^3)**	**Mean Site Volume (±SD) (in Å^3)**
A	97	1.099	1.078	0.895	0.987 (±0.039)	425.32	203.515	203.515 (±68.430)
B	27	0.994	0.972	0.655	0.816 (±0.089)	322.42	141.316	205.470 (±51.183)
C	31	0.912	1.048	0.814	0.922 (±0.099)	191.737	43.218	94.668 (±55.634)
D	53	0.975	0.967	0.630	0.777 (±0.094)	363.58	87.122	161.890 (±62.046)
E	22	0.887	1.042	0.754	0.890 (±0.072)	173.215	78.204	114.422 (±17.684)
F	12	0.736	0.747	0.552	0.616 (±0.052)	140.287	72.373	85.979 (±22.398)
G	38	0.713	0.719	0.513	0.612 (±0.043)	142.345	55.223	85.208

								(±15.567)
H	35	0.691	0.746	0.535	0.647 (±0.045)	116.963	69.629	91.267 (±11.006)
I	14	0.903	0.930	0.786	0.858 (±0.043)	62.083	32.928	45.08 (±8.073)

Key:

§ *number of conformers out of 100, in which sites were identified by SiteMap2.3.*

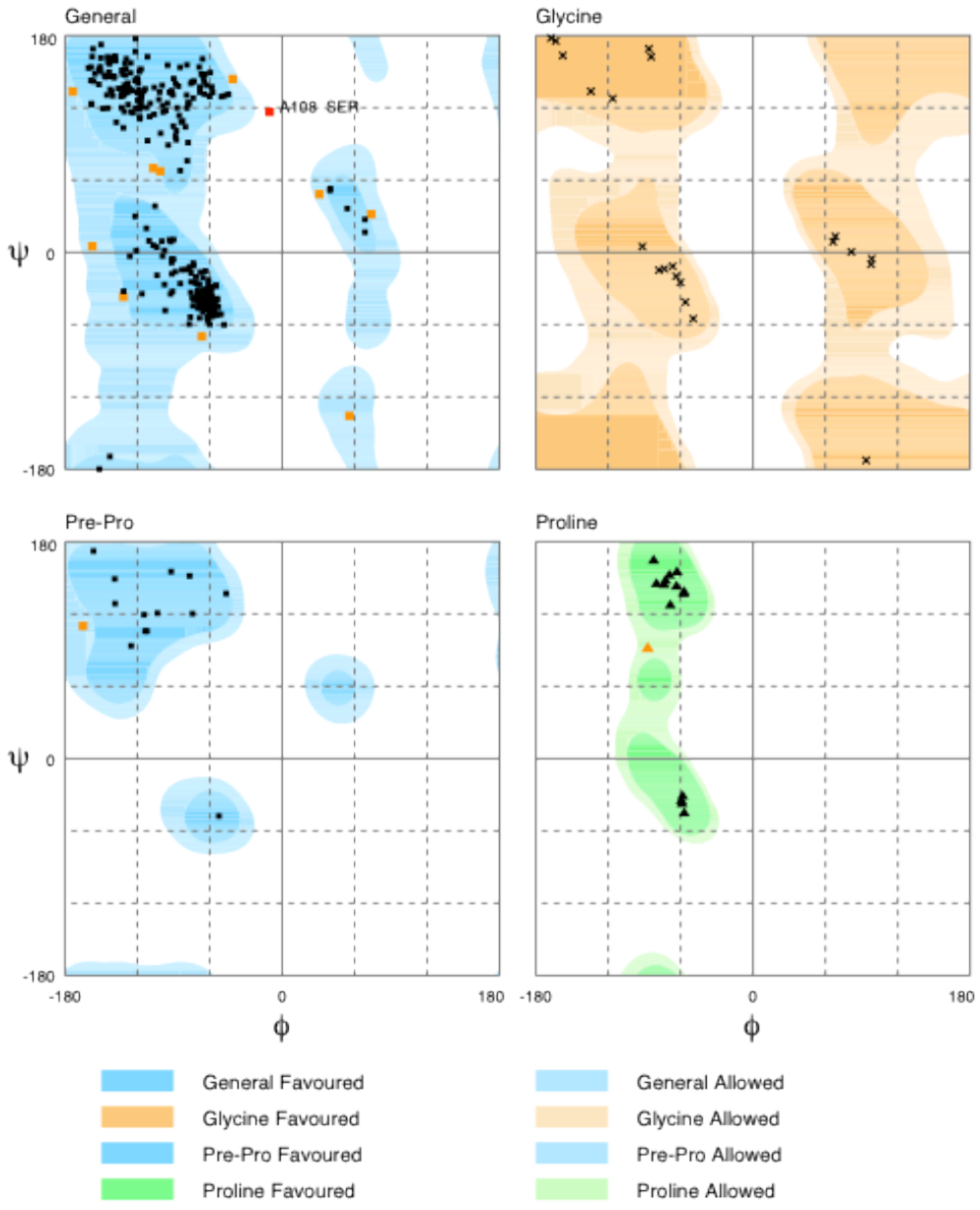
G'. – Procheck Results for 3ne4



Number of residues in favoured region (~98.0% expected) : 355 (96.5%)
Number of residues in allowed region (~2.0% expected) : 12 (3.3%)
Number of residues in outlier region : 1 (0.3%)

RAMPAGE by Paul de Bakker and Simon Lovell available at http://www-cryst.bioc.cam.ac.uk/rampage/

Please cite: S.C. Lovell, I.W. Davis, W.B. Arendall III, P.I.W. de Bakker, J.M. Word, M.G. Prisant, J.S. Richardson & D.C. Richardson (2002) Structure validation by Cα geometry: ϕ,ψ and Cβ deviation. *Proteins: Structure, Function & Genetics*. **50**: 437-450



Number of residues in favoured region (~98.0% expected) : 355 (96.5%)
Number of residues in allowed region (~2.0% expected) : 12 (3.3%)
Number of residues in outlier region : 1 (0.3%)

RAMPAGE by Paul de Bakker and Simon Lovell available at http://www-cryst.bioc.cam.ac.uk/rampage/

Please cite: S.C. Lovell, I.W. Davis, W.B. Arendall III, P.I.W. de Bakker, J.M. Word, M.G. Prisant, J.S. Richardson & D.C. Richardson (2002) Structure validation by Cα geometry: φ/ψ and Cβ deviation. *Proteins: Structure, Function & Genetics*, **50**: 437-450

"Judge not, that you be not judged" Bible, Matthew, ch.7

Structure Factor Check

3NE4

Title: 1.8 ANGSTROM STRUCTURE OF INTACT NATIVE WILD-TYPE ALPHA-1-AN
Date: 08-JUN-10
PDB code: 3NE4

Crystal

Cell parameters:

a: 114.38 A	b: 38.94 A	c: 88.83 A
α: 90.00	β: 104.29	γ: 90.00

Space group: C 1 2 1

Model

3167 atoms (217 water molecules)

Number of chains:	2
Volume not occupied by model:	30.1 %
<B> (for atomic model):	25.9 A^2
σ(B):	10.91 A^2
Matthews coefficient:	2.16
Corresponding solvent % :	42.66

Refinement

Program:	REFMAC 5.6.0077
Nominal resolution range:	86.1 – 1.81 A
Reported R-factor:	0.187
Number of reflections used:	32406
Reported Rfree:	0.23
Sigma cut-off:	N.A.

Structure Factors

Input

Nominal resolution range:		42.1 – 1.81 A
Reflections in file:		34123
Unique reflections	above 0:	34123
	above 1 σ:	34084
	above 3 σ:	26078

SFCHECK

Nominal resolution range: 42.1 – 1.81 A
(max. from input data, min. from author)

Used reflections:	34123
Completeness:	97.8 %
R_stand(F) = <σ(F)>/<F> :	0.056

Anisotropic distribution of Structure Factors
ratio of eigen values: 0.9711 1.0000 0.9347

B_overall (by Patterson):	26.A^2
Optical resolution:	1.48 A
Expected opt. resol. for complete data set:	1.48 A
Estimated minimal error:	0.049 A

Model vs. Structure Factors

R-factor for all reflections:	0.202
Correlation factor:	0.942
R-factor:	0.206
for F > 2.0 σ	
nom. resolution range:	86.08 – 1.81A
reflections used:	34084
Rfree:	0.249
Nfree:	1713
R-factor without free-refl.:	0.204
Non free-reflections:	32371
<u> (error in coords by Luzzati plot):	0.197 A
Estimated maximal error:	0.207 A
DPI:	0.130 A

Scaling

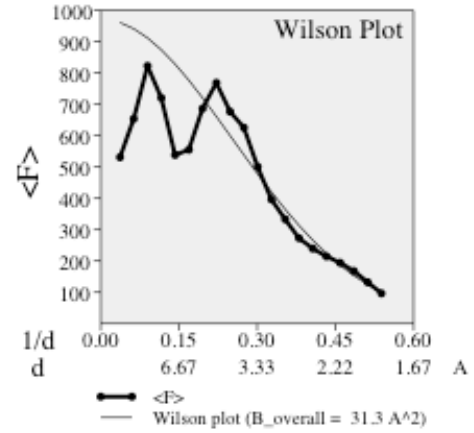
Scale:	0.930
Bdiff:	-2.78

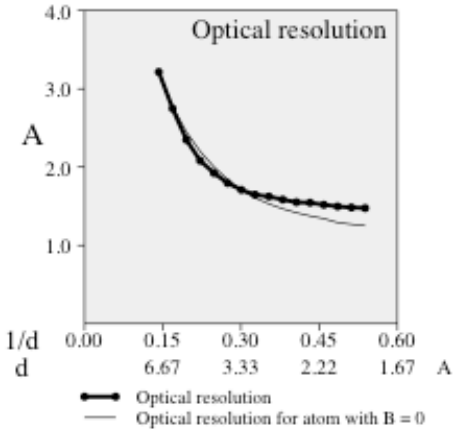
Anisothermal Scaling (Beta):
0.9655 0.9756 0.8096 0.0000 0.3963 0.0000

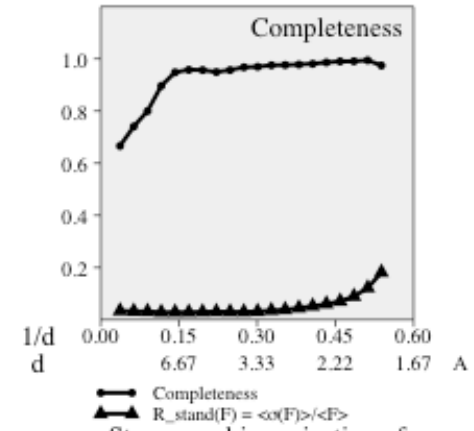
Solvent correction – Ks,Bs: 0.900 250.010

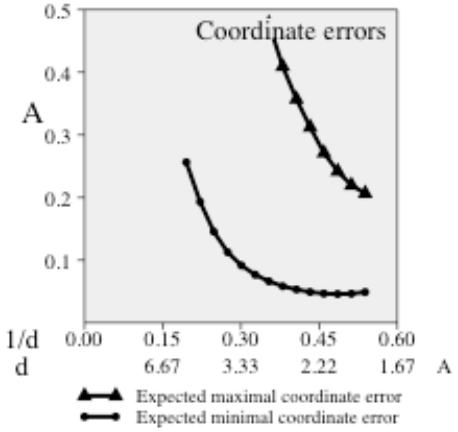
Structure Factor Check

3NE4

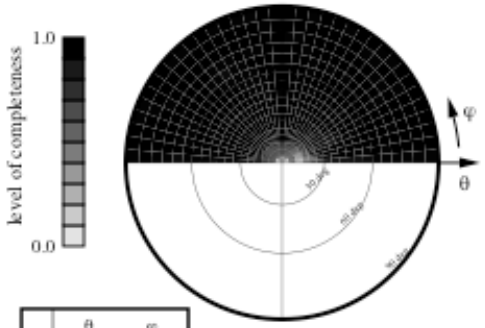






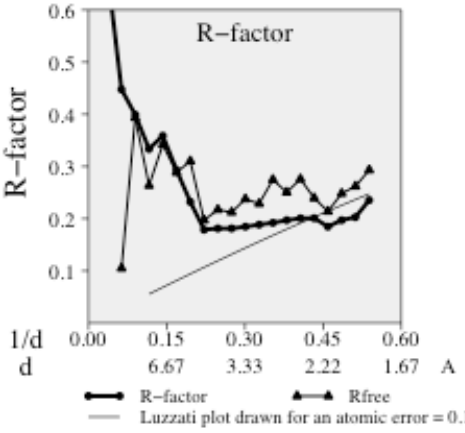


Stereographic projection of
the averaged radial completeness



	θ	φ
h	75.71	0.00
k	90.00	90.00
l	0.00	0.00

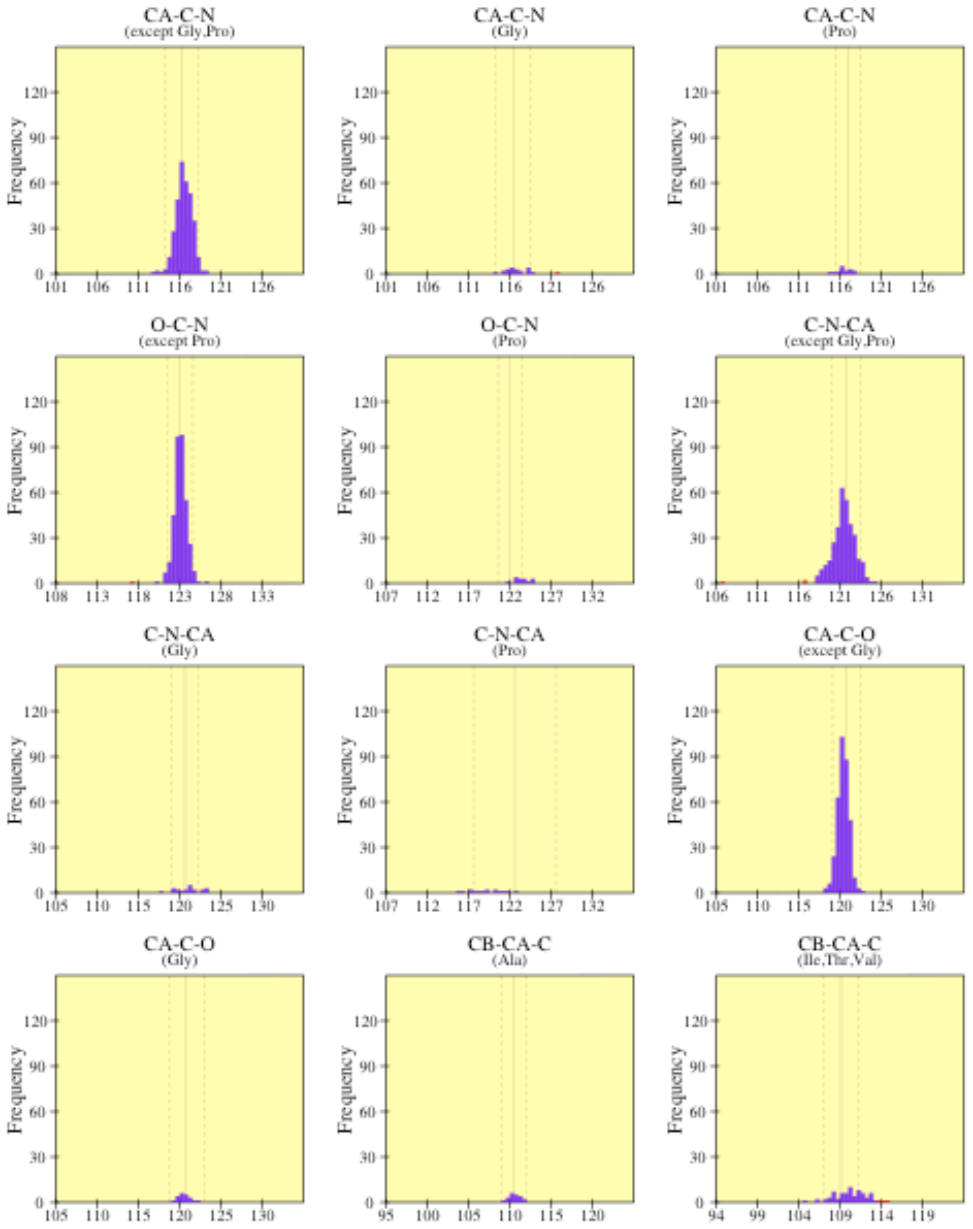
polar coordinates of the
crystallographical axes



PROCHECK

Main-chain bond angles

rcsb059714_refmac2-coot-1

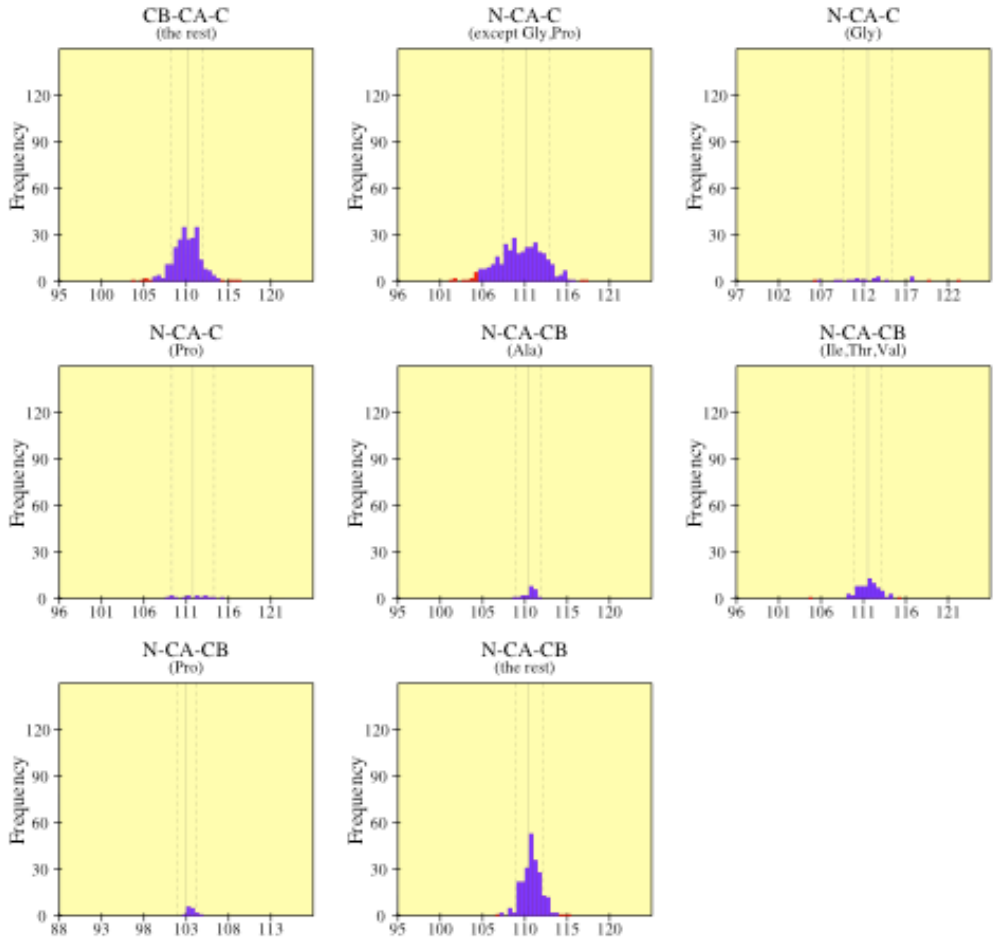


Black bars > 2.0 st. devs. from mean.

Solid and dashed lines represent the mean and standard deviation values as per Engh & Huber small-molecule data.

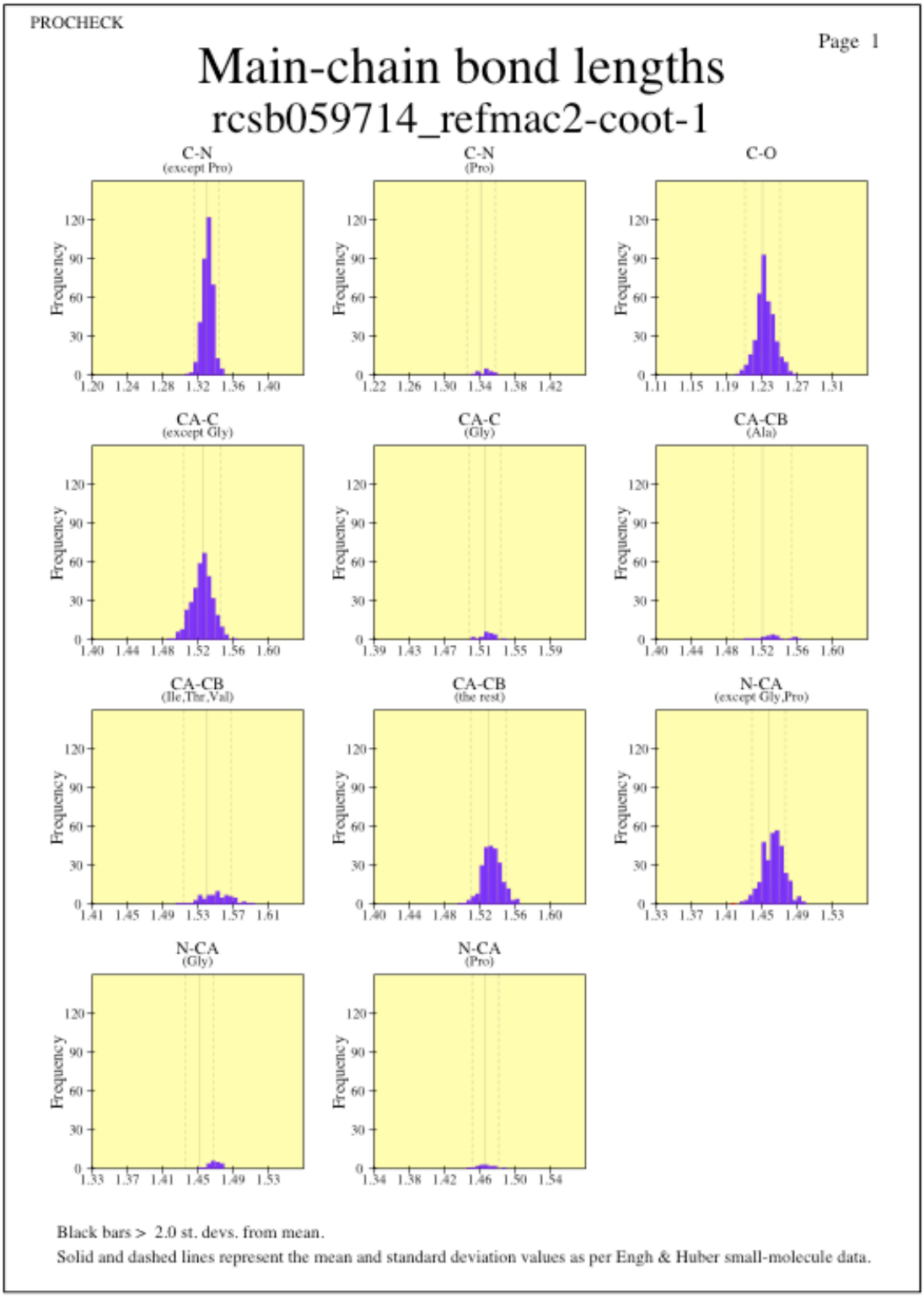
Main-chain bond angles

rcsb059714_refmac2-coot-1

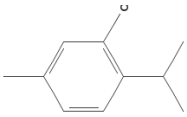
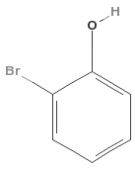
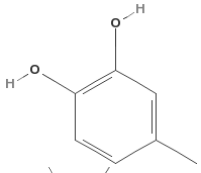
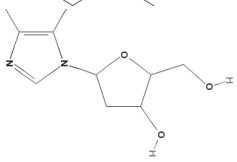
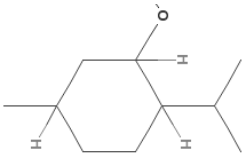
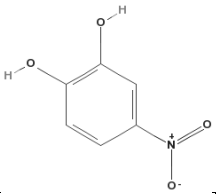
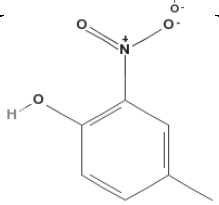
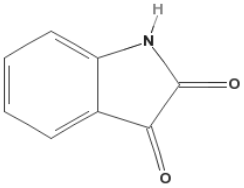
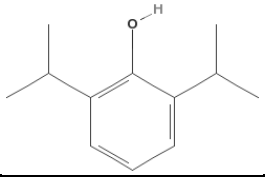


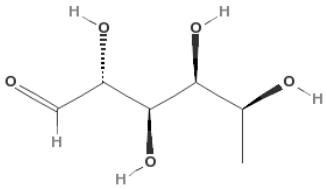
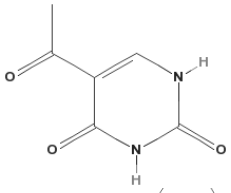
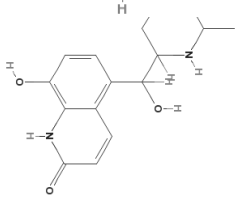
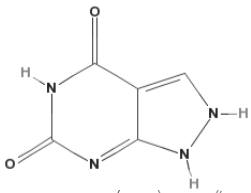
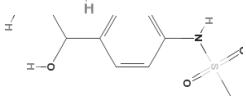
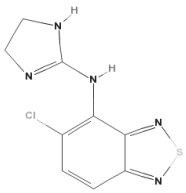
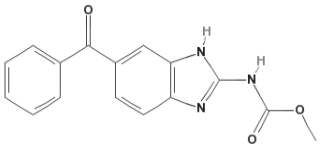
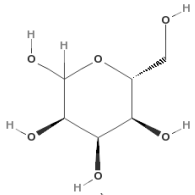
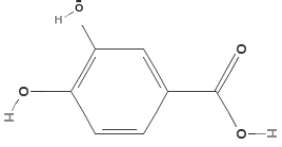
Black bars > 2.0 st. devs. from mean.

Solid and dashed lines represent the mean and standard deviation values as per Engh & Huber small-molecule data.



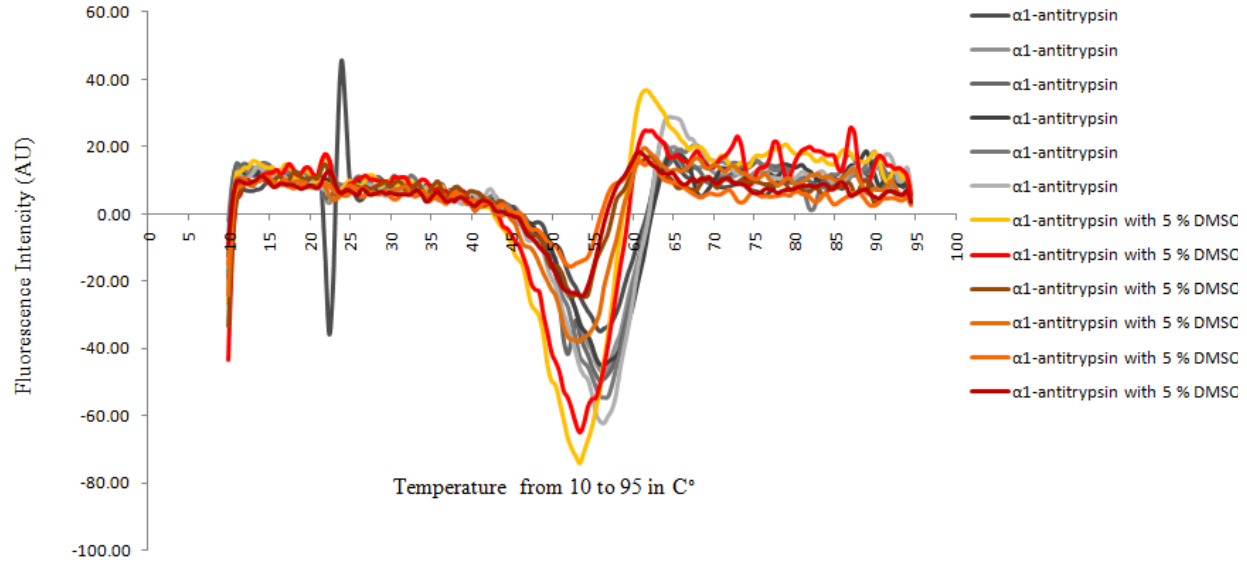
H'.– 18 Compounds from DrugBank Library 'Hits' used for the ThermoFluor Assay

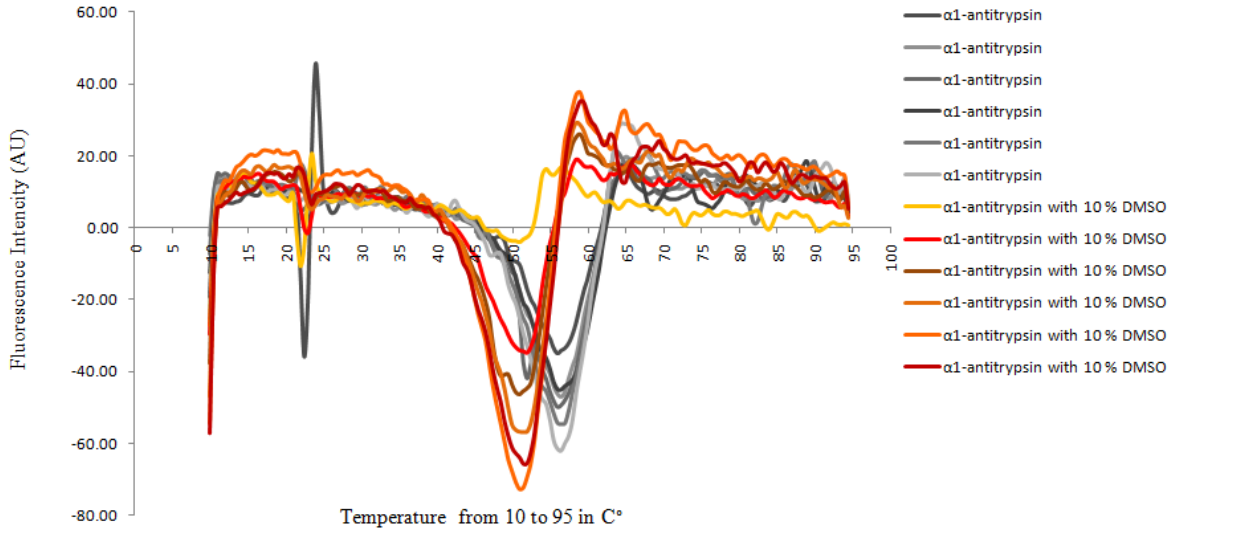
Compound	Structure	Molecular Weight in $g.mol^{-1}$	GlideScore in $kcal.mol^{-1}$	Site Target
Thymol		150.22	-6.85	I
2-Bromophenol		173.01	-6.17	I
4-Methylcatechol		124.14	-6.18	I
2'-Deoxyinosine		252.23	-7.69	I
(−)-Menthol		156.27	-6.75	I
4-Nitrocatechol		155.11	-6.41	I
4-Methyl-2-Nitrophenol		153.14	-6.31	I
Isatin		147.13	-6.30	I
2,6-Diisopropylphenol		178.27	-6.53	I

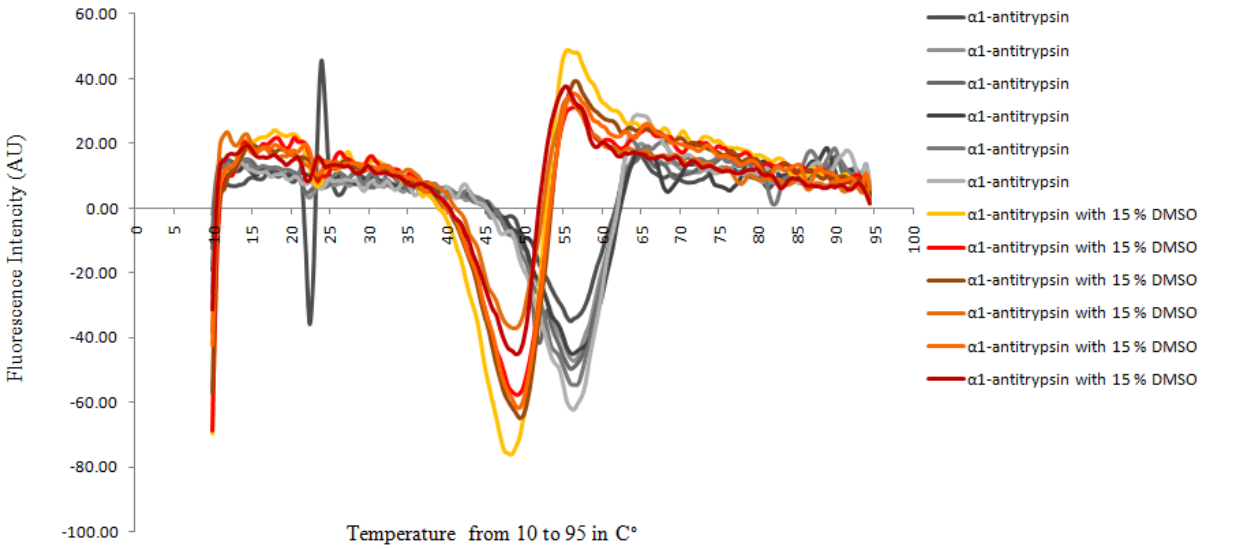
L-Rhamnose Monohydrate		182.17	-8.10	C
5-Acetyluracil		154.12	-6.73	I
Procaterol Hydrochloride		326.82	-7.11	D
Oxypurinol		152.11	-7.16	D
(±)-Sotalol Hydrochloride		308.82	-7.16	E
Tizanidine Hydrochloride		290.17	-7.12	E
Mebendazole		295.29	-8.93	C
D-Allose		180.16	-7.67	C
3,4-Dihydroxybenzoic Acid		154.12	-6.32	I

I'. – ThermoFluor Data – Each Graph Contains the Repeats for Each Compound Assayed

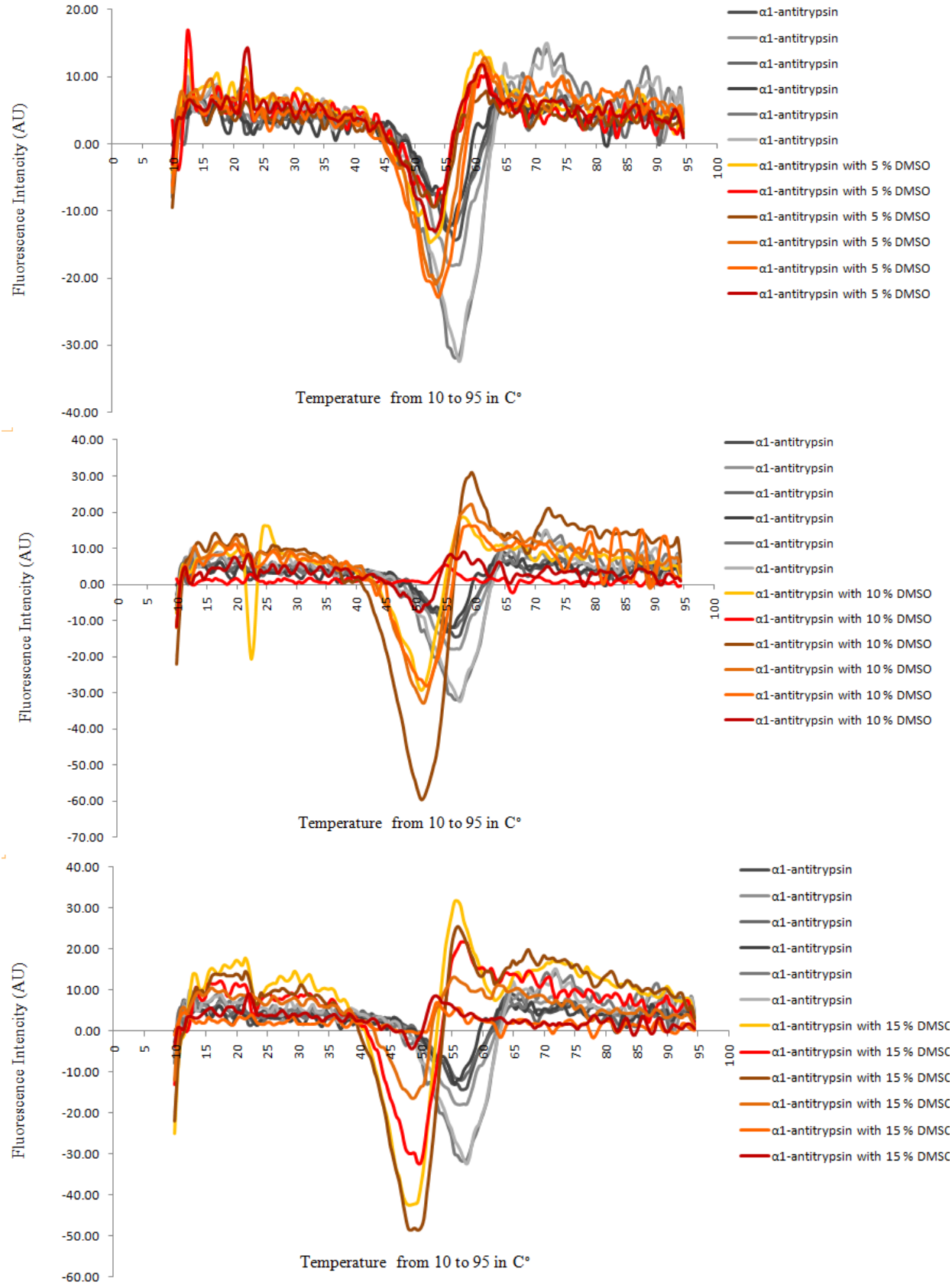
Controls 1 - 96 Well Plate 1 (α_1-Antiprypsin vs. α_1-Antiprypsin in either 5 %, 10 % or 15 % DMSO)



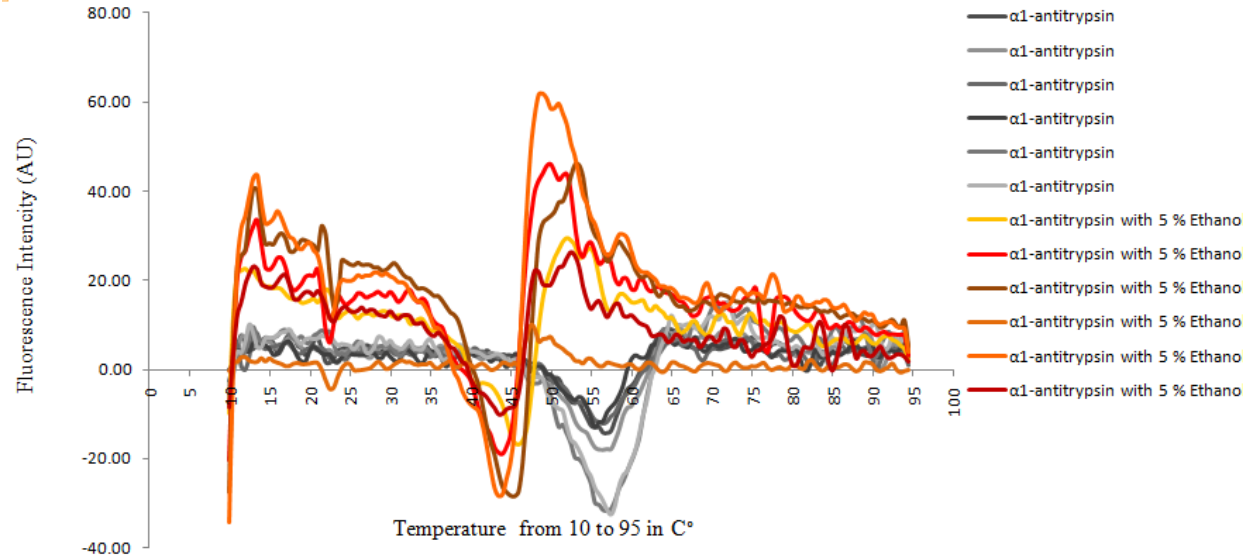


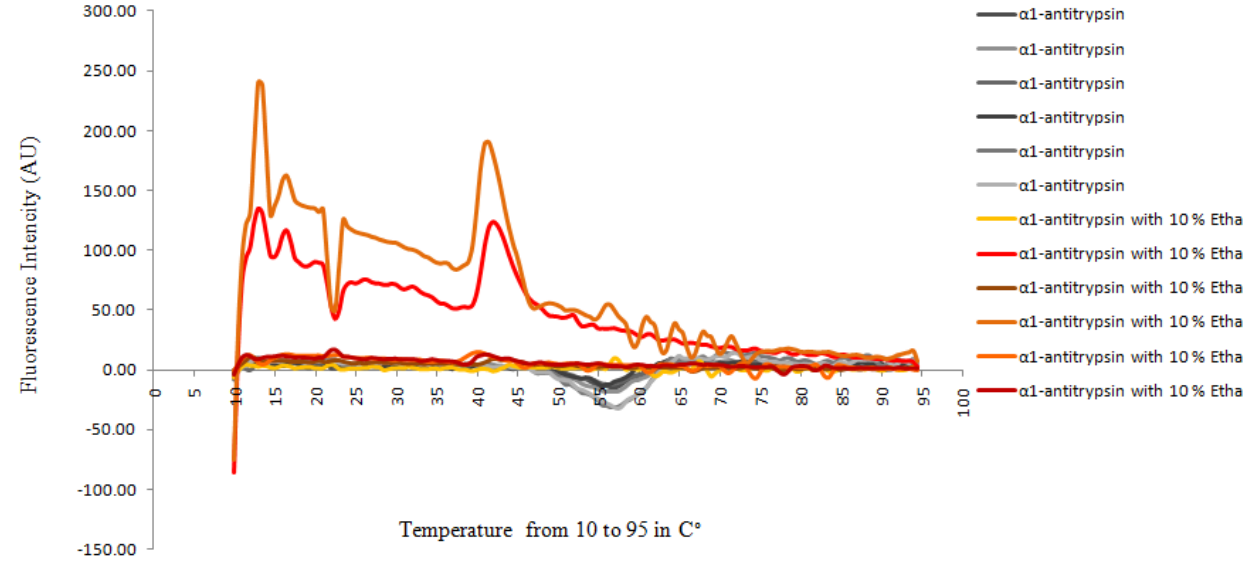


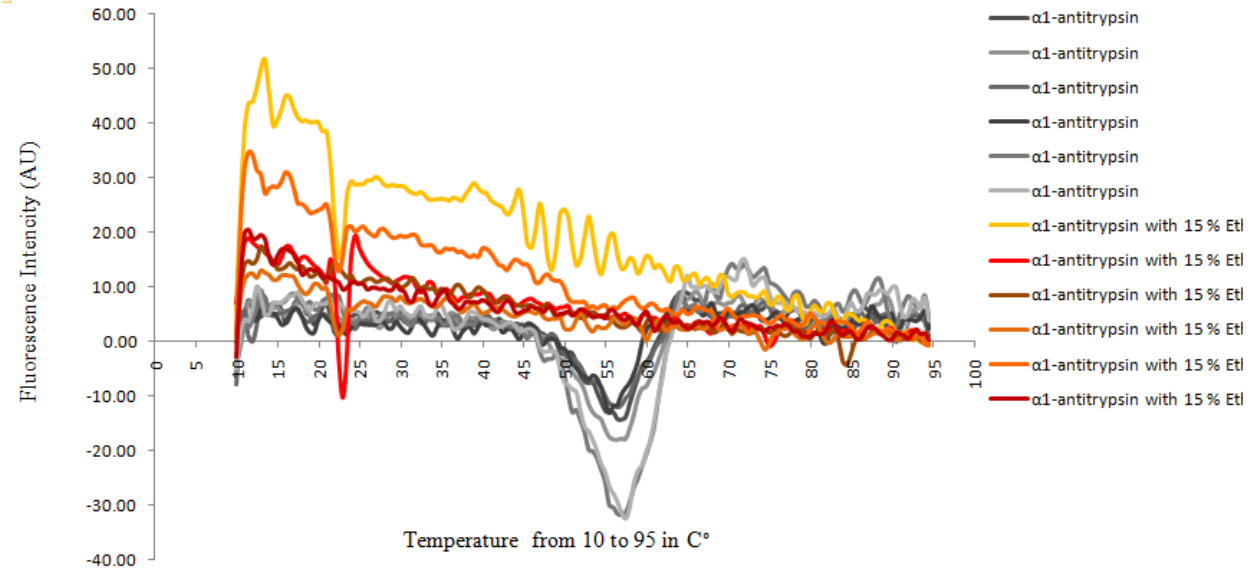
Controls 2 - 96 Well Plate 2 (α_1-Antiprypsin vs. α_1-Antiprypsin in either 5 %, 10 % or 15 % DMSO)



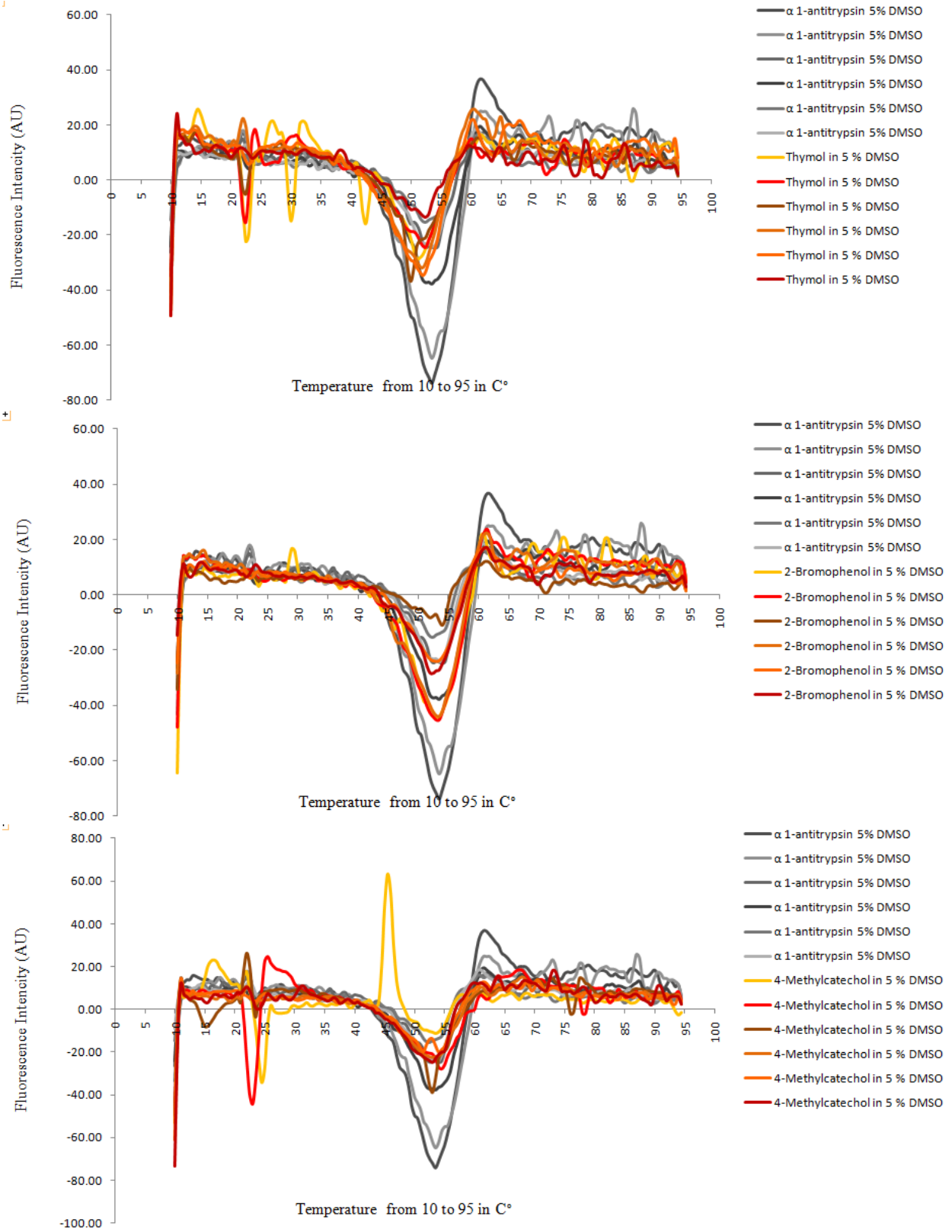
Controls 2 - 96 Well Plate 2 – Ethanol (α_1-Antiprypsin vs. α_1-Antiprypsin in either 5 %, 10 % or 15 % Ethanol)

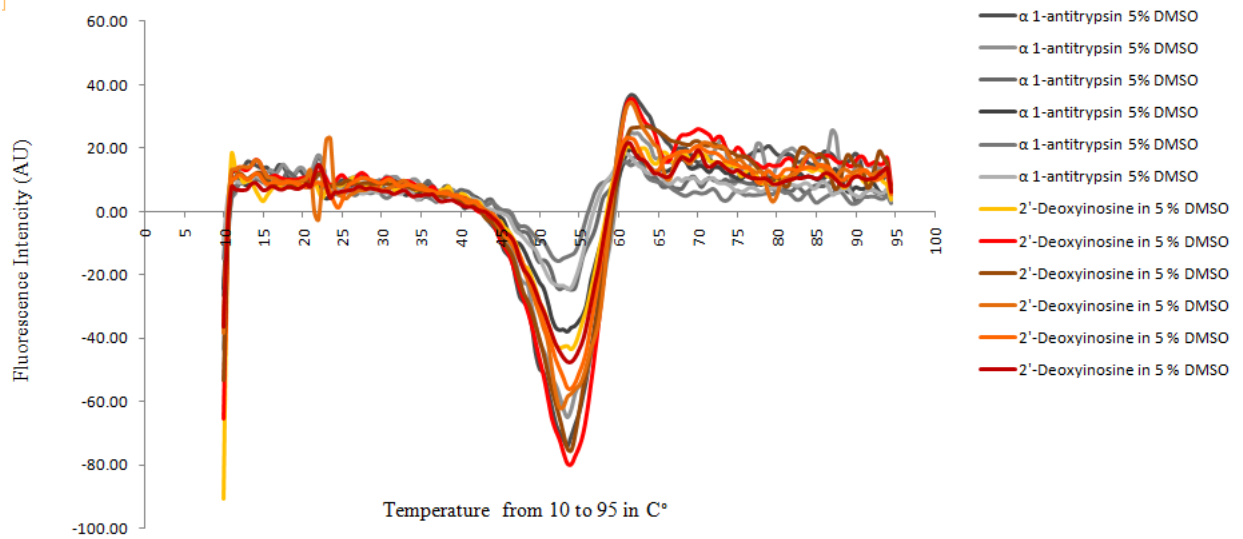


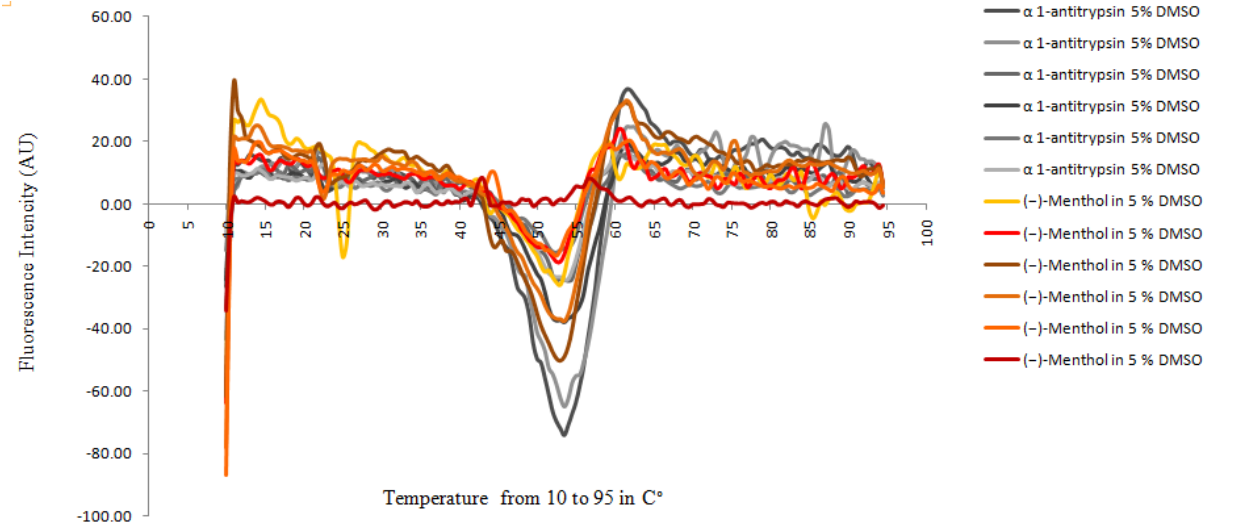


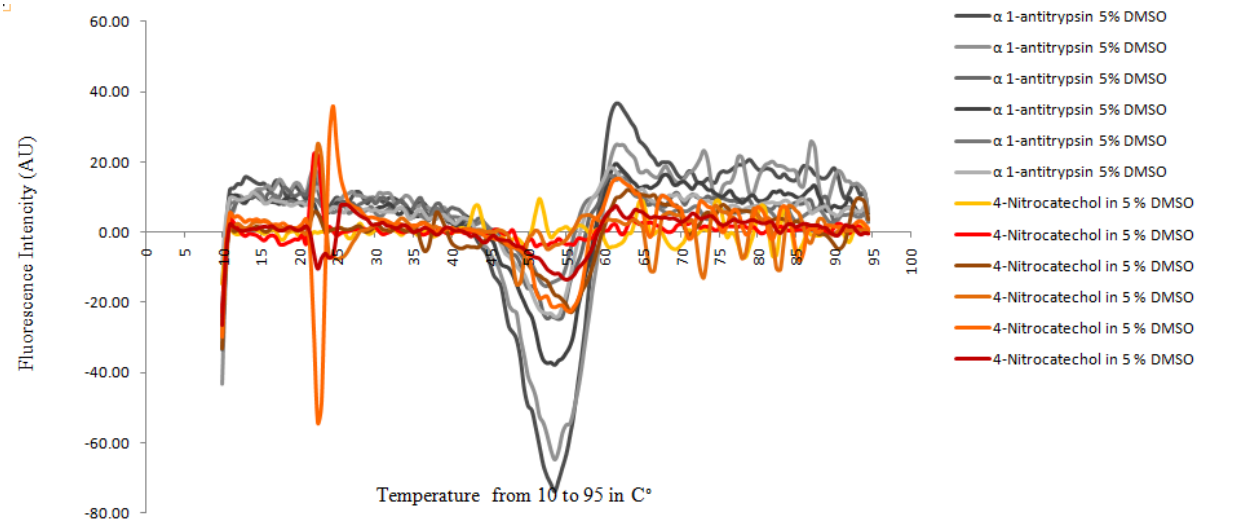


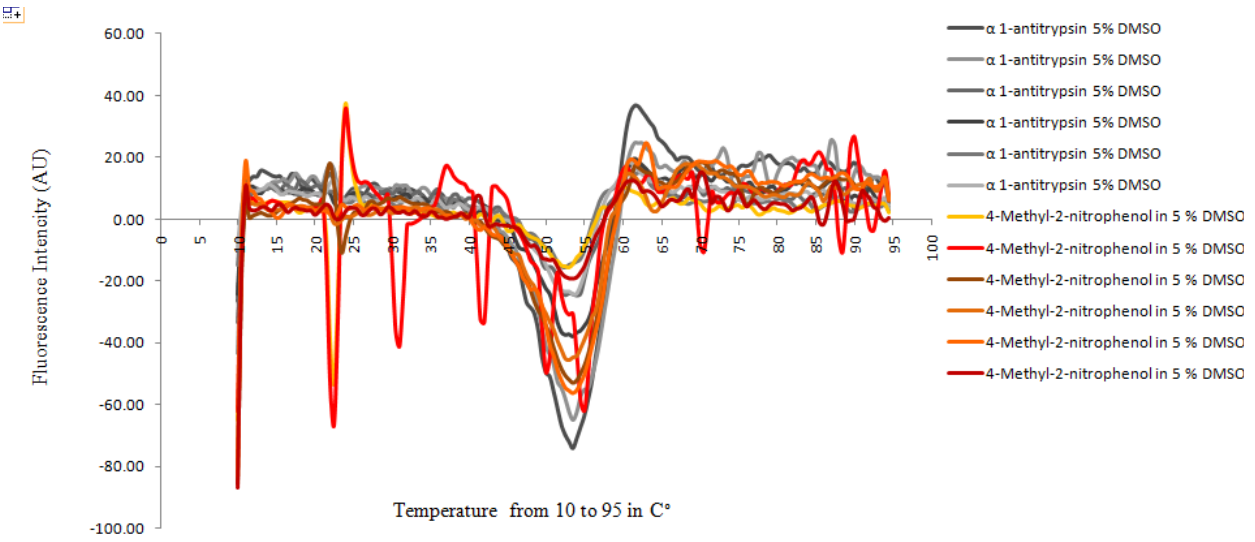
Compounds - 96 Well Plate 1 (α_1-Antiprypsin in 5 % DMSO vs. α_1-Antiprypsin and Compound at 1 mM in 5 % DMSO)

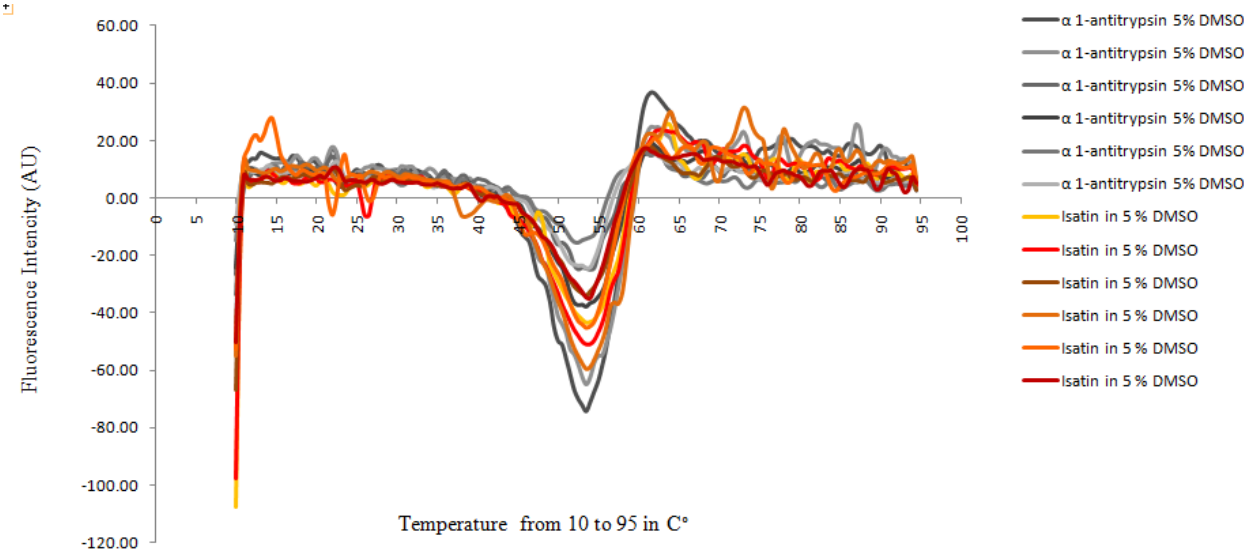


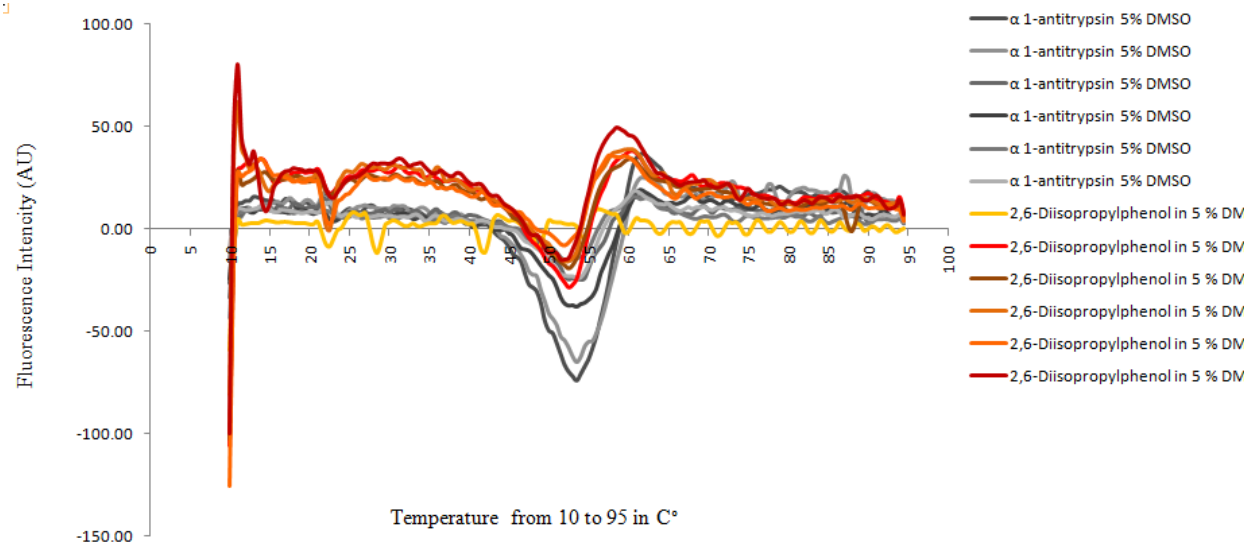

Fluorescence Intensity (AU)
60.00
40.00
20.00
0.00
-20.00
-40.00
-60.00
-80.00
-100.00
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100
Temperature from 10 to 95 in C°
α 1-antitrypsin 5% DMSO
α 1-antitrypsin 5% DMSO
α 1-antitrypsin 5% DMSO
α 1-antitrypsin 5% DMSO
α 1-antitrypsin 5% DMSO
α 1-antitrypsin 5% DMSO
2'-Deoxyinosine in 5 % DMSO
2'-Deoxyinosine in 5 % DMSO
2'-Deoxyinosine in 5 % DMSO
2'-Deoxyinosine in 5 % DMSO
2'-Deoxyinosine in 5 % DMSO
2'-Deoxyinosine in 5 % DMSO

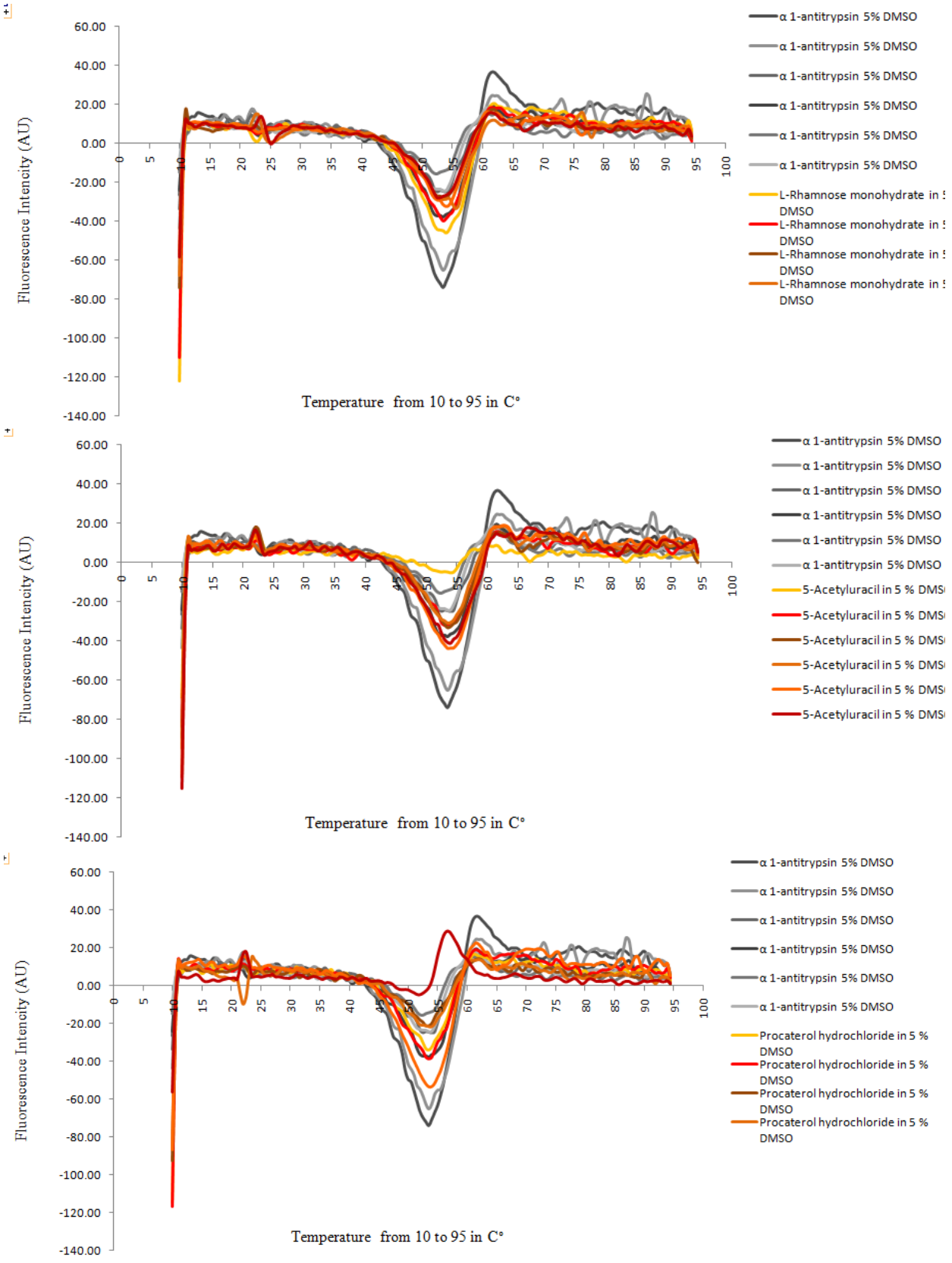

Fluorescence Intensity (AU)
60.00
40.00
20.00
0.00
-20.00
-40.00
-60.00
-80.00
-100.00
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100
Temperature from 10 to 95 in C°
α 1-antitrypsin 5% DMSO
α 1-antitrypsin 5% DMSO
α 1-antitrypsin 5% DMSO
α 1-antitrypsin 5% DMSO
α 1-antitrypsin 5% DMSO
α 1-antitrypsin 5% DMSO
(−)-Menthol in 5 % DMSO
(−)-Menthol in 5 % DMSO
(−)-Menthol in 5 % DMSO
(−)-Menthol in 5 % DMSO
(−)-Menthol in 5 % DMSO
(−)-Menthol in 5 % DMSO


Fluorescence Intensity (AU)
60.00
40.00
20.00
0.00
-20.00
-40.00
-60.00
-80.00
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100
Temperature from 10 to 95 in C°
α 1-antitrypsin 5% DMSO
α 1-antitrypsin 5% DMSO
α 1-antitrypsin 5% DMSO
α 1-antitrypsin 5% DMSO
α 1-antitrypsin 5% DMSO
α 1-antitrypsin 5% DMSO
4-Nitrocatechol in 5 % DMSO
4-Nitrocatechol in 5 % DMSO
4-Nitrocatechol in 5 % DMSO
4-Nitrocatechol in 5 % DMSO
4-Nitrocatechol in 5 % DMSO
4-Nitrocatechol in 5 % DMSO

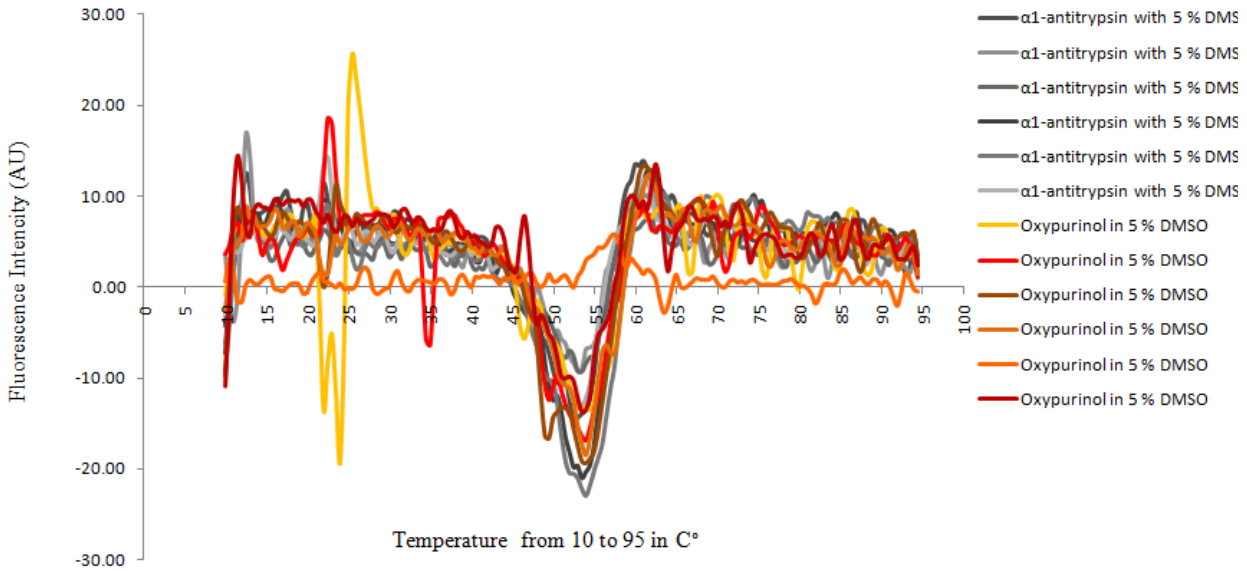

Fluorescence Intencity (AU)
60.00
40.00
20.00
0.00
-20.00
-40.00
-60.00
-80.00
-100.00
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100
Temperature from 10 to 95 in C°
α 1-antitrypsin 5% DMSO
α 1-antitrypsin 5% DMSO
α 1-antitrypsin 5% DMSO
α 1-antitrypsin 5% DMSO
α 1-antitrypsin 5% DMSO
α 1-antitrypsin 5% DMSO
4-Methyl-2-nitrophenol in 5 % DMSO
4-Methyl-2-nitrophenol in 5 % DMSO
4-Methyl-2-nitrophenol in 5 % DMSO
4-Methyl-2-nitrophenol in 5 % DMSO
4-Methyl-2-nitrophenol in 5 % DMSO
4-Methyl-2-nitrophenol in 5 % DMSO

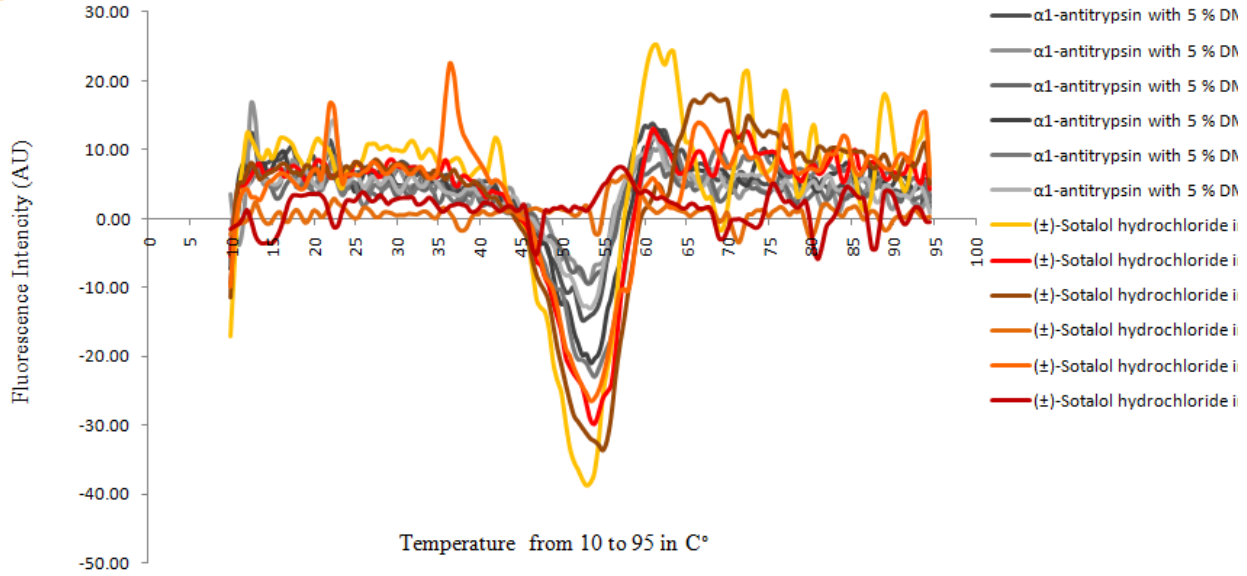

Fluorescence Intencity (AU)
60.00
40.00
20.00
0.00
-20.00
-40.00
-60.00
-80.00
-100.00
-120.00
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100
Temperature from 10 to 95 in C°
α 1-antitrypsin 5% DMSO
α 1-antitrypsin 5% DMSO
α 1-antitrypsin 5% DMSO
α 1-antitrypsin 5% DMSO
α 1-antitrypsin 5% DMSO
α 1-antitrypsin 5% DMSO
Isatin in 5 % DMSO
Isatin in 5 % DMSO
Isatin in 5 % DMSO
Isatin in 5 % DMSO
Isatin in 5 % DMSO
Isatin in 5 % DMSO

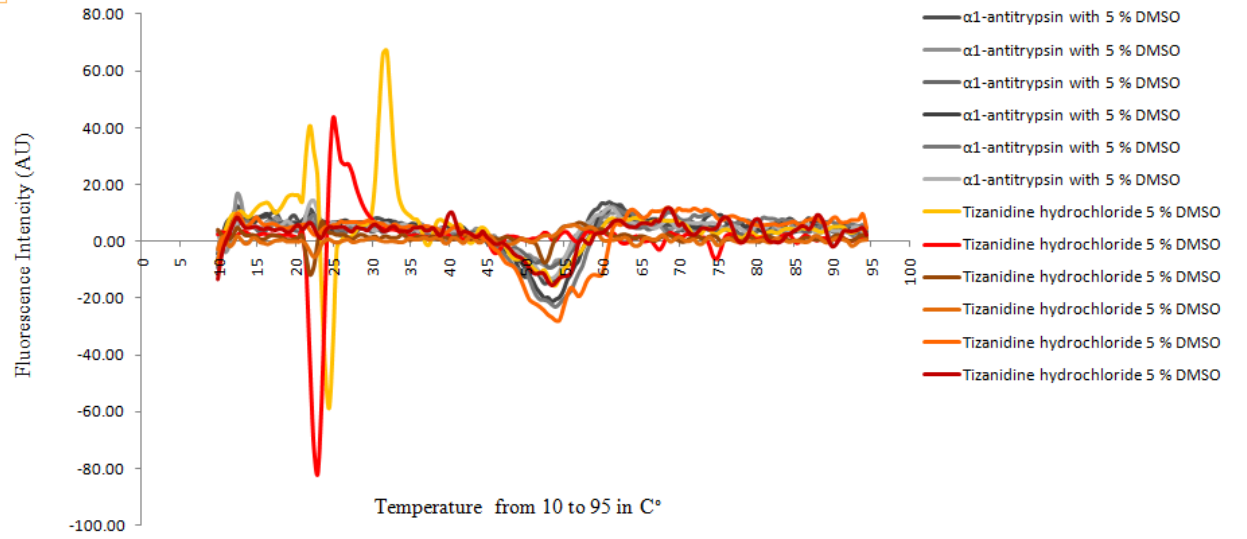

Fluorescence Intencity (AU)
100.00
50.00
0.00
-50.00
-100.00
-150.00
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100
Temperature from 10 to 95 in C°
α 1-antitrypsin 5% DMSO
α 1-antitrypsin 5% DMSO
α 1-antitrypsin 5% DMSO
α 1-antitrypsin 5% DMSO
α 1-antitrypsin 5% DMSO
α 1-antitrypsin 5% DMSO
2,6-Diisopropylphenol in 5 % DM
2,6-Diisopropylphenol in 5 % DM
2,6-Diisopropylphenol in 5 % DM
2,6-Diisopropylphenol in 5 % DM
2,6-Diisopropylphenol in 5 % DM
2,6-Diisopropylphenol in 5 % DM

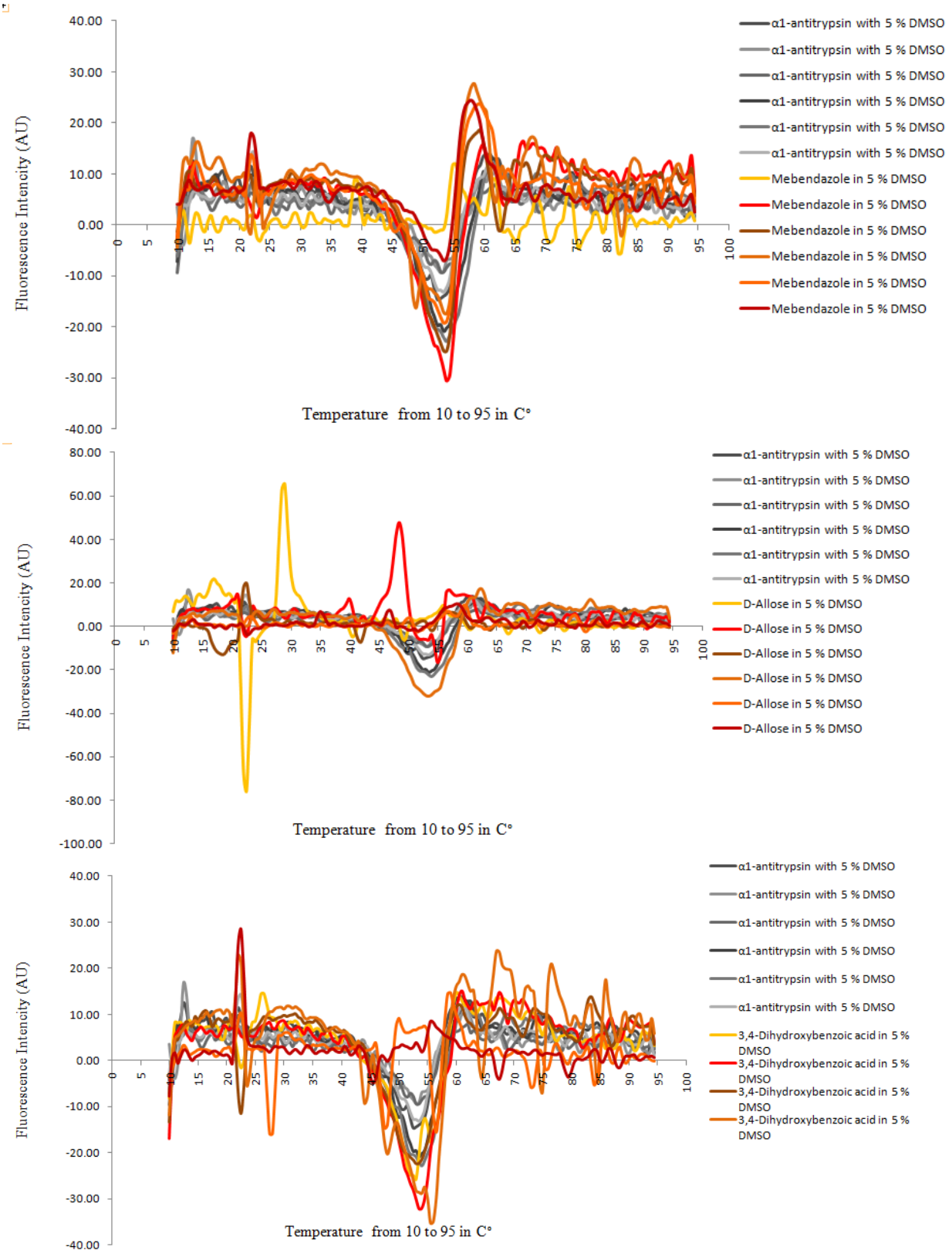


Compounds - 96 Well Plate 2 (α_1-Antiprypsin in 5 % DMSO vs. α_1-Antiprypsin and Compound at 1 mM in 5 % DMSO)

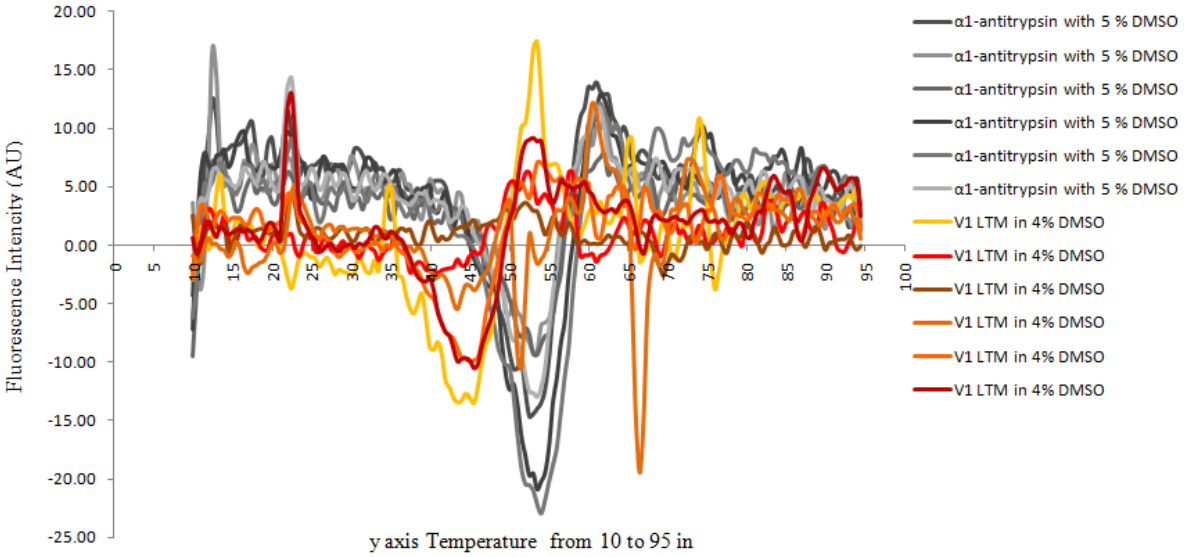

Fluorescence Intencity (AU)
Temperature from 10 to 95 in C°
α1-antitrypsin with 5 % DMSO
Oxypurinol in 5 % DMSO

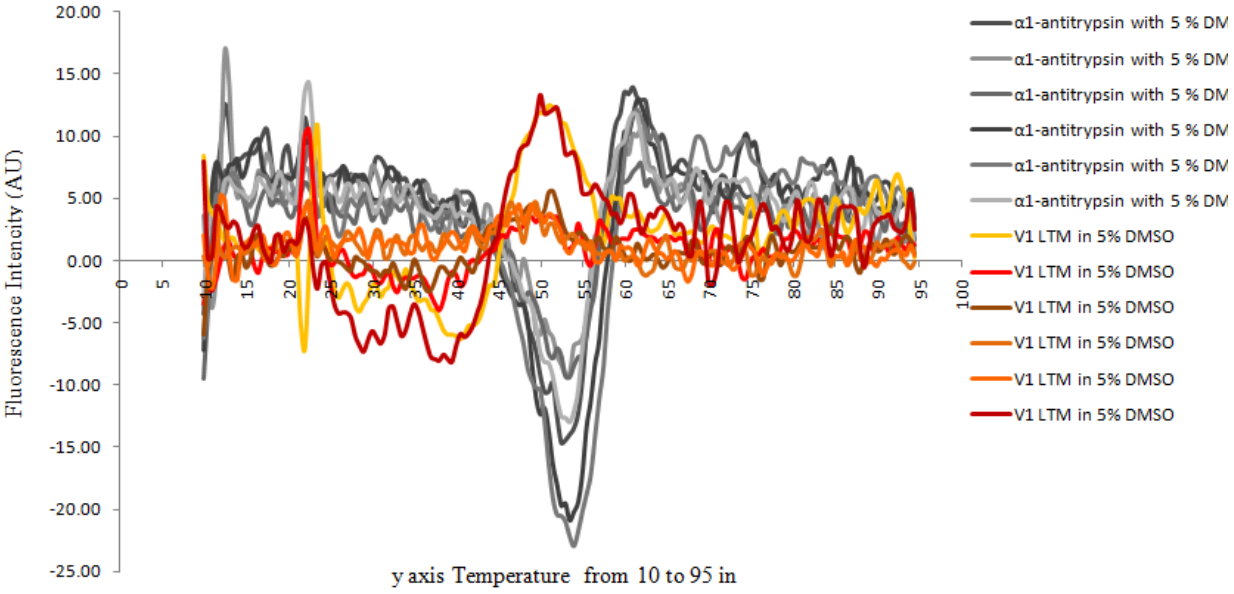

Fluorescence Intencity (AU)
Temperature from 10 to 95 in C°
α1-antitrypsin with 5 % DMSO
(±)-Sotalol hydrochloride


Fluorescence Intencity (AU)
Temperature from 10 to 95 in C°
α1-antitrypsin with 5 % DMSO
Tizanidine hydrochloride 5 % DMSO



Compounds - 96 Well Plate 2 (α_1-Antiprypsin in 5 % DMSO vs. α_1-Antiprypsin and V1-LTM in 5 % DMSO)


Fluorescence Intencity (AU)
20.00
15.00
10.00
5.00
0.00
-5.00
-10.00
-15.00
-20.00
-25.00
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100
α1-antitrypsin with 5 % DMSO
α1-antitrypsin with 5 % DMSO
α1-antitrypsin with 5 % DMSO
α1-antitrypsin with 5 % DMSO
α1-antitrypsin with 5 % DMSO
α1-antitrypsin with 5 % DMSO
V1 LTM in 4% DMSO
V1 LTM in 4% DMSO
V1 LTM in 4% DMSO
V1 LTM in 4% DMSO
V1 LTM in 4% DMSO
V1 LTM in 4% DMSO
y axis Temperature from 10 to 95 in


Fluorescence Intencity (AU)
20.00
15.00
10.00
5.00
0.00
-5.00
-10.00
-15.00
-20.00
-25.00
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100
α1-antitrypsin with 5 % DM
α1-antitrypsin with 5 % DM
α1-antitrypsin with 5 % DM
α1-antitrypsin with 5 % DM
α1-antitrypsin with 5 % DM
α1-antitrypsin with 5 % DM
V1 LTM in 5% DMSO
V1 LTM in 5% DMSO
V1 LTM in 5% DMSO
V1 LTM in 5% DMSO
V1 LTM in 5% DMSO
V1 LTM in 5% DMSO
y axis Temperature from 10 to 95 in