uOttawa

L'Université canadienne
Canada's university

FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTOCTORALES

uOttawa

L'Université canadienne
Canada's university

FACULTY OF GRADUATE AND POSDOCTORAL STUDIES

Sam Khalouei

AUTEUR DE LA THÈSE / AUTHOR OF THESIS

M.Sc. (Biology)

GRADE / DEGREE

Department of Biology

FACULTÉ, ÉCOLE, DÉPARTEMENT / FACULTY, SCHOOL, DEPARTMENT

Translation Initiation in Human Immunodeficiency Virus Type 1 (HIV-1)

TITRE DE LA THÈSE / TITLE OF THESIS

Dr. Xuhua Xia

DIRECTEUR (DIRECTRICE) DE LA THÈSE / THESIS SUPERVISOR

CO-DIRECTEUR (CO-DIRECTRICE) DE LA THÈSE / THESIS CO-SUPERVISOR

EXAMINATEURS (EXAMINATRICES) DE LA THÈSE / THESIS EXAMINERS

Dr. Michel Dumontier

Dr. Stéphane Aris-Brosou

Dr. Guy Drouin

Gary W. Slater

Le Doyen de la Faculté des études supérieures et postdoctorales / Dean of the Faculty of Graduate and Postdoctoral Studies

Translation Initiation in Human Immunodeficiency Virus Type 1 (HIV-1)

Sam Khalouei

Thesis submitted to the
Faculty of Graduate and Postdoctoral Studies
University of Ottawa
In partial fulfillment of the requirements for the Masters degree
In the Ottawa-Carleton Institute of Biology

Thèse soumise à
Faculté des études supérieures et postdoctorales
Université d'Ottawa
en vue de l'obtention de la maîtrise ès sciences

L'Institut de biologie d'Ottawa-Carleton

Your file Votre référence
ISBN: 978-0-494-49224-6
Our file Notre référence
ISBN: 978-0-494-49224-6

Table of Content:

Acknowledgements.....iii

Abstract.....iv

Résumé.....v

IUPAC Code Table.....vi

Universal Genetic Code.....vii

Abbreviations..... vii

List of Chapters.....viii

List of Tables.....x

List of Figures.....xii

Chapter 1.....1

Chapter 2.....35

Chapter 3.....47

Chapter 4.....56

Chapter 5.....75

Chapter 6.....83

Chapter 7.....92

Contribution of Collaborators.....95

References.....96

Copyright Permissions.....109

Acknowledgements

First of all, I would like to thank my supervisor, Dr. Xuhua Xia, for all his support and for providing me with the opportunity to pursue my graduate studies. I would like to thank him for patiently answering my frequent questions and most importantly for teaching me how to be a better student.

I am grateful to my advisory committee members, Dr. Natalie Goto, Dr. Guy Drouin, and Dr. Michel Dumontier, for attending my committee meetings, resolving my questions, and providing me with insightful guidance throughout my Masters degree. I am also grateful to Dr. Drouin, Dr. Dumontier, and Dr. Stéphane Aris-Brosou for examining my thesis and oral defense and to Dr. Fabien Avaron for helping me with the French translation of the abstract.

I am thankful to the members of Dr. Xia's laboratory, for interesting discussions and feedbacks and for reviewing my manuscripts.

Various public databases' staff members helped me with my search and download problems, which is greatly appreciated, in particular Dr. Ivo Pedruzzi (Swiss-Prot), Dr. Brian Foley (Los Alamos HIV database), and Dr. Giulietta Spudich (Ensembl database).

I am grateful to the University of Ottawa and the Department of Biology for providing me with financial assistance, and the granting agencies for their contribution to my research.

I would like to thank my family for their continuous support and encouragement, both in the past and during my Masters studies. And last but not least, I would like to thank my wife, Samaneh, for her great support and understanding.

Abstract

Translation of human immunodeficiency virus type 1 (HIV-1) mRNAs is entirely dependent on the host translation machinery. There are two prevailing hypotheses regarding the translation initiation mechanism in HIV-1; conventional cap-dependent ribosomal scanning mechanism (CDRSM) and cap-independent entry of the ribosome, usually through an internal ribosome entry site (IRES). The first mechanism makes use of the Kozak consensus sequence in locating the translation initiation codon, similar to the mechanism observed in human mRNAs. Therefore, a thorough understanding of the Kozak consensus and translation initiation in human would also shed light on the mechanism of translation initiation in HIV-1. The role of Kozak +4G site in translation initiation has been controversial, with the alternative hypothesis explaining the prevalence of +4G by invoking the observation that small amino acids, coded by G-starting codons, which are efficient for N-terminal methionine excision (NME), are preferred at the penultimate (second) position. Using two bioinformatics approaches we provide strong support for this alternative hypothesis and provide evidence contradicting the involvement of +4G in translation initiation.

One of the predictions of the CDRSM hypothesis is a high conservation of Kozak consensus sequence in different HIV-1 sequences. Our results presented here validate this prediction. The CDRSM hypothesis also predicts that there should be a selective pressure against ATG usage in optimal context in the HIV-1 5'-UTR to avoid their erroneous detection by the scanning ribosome, whereas the IRES-dependent mechanism in the presence of stable secondary structures, predicts no such selective pressure because these ATGs would be embedded in the secondary structures. Here we demonstrate this selective pressure in the HIV-1 5'-UTR which further supports the CDRSM hypothesis. Finally, we present evidence for strong site conservation in the 5'-UTR of HIV-1 sequences, which not only point to as yet unknown mechanisms of translation initiation, but also provide a mean to separate HIV-1 and human mRNAs. This implies that it is theoretically possible to design HIV-1-specific translation inhibition drugs.

Résumé

Le processus de traduction chez le virus d'immunodéficience humaine de type 1 (HIV-1) est entièrement dépendant de la machinerie traductionnelle de l'hôte infecté. Il existe deux hypothèses principales concernant l'initiation de la traduction chez HIV-1 : Le mécanisme conventionnel de balayage de la coiffe par les ribosomes (CDRSM), et l'entrée « coiffe-indépendante » généralement par le biais d'un site d'entrée interne d'un ribosome (IRES). Le premier mécanisme utilise la séquence consensus de Kozak pour l'identification du codon d'initiation, d'une façon similaire au mécanisme utilisé chez les ARNm humains. Aussi, une connaissance approfondie de la séquence consensus de Kozak ainsi que de l'initiation de la traduction chez l'humain permettrait de mieux comprendre le mécanisme d'initiation de la traduction chez HIV-1. Le rôle joué par le site Kozak +4G dans l'initiation de la traduction est sujet à controverse, et une hypothèse alternative explique la prévalence du +4G par le fait que les acides aminés de petite taille, codés par des codons débutant par G, et qui sont efficaces pour l'excision des méthionines N-terminales (NME), sont préférés en position pénultième (seconde position). En utilisant deux méthodes bioinformatiques, nous avons apporté des éléments fortement en faveur de cette hypothèse alternative, ainsi que des résultats contestant l'implication des +4G dans l'initiation de la traduction.

Une des prédictions de l'hypothèse CDRSM est une forte conservation des séquences de Kozak dans les différentes séquences de HIV-1. Les résultats présentés dans cette étude valident cette prédiction. L'hypothèse CDRSM prévoit également l'existence d'une pression de sélection contre l'utilisation de codons ATG en contexte optimal dans la région 5'-UTR de HIV-1 afin d'éviter leur détection erronée par les ribosomes au moment du balayage, tandis que le mécanisme faisant intervenir les IRES en présence de structures secondaires stables ne prédit pas de telle pression de sélection, car dans ce cas les ATG se retrouvent enfermés à l'intérieur des structures secondaires. Nous démontrons ici l'existence d'une pression sélective au niveau du 5'-UTR de HIV-1, ce qui appuie l'hypothèse du CDRSM. Finalement nous apportons des éléments en faveur de l'existence de sites conservés dans les séquences 5'-UTR de HIV-1, ce qui non seulement suggère une implication dans des mécanismes d'initiation de la traduction jusqu'alors inconnus, mais aussi pourrait apporter un moyen de distinguer les ARNm humain et de ceux de HIV-1. Ceci implique qu'il est théoriquement possible de concevoir des médicaments dirigés spécifiquement contre les mécanismes de traduction de HIV-1.

IUPAC Code Table

One Letter Code	Three Letter Code	Name
A	Ala	Alanine
C	Cys	Cysteine
D	Asp	Aspartate
E	Glu	Glutamate
F	Phe	Phenylalanine
G	Gly	Glycine
H	His	Histidine
I	Ile	Isoleucine
K	Lys	Lysine
L	Leu	Leucine
M	Met	Methionine
N	Asn	Asparagine
P	Pro	Proline
Q	Gln	Glutamine
R	Arg	Arginine
S	Ser	Serine
T	Thr	Threonine
V	Val	Valine
W	Trp	Tryptophan
Y	Tyr	Tyrosine
B		Aspartate or Asparagine
U		Selenocysteine
Z		Glutamate or Glutamine

Code	Nucleotides Represented	Code	Nucleotides Represented
A	Adenosine	M	A C (amino)
C	Cytidine	S	G C (strong)
G	Guanine	W	A T (weak)
T	Thymidine	B	G T C
U	Uridine	D	G A T
R	G A (purine)	H	A C T
Y	T C (pyrimidine)	V	G C A
K	G T (keto)	N	A G C T (any)

Universal Genetic Code

First (5') letter		Second letter: U		C		A		G		Third (3') letter
	U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
		UUC		UCC		UAC		UGC		C
		UUA	Leu	UCA		UAA	Stop	UGA	Stop	A
		UUG		UCG		UAG		UGG	Trp	G
	C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
		CUC		CCC		CAC		CGC		C
		CUA		CCA		CAA	Gln	CGA		A
		CUG		CCG		CAG		CGG		G
	A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
		AUC		ACC		AAC		AGC		C
		AUA		ACA		AAA	Lys	AGA	Arg	A
		AUG	Met	ACG		AAG		AGG		G
	G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
		GUC		GCC		GAC		GGC		C
		GUA		GCA		GAA	Glu	GGA		A
		GUG		GCG		GAG		GGG		G

Abbreviations

CAI	Codon Adaptation Index
CDRSM	Conventional Cap-Dependent Ribosomal Scanning Mechanism
CDS	Coding Sequence
HAART	Highly Active Antiretroviral Therapy
IRE	Internal Ribosome Entry
IRES	Internal Ribosome Entry Site
NME	N-terminal Methionine Excision
ORF	Open Reading Frame
PWM	Positional Weighting Matrix
TFBS	Transcription Factor Binding Site
LTR	Long Terminal Repeat
UTR	Untranslated Region
MA	Matrix
CA	Capsid
NC	Nucleocapsid
PR	Protease
RT	Reverse Transcriptase
IN	Integrase
SU	Surface glycoprotein (gp120)
TM	Transmembrane protein (gp41)
TAR	Transactivation Response Region
polyA	Polyadenylation signal
PBS	Primer Binding Site
DIS	RNA Dimer Initiation signal
SD1	Splice Donor site 1

List of Chapters

CHAPTER 1 INTRODUCTION 1

1.1 Viruses 1
1.1.1 Retroviruses 2

1.2 AIDS 2

1.3 Human Immunodeficiency Virus Type 1 (HIV-1) 3
1.3.1 HIV-1 Subtypes 4
1.3.2 HIV-1 Infection Life-History 5
1.3.3 HIV-1 Genome 7
1.3.4 HIV-1 Particle 13
1.3.5 HIV-1 Life Cycle 14
1.3.5.1 Cell Entry 15
1.3.5.2 Reverse Transcription 16
1.3.5.3 Integration 17
1.3.5.4 Transcription 17
1.3.5.5 Translation 21
1.3.5.6 Viral Particle Formation and Release 25
1.3.5.7 Viral Particle Maturation 26
1.3.6 HIV-1 Mutation 27

1.4 HIV-1 resources used in this project 28
1.4.1 Public HIV Databases 28
1.4.2 LANL Database 29
1.4.3 HIV Structures in Protein Data Bank (PDB) 32

1.5 Overview of subsequent chapters 33

CHAPTER 2 TRANSLATION INITIATION IN HUMAN: THE ROLE OF +4G OF THE KOZAK CONSENSUS 35

2.1 Abstract 35

2.2 Introduction 36

2.3 Materials and Methods 39

2.4 Results and Discussion 41

CHAPTER 3 KOZAK CONSENSUS IN HIV-1 GENES 47

3.1 Abstract 47

3.2 Introduction 47

3.3 Materials and Methods 48

3.4 Results and Discussion 50

CHAPTER 4 STRONG SITE CONSERVATIONS IN HIV-1 5'-UTR POINT TO FUNCTIONAL ROLE IN TRANSLATION INITIATION 56

4.1 Abstract 56

4.2 Introduction 56

4.3 Materials and Methods 58

4.4 Results and Discussion 64

CHAPTER 5 HIV-1 TRANSLATION INITIATION AND SELECTIVE PRESSURE AGAINST ATG USAGE IN 5'-UTR 75

5.1 Abstract 75

5.2 Introduction 75

5.3 Materials and Methods 76

5.4 Results and Discussion 79

CHAPTER 6 BIOINFORMATIC APPROACH TO IDENTIFY PENULTIMATE AMINO ACIDS EFFICIENT FOR N-TERMINAL METHIONINE EXCISION 83

6.1 Abstract 83

6.2 Introduction 83

6.3 Methods and Materials 85

6.4 Results and Discussion 86

CHAPTER 7 CONCLUDING REMARKS 92

CONTRIBUTION OF COLLABORATORS 95

REFERENCES 96

List of Tables

Table 1-1 HIV-1 Genes and their function. Adapted and modified from HIV Sequence Compendium 2005 (Leitner et al., 2005). 8
Table 1-2 Different HIV-1 mRNAs generated from different combinations of exons. Exon numbers and combinations are according to figure 1(C) and adapted from Purcell and Martin, 1993. 21
Table 1-3 HIV-1 molecular and structural resources. 29
Table 1-4 Available Number of HIV-1 Sequences on LANL based on Geographic Region. Table compiled on June 25th, 2007. 30
Table 1-5 Available number of HIV-1 subtypes on LANL. Table compiled on June 25th, 2007 31
Table 1-6 Available Number of HIV-1 sequences on LANL. Table compiled on June 25th, 2007 31
Table 2-1 The molecular weight (MW) and Gyration radius of the amino acids categorized into being coded by G-starting codons and NonG-starting codons 39
Table 2-2 The nucleotide frequencies of the -3 and +4 sites of the human CDSs. The percentages in the brackets, represent the proportion of +4 sites in each -3 site category so that the sum of the percentages in each row is equal to 100. 42
Table 2-3 The nucleotide frequencies of the -3 and +4 sites of the human CDSs in the low-CAI and high-CAI groups. The percentages in the brackets, represent the proportion of +4 sites in each -3 site category so that the sum of the percentages in each row is equal to 100. 44
Table 3-1 Kozak sequences in NCBI HIV-1 type sequence (NC_001802) and their comparison with Kozak consensus. Positions -3 and +4 of the Kozak consensus are shown in bold. Individual nucleotides that resemble the corresponding Kozak consensus positions are shown in capital letters. 50
Table 3-2 Nucleotide frequencies of the -3 and +4 sites in Kozak consensus of HIV-1 sequences downloaded from Los Alamos National Library Database (LANL). 53
Table 3-3 Pooled results of Kozak consensus analysis in HIV-1 genes of the five most abundant subtypes obtained from LANL database. The most abundant combination of nucleotides at positions -3 and +4, and the percentage of sequences with -3R and +4G (last column) are shown. N/A in *vpu* refers to the fact that the most abundant combination of *vpu*, as opposed to other genes, is not the same among the five subtypes. 54
Table 4-1 Site-specific frequency distribution and position weight matrix (computed with the logarithm of base 2) for the 50 nucleotides upstream of the coding genes in 86 HIV-1 genomes. Highest scores for sites -17 and -25 are in bold. 65
Table 4-2 Summary statistics of position weight matrix scores by HIV-1 genes 67
Table 4-3 Weighting matrix from perceptron with the maximum value in each row shown in the last column. 68
Table 4-4 Delete-half jackknifing. Number of perceptron iterations before the convergence was reached are shown for each of the 20 runs. Light-gray boxes indicate that the site was among the six highest perceptron weight matrix (PTWM) values and the dark-gray boxes indicate that the site had the highest PTWM value. 71
Table 4-5 Perceptron performed on smaller subsets of HIV-1 and human dataset. Perceptron was run five times for each subset category. Light-gray boxes indicate that the site was among the six highest perceptron weight matrix (PTWM) values and the dark-gray boxes indicate that the site had the highest PTWM value. 72

Table 4-6 Perceptron performed on identical amount of HIV-1 and human sequences. All 331 unique HIV-1 50-mers were used as the POS group, and 331 randomly selected human 50-mers were used as the NEG group in each run. Light-gray boxes indicate that the site was among the six highest perceptron weight matrix (PTWM) values and the dark-gray boxes indicate that the site had the highest PTWM value. 73
Table 5-1 The HIV-1 coding sequences. The position of start and stop codons are indicated according to the NCBI HIV-1 type sequence (NCBI Genome: NC_001802). Start position corresponds to the position of A in ATG start codon and stop position corresponds to the third base position of the stop codon. 77
Table 5-2 Nucleotide frequencies of the 5' UTR of major HIV-1 mRNAs 80
Table 5-3 The expected and observed number of ATGs in HIV-1 5'-UTRs of major transcripts in different optimal contexts 80
Table 5-4 The expected and observed number of ATG_{-3R} and Non-ATG_{-3R} triplets in the concatenated HIV-1 5'-UTR 81
Table 6-1 Details of computing NME efficiency ($E_{NME\cdot i}$) for amino acid i, based on yeast and human proteins. AA – amino acids in 3-letter code, N_i – number of amino acid i at the penultimate site of proteins before any N-terminal processing, M_i – number of amino acid i in the penultimate site of proteins known to undergo NME. $E_{NME\cdot i}$ is specified in Eq. (6.1). 89

List of Figures

Figure 1.1 Global view of HIV infection. Reproduced by kind permission of UNAIDS (UNAIDS, 2006). 3
Figure 1.2 Regional distribution of HIV-1 subtypes and recombinants in 2004. Adapted and modified from (Hemelaar et al., 2006). Reproduced by kind permission of UNAIDS. 5
Figure 1.3 HIV-1 course of infection. $CD4^+$ T cell count is shown with darker color and HIV RNA copies per ml of plasma is shown with a lighter color. Adapted from Wikipedia (http://en.wikipedia.org), under the terms of the GNU free documentation license, version 1.2. 6
Figure 1.4 Genomic Organization of HIV-1. Adapted with permission from (http://www.stanford.edu/group/virus/retro/2005gongishmail/HIV-1b.jpg) 9
Figure 1.5 The complete genome of NCBI HIV-1 Type sequence (NC_001802). Translation initiation codons (highlighted in green) and termination codons (highlighted in light blue) are shown for different HIV-1 genes. Splice acceptor (SA) and donor (SD) sites are shown in bold in orange and blue font colors, respectively. Different exons are underlined in red color, with exons generated by SA4, SA4c, SA4a, and SA4b splice sites shown by discontinuous underlines. Annotations are included in the left and right margins of the graph with matching colors. 10
Figure 1.6 A Schematic diagram of mature HIV-1 particle. RT: reverse transcriptase, NC: nucleocapsid, PR: protease, IN: integrase, MA: matrix protein, CA: capsid protein. 14
Figure 1.7 HIV-1 Virus entry into CD4+ T-cells. Interaction of gp120 on the surface of the virus with CD4 receptor in the presence of a coreceptor, CXCR4 or CCR5, induces a membrane fusion followed by the injection of viral particle contents into the cell cytoplasm. 16
Figure 1.8 U3 region of HIV-1 LTR. Various transcriptional promoter elements are shown; activator protein-1 (AP-1); nuclear factor of activated T cells (NF-AT); upstream stimulatory factor-1 (USF-1); nuclear factor kappa B (NF-kB); SP-1, and TATA box. Figure adapted and modified from (Morrow et al., 1994). 18
Figure 1.9 Basal transcription of HIV-1 genome from integrated provirus. 18
Figure 1.10 HIV-1 exons and splice sites. (A) HIV-1 genome and different open reading frames (ORF). The vertical lines inside *gag/pol* and *env* ORFs indicate the site of viral and cellular protease cleavage, respectively. *tat* and *rev* ORFs, made up of two coding exons, are shown in darker shades. Adapted and modified from Leitner *et al.* (Leitner et al., 2005). (B) Location of the splice donor (SD, with a L shape flipped vertically) and acceptor (SA, with a L shape flipped horizontally) sites in the NCBI HIV-1 type genome (NCBI Genome: NC_001802), corresponding to the numbers done by Schwartz *et al.* (Schwartz et al., 1990). The first position of the 5'-R region is designated as +1 (see text for more details). (C) Various HIV-1 exons generated by alternative use of splice sites found in Purcell and Martin report (Purcell and Martin, 1993), and numbered according to Muesing *et al.* (Muesing et al., 1985). Figures B and C, adapted and modified from Purcell and Martin (Purcell and Martin, 1993). 20
Figure 1.11 Eukaryotic cap-dependent ribosomal scanning mechanism (CDRSM) of translation initiation. The 40S ribosomal subunit, along with the eIF4F translation initiation complex, binds at the 5'cap and starts scanning the mRNA molecule until it reaches an AUG in optimal context (Kozak consensus). The 60S ribosomal subunit then binds to the complex and the translation of the polypeptide starts. 23

Figure 1.12 Internal Ribosome Entry Site (IRES)-dependent translation initiation mechanism. The 40S ribosomal subunit binds at IRES with the help of IRES Transacting Factors (ITAF). This mechanism is independent of the 5'cap and the 5'-UTR stable secondary structures are bypassed. The depicted secondary structure corresponds to the structure prediction of *vif* shown in (Yilmaz et al., 2006). The IRES location is hypothetical. 25
Figure 4.1 The 50 nucleotides upstream of the translation initiation sites in NCBI HIV-1 type genome (NC_001802). The conserved -17 and -25 sites are shown in bold. 57
Figure 4.2 Location of splicing acceptor sites (SA) 4c, 4a, 4b, 5, and the Rev start codon on the NCBI HIV-1 type sequence (NC_001802). The region is from nucleotides 5456 to 5532 of the complete genome. The small rectangle and circle show the position of the -17 and -25, respectively, relative to the A in Rev start codon, designated as +1..... 60
Figure 4.3 Sequence logo of the 331 unique HIV-1 sequences in this study produced by WebLogo (Crooks et al., 2004)..... 65
Figure 6.1 Different amino acids at the penultimate site result in dramatically different NME Efficiencies in yeast proteins. p_i and q_i are proportions of amino acid i at the penultimate site of unprocessed proteins and NME-processed proteins, respectively. The line indicates the position when $q_i = p_i$..... 87
Figure 6.2 Different amino acids at the penultimate site result in dramatically different NME efficiencies in human proteins, with the meaning of symbols as in Figure 6.1. 91

Chapter 1 Introduction

1.1 Viruses

Viruses are obligate intracellular parasites whose reproduction depends on the host cell. Having a relatively small genome, they are typically constructed from a nucleic acid and a protein coat also known as the capsid. The nucleic acid can be either RNA or DNA (but not both), single or double stranded, linear or circular, and fragmented or continuous.

Among RNA viruses, human immunodeficiency virus type 1 (HIV-1) and hepatitis C virus (HCV) have been extensively analysed, from their genome structure to replication and pathogenesis. Despite the world-wide efforts in gaining knowledge on these deadly pathogens, substantial amount of information is lacking in some aspects of their life cycle. The mechanism of translation initiation, for example, which is the subject of this thesis, has raised many as yet unanswered questions.

Almost all emerging viral pathogens have RNA rather than DNA genome and almost all of them have an animal reservoir, such that cross-species transmission can generally be considered as the underlying mechanism of viral emergence (Cleaveland et al., 2001; Holmes and Rambaut, 2004). Both RNA and DNA viruses can cause chronic disease in humans. Relative to DNA viruses, RNA viruses undergo much more genetic variation throughout their chronic infection due to their short generation times, large population size, high rates of mutation caused by error-prone RNA polymerase (Katz and Skalka, 1990), and recombination (Lai, 1992). These genetic variations can give rise to epidemiological (among-host) and infection (within-host) dynamics, which make prediction of the disease outcome and treatment extremely difficult.

1.1.1 Retroviruses

Retroviruses are a group of RNA viruses with reverse transcription being the hallmark of their replication cycle. The first publication in 1970 which identified an RNA-dependent DNA polymerase in retroviruses came as a surprise to the scientific community which was accustomed to the flow of information from DNA to RNA to protein (Baltimore, 1992; Temin and Mizutani, 1992).

According to the NCBI Taxonomy, retro-transcribing viruses include the Caulimoviridae, Hepadnaviridae, Metaviridae, and Retroviridae families. The Retroviridae family, in turn, includes the Orthoretrovirinae, and Spumaretrovirinae subfamilies. The Lentivirus genus belongs to the Orthoretrovirinae subfamily and includes bovine, equine, feline, ovine, and primate lentivirus groups. Finally, the primate lentivirus group includes the Human Immunodeficiency Virus type 1 (HIV-1) and 2 (HIV-2), Simian Immunodeficiency Virus (SIV), and Simian-Human Immunodeficiency Virus (SHIV) species.

Retroviruses contain two copies of a linear, single, and positive-stranded RNA molecule. Even though the life cycle of different retroviruses can include virus-specific stages, the general aspects are common among various members of this family and similar to that of HIV-1, which is explained in details in section 1.3.5.

1.2 AIDS

Acquired Immunodeficiency Syndrome (AIDS) is a deadly disease with worldwide complications. In 2006 alone 2.9 million people worldwide died from this disease. At the same year about 4.3 million people became newly infected with Human Immunodeficiency Virus (HIV), the causative agent of AIDS. This increased the total number of people infected by this virus to about 39.5 million (UNAIDS, 2006). One of the main problems with this

disease is that the human immune system, which normally protects the body from opportunistic infections, is the main target of HIV (Fauci, 1993; Levy, 1993). Being an intracellular parasite, the HIV virus enters the CD4+ T-lymphocytes of the immune system to use its enzymatic machinery to replicate and form new particles, killing the infected cells in the process (see section 1.3.5).

Among different geographical regions, sub-Saharan Africa has experienced the hardest hit in the AIDS epidemic with approximately 24.7 million adults and children infected (Figure 1.1) (UNAIDS, 2006). Despite the recent advances in disease prevention and treatment, the total number of people living with HIV continues to increase. This is due to both the current rate of about 4.5 million new infections per years and a prolonged life expectancies of those infected, with the aid of potent antiviral drugs (UNAIDS, 2006).

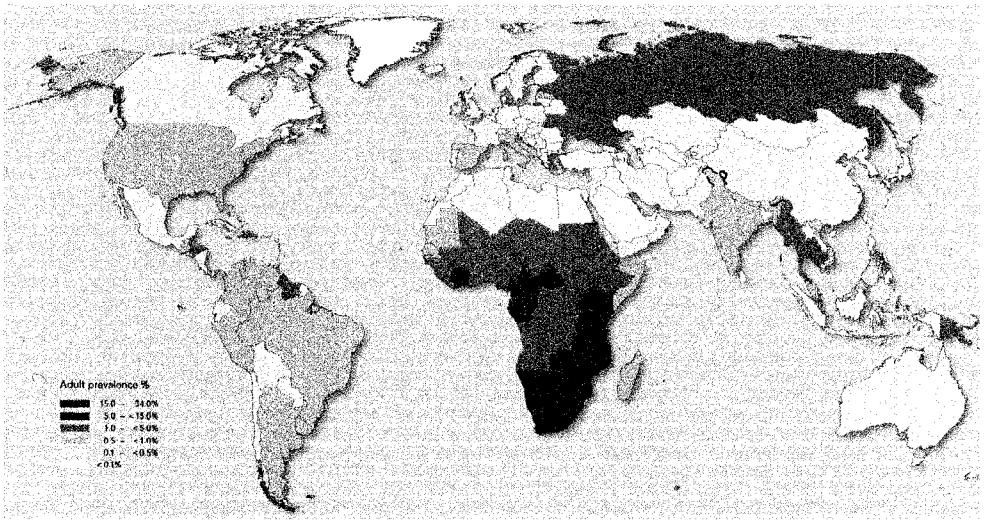


Figure 1.1 Global view of HIV infection. Reproduced by kind permission of UNAIDS (UNAIDS, 2006).

1.3 Human Immunodeficiency Virus Type 1 (HIV-1)

HIV originates from Simian Immunodeficiency Virus (SIV) that infects nonhuman primates (Hahn et al., 2000). It has two forms; HIV-1 and HIV-2. The latter is more prevalent in western Africa and is less virulent than HIV-1, which is responsible for the

global AIDS epidemic. HIV-1 was first discovered in 1983 (Barre-Sinoussi, 1983; Chermann et al., 1983) and has been under intensive study in the past two decades. Even though a vast amount of knowledge has been gained in understanding the life cycle of this complex virus, also contributing to a significant progress in the field of immunology, a cure for AIDS has not yet been attained. The AIDS epidemic has been identified by the World Health Organization (WHO) as the world's most urgent public health challenge (WHO, 2004).

1.3.1 HIV-1 Subtypes

The HIV-1 viruses belong to one of the three phylogenetic groups; M (main), O (outgroup), and N (non-M/non-O) (Ayouba et al., 2000; Simon et al., 1998). Group M can further be divided into 10 different phylogenetic subtypes (A-D, F-K) (Robertson et al., 2000). This group is responsible for most of the HIV-1 infections worldwide, while groups N and O infections are mainly confined to central Africa (Hemelaar et al., 2006). In individuals that are infected by more than one HIV-1 subtype, the reverse transcription process, which is a crucial step in HIV-1 life cycle (see subsection 1.3.5.2), can result in the production of recombinant strains. This is due to the fact that the HIV-1 reverse transcriptase can jump back and forth between the two RNA templates contained in the particle, which can in some cases originate from different HIV-1 subtypes (Temin, 1991). If the resulting recombinant forms are unique, they are referred to as "unique recombinant forms" (URF) (McCutchan, 2006), whereas if they are observed in at least three epidemiologically unlinked people, they are referred to as "circulating recombinant forms" (CRF) (Peeters, 2001).

Global and regional distribution of HIV-1 genetic subtypes and recombinants in 2004 were investigated in a study by WHO and UNAIDS (Hemelaar et al., 2006). In this study, 23,874 HIV-1 samples from 70 counties were collected. The results showed that subtype C is

the most prevalent subtype, accounting for almost 50% of all infections worldwide, and is mainly observed in South Africa and India (Figure 1.2). Subtypes A, B, D, and G account for 12%, 10%, 3%, and 6% of infections, respectively. The HIV-1 infections in North America are primarily by subtype B strains. Other subtypes, such as F, H, J, and K, together account for less that 1% of infections. The study also showed that the most prevalent circulating recombinant forms are CRF01_AE and CRF02_AG, each accounting for about 5% of infections. Other recombinant forms contribute much less to the infection profile but together are responsible for 18% of infections.

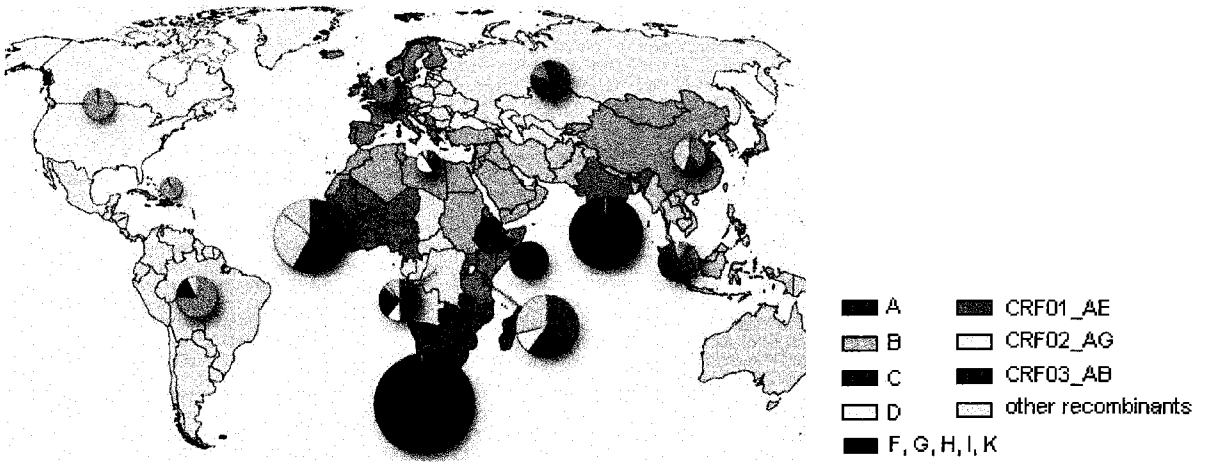


Figure 1.2 Regional distribution of HIV-1 subtypes and recombinants in 2004. Adapted and modified from (Hemelaar et al., 2006). Reproduced by kind permission of UNAIDS.

1.3.2 HIV-1 Infection Life-History

The number of CD4+ cells has been generally accepted as the best predictive measure of the progress of the HIV-1 infection (Stein et al., 1992). After the onset of primary infection, rapid replication of HIV-1 virions causes a swift reduction in the number of CD4+ T-lymphocytes (Figure 1.3) (Ho et al., 1995). This stage, which lasts about two to six weeks, is associated with both a rapid increase in HIV-1 RNA copy numbers (Daar et al., 1991), and acute infection symptoms such as fever, headache, and rash (Saag, 1994). After this initial

stage, the immune system takes HIV-1 replication under control and resumes the CD4+ cell counts, although they never reach the original amount (Stein et al., 1992). This stage is accompanied by a simultaneous reduction in HIV-1 RNA copy numbers. The suppression of HIV-1 replication by the immune system continues for the subsequent 7-10 years, known as the latency period. It has, however, been shown that a high amount of viral replication still persists in this period within the lymphatic tissues (Pantaleo et al., 1993). Over the years of clinical latency period, there is a gradual decrease of CD4+ cells (on average 40-80 cell/mm^3/year) and a simultaneous increase in HIV-1 RNA copy numbers in the plasma of infected people (Figure 1.3) (Saag, 1994).

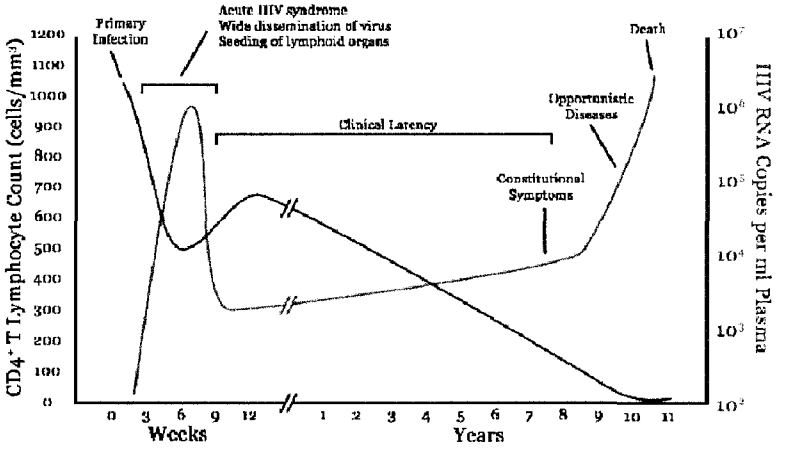


Figure 1.3 HIV-1 course of infection. CD4$^+$ T cell count is shown with darker color and HIV RNA copies per ml of plasma is shown with a lighter color. Adapted from Wikipedia (http://en.wikipedia.org), under the terms of the GNU free documentation license, version 1.2.

AIDS is defined as the stage when the number of CD4+ cells drops below 200 cells/mm^3, and opportunistic infections such as *Pneumocystis carinii* pneumonia (PCP) and tuberculosis (TB) emerge (MacDonell et al., 1990). Even though, without antiviral drugs, the median of the period between primary infection and AIDS-defining stage is 10-12 years, some individuals, known as fast progressors, develop AIDS within 18 months, whereas

others known as slow progressors, do not reach this stage until up to 18 years (Moss and Bacchetti, 1989; Saag, 1994).

Individuals infected by the HIV-1 virus, especially those infected through sexual contact with multiple partners (e.g. prostitution) and drug-users, are likely to be infected with multiple HIV-1 strains after the primary infection (Chohan et al., 2005; Smith et al., 2004; van der Kuyl et al., 2005), which is referred to as superinfection. In an investigation of the incidence of HIV-1 superinfection in a cohort of 36 high-risk women in Mombasa, Kenya, seven cases of HIV-1 superinfection were observed corresponding to an incidence of 3.7% per person per year (Piantadosi et al., 2007).

1.3.3 HIV-1 Genome

HIV-1 contains two genomic RNA molecules per virion. Each strand codes a total of three structural proteins: matrix (MA), capsid (CA), nucleocapsid (NC), two envelope glycoproteins; gp41 and gp120, six accessory and regulatory proteins; Nef, Rev, Tat, Vpr, Vpu, and Vif, and three enzymes: protease (PR), reverse transcriptase (RT), and integrase (IN) (Frankel and Young, 1998; Turner and Summers, 1999). The functions of these proteins are indicated in Table 1-1. More detailed description of these proteins is presented in the HIV-1 life cycle in section 1.3.5.

Among HIV-1 proteins, gp120, PR, RT, and IN, have been under extensive research as suitable targets for antiviral drugs. A large number of these drugs have proven to be potent inhibitors of different stages of virus life cycle. One of the main obstacles, however, in the design and administration of these drugs is the high rate of mutation and genetic variability observed in HIV-1 viruses. For a complete list of current HIV-1 antiviral drugs and associated mutations, see the list compiled by Clark and colleagues (Clark et al., 2005).

Table 1-1 HIV-1 Genes and their function. Adapted and modified from HIV Sequence Compendium 2005 (Leitner et al., 2005).

Name		Size (kD)	Start	End	Function
Gag	MA (Matrix)	p17	790	1186	Membrane anchoring; Env interaction; nuclear transport of viral core (myristoylated protein)
	CA (Capsid)	p24	1186	1879	Core capsid
	NC (Nucleoacpsid)	p7	1921	2086	Nucleocapsid, binds RNA
		p6	2134	2292	Binds Vpr
Pol	PR (Protease)	p10	2253	2550	Gag/Pol cleavage and maturation
	RT (Reverse Transcriptase)	p51	2550	3870	Reverse Transcription
	RNase H	p15	3870	4230	RNase H activity
	IN (Integrase)	p31	4230	5096	DNA provirus integration
Env	Gp120	gp120	6225	7758	External viral glycoprotein bind to CD4 and secondary receptors
	Gp41	gp41	7758	8795	External viral glycoprotein, anchors gp120 to viral membrane
Regulatory	Tat	p16 2 exons	5831 8379	6045 8469	Viral transcriptional transactivator
	Rev	p19 2 exons	5970 8379	6045 8653	RNA transport, stability and utilization factor (phosphoprotein)
	Vif	p23	5041	5619	Promotes virion maturation and infectivity
	Vpr	p10-15	5559	5850	Promotes nuclear localization of preintegration complex, inhibits cell division, arrests infected cells at G2/M
	Vpu	p16	6062	6310	Promotes extracellular release of viral particles; degrades CD4 in the ER
	Nef	p25-27	8797	9417	CD4 and class I downregulation (myristoylated protein)

Figure 1.4 shows the genomic organization of HIV-1. At each end of the genome, there are identical long terminal repeats (LTR) (see section 1.3.5.2). The *gag* and *pol* open reading frames (ORF), which span the first half of the genome, code for the aforementioned structural proteins and enzymes, respectively. The second half of the genome contains the *env* ORF, which codes for the gp120 and gp41, and the ORFs for the six regulatory proteins. HIV-1 has a complex genome with overlapping ORFs in different reading frames.

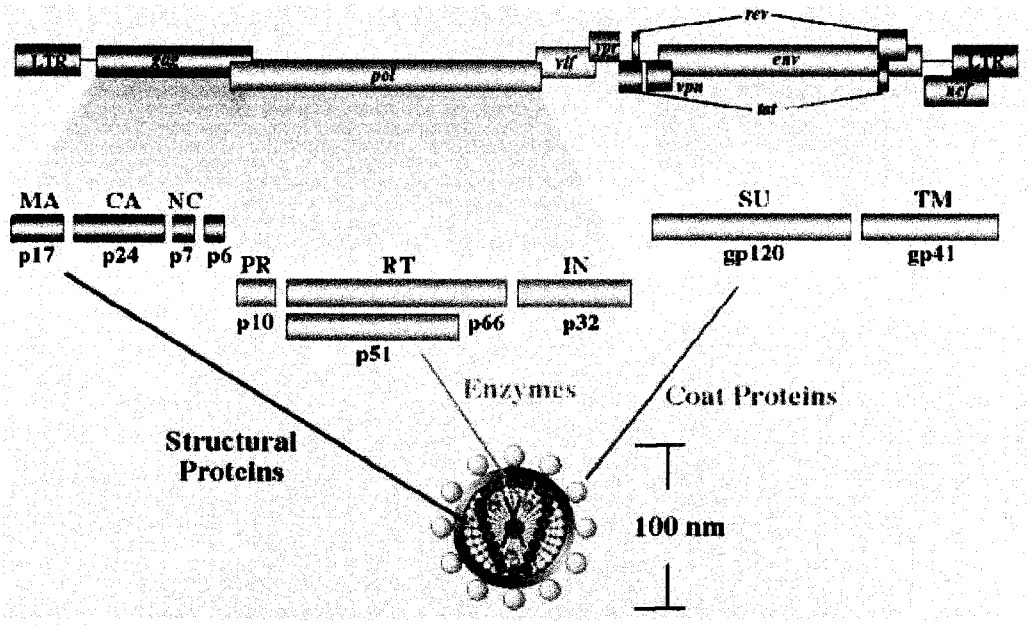


Figure 1.4 Genomic Organization of HIV-1. Adapted with permission from (http://www.stanford.edu/group/virus/retro/2005gongishmail/HIV-1b.jpg)

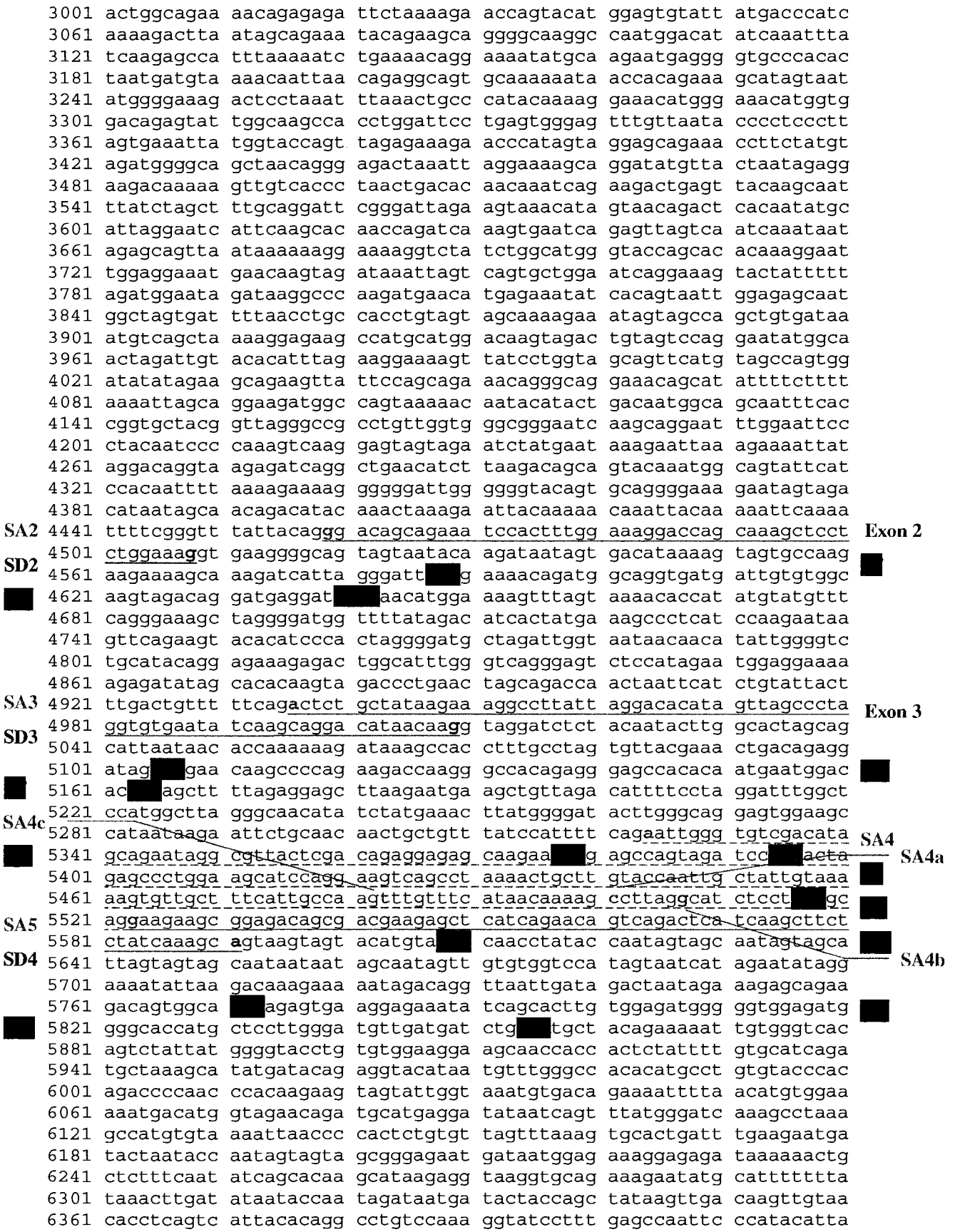
Figure 1.5 shows the sequence of NCBI HIV-1 type genome (NC_001802). Different exons, and splice acceptor (SA) and donor (SD) sites are shown, which are explained in HIV-1 transcription process (section 1.3.5.4). Also shown are the translation initiation and termination codons of the eight ORFs with independent translation initiation sites; *gag*, *vif*, *vpr*, *tat*, *rev*, *vpu*, *env*, and *nef*. The NC_001802 genome is composed of 9181 nucleotides and is A-rich and C-poor (36% A, 18% C, 24% G, and 22% T), typical of all HIV-1 viruses.

Figure 1.5 The complete genome of NCBI HIV-1 Type sequence (NC_001802). Translation initiation codons (highlighted in green) and termination codons (highlighted in light blue) are shown for different HIV-1 genes. Splice acceptor (SA) and donor (SD) sites are shown in bold in orange and blue font colors, respectively. Different exons are underlined in red color, with exons generated by SA4, SA4c, SA4a, and SA4b splice sites shown by discontinuous underlines. Annotations are included in the left and right margins of the graph with matching colors.

```
   1 ggtctctctg gttagaccag atctgagcct gggagctctc tggctaacta gggaacccac
  61 tgcttaagcc tcaataaagc ttgccttgag tgcttcaagt agtgtgtgcc cgtctgttgt
 121 gtgactctgg taactagaga tccctcagac cctttttagtc agtgtggaaa atctctagca
 181 gtggcgcccg aacagggacc tgaaagcgaa agggaaacca gaggagctct ctcgacgcag
 241 gactcggctt gctgaagcgc gcacggcaag aggcgagggg cggcgactgg tgagtacgcc
 301 aaaaattttg actagcggag gctagaagga gagag[redacted]gg tgcgagagcg tcagtattaa
 361 gcgggggaga attagatcga tgggaaaaaa ttcggttaag gccaggggga aagaaaaaat
 421 ataaattaaa acatatagta tgggcaagca gggagctaga acgattcgca gttaatcctg
 481 gcctgttaga aacatcagaa ggctgtagac aaatactggg acagctacaa ccatcccttc
 541 agacaggatc agaagaactt agatcattat ataatacagt agcaaccctc tattgtgtgc
 601 atcaaaggat agagataaaa gacaccaagg aagctttaga caagatagag gaagagcaaa
 661 acaaaagtaa gaaaaaagca cagcaagcag cagctgacac aggacacagc aatcaggtca
 721 gccaaaatta ccctatagtg cagaacatcc aggggcaaat ggtacatcag gccatatcac
 781 ctagaacttt aaatgcatgg gtaaaagtag tagaagagaa ggctttcagc ccagaagtga
 841 tacccatgtt ttcagcatta tcagaaggag ccaccccaca agatttaaac accatgctaa
 901 acacagtggg gggacatcaa gcagccatgc aaatgttaaa agagaccatc aatgaggaag
 961 ctgcagaatg ggatagagtg catccagtgc atgcagggcc tattgcacca ggccagatga
1021 gagaaccaag gggaagtgac atagcaggaa ctactagtac ccttcaggaa caaataggat
1081 ggatgacaaa taatccacct atcccagtag gagaaattta taaaagatgg ataatcctgg
1141 gattaaataa aatagtaaga atgtatagcc ctaccagcat tctggacata agacaaggac
1201 caaaggaacc ctttagagac tatgtagacc ggttctataa aactctaaga gccgagcaag
1261 cttcacagga ggtaaaaaat tggatgacag aaaccttgtt ggtccaaaat gcgaacccag
1321 attgtaagac tattttaaaa gcattgggac cagcggctac actagaagaa atgatgacag
1381 catgtcaggg agtaggagga cccggccata aggcaagagt tttggctgaa gcaatgagcc
1441 aagtaacaaa ttcagctacc ataatgatgc agagaggcaa ttttaggaac caaagaaaga
1501 ttgttaagtg tttcaattgt ggcaaagaag ggcacacagc cagaaattgc agggccccta
1561 ggaaaaaggg ctgttggaaa tgtggaaagg aaggacacca aatgaaagat tgtactgaga
1621 gacaggctaa ttttttaggg aagatctggc cttcctacaa gggaaggcca gggaattttc
1681 ttcagagcag accagagcca acagccccac cagaagagag cttcaggtct ggggtagaga
1741 caacaactcc ccctcagaag caggagccga tagacaagga actgtatcct ttaacttccc
1801 tcaggtcact ctttggcaac gacccctcgt cacaataaag atagggggggc aactaaagga
1861 agctctatta gatacaggag cagatgatac agtattagaa gaaatgagtt tgccaggaag
1921 atggaaacca aaaatgatag ggggaattgg aggttttatc aaagtaagac agtatgatca
1981 gatactcata gaaatctgtg gacataaagc tataggtaca gtattagtag gacctacacc
2041 tgtcaacata attggaagaa atctgttgac tcagattggt tgcactttaa attttcccat
2101 tagccctatt gagactgtac cagtaaaatt aaagccagga atggatggcc caaaagttaa
2161 acaatggcca ttgacagaag aaaaaataaa agcattagta gaaatttgta cagagatgga
2221 aaaggaaggg aaaatttcaa aaattgggcc tgaaaatcca tacaatactc cagtatttgc
2281 cataaagaaa aaagacagta ctaaatggag aaaattagta gatttcagag aacttaataa
2341 gagaactcaa gacttctggg aagttcaatt aggaatacca catcccgcag ggttaaaaaa
2401 gaaaaaatca gtaacagtac tggatgtggg tgatgcatat ttttcagttc ccttagatga
2461 agacttcagg aagtatactg catttaccat acctagtata aacaatgaga caccagggat
2521 tagatatcag tacaatgtgc ttccacaggg atggaaagga tcaccagcaa tattccaaag
2581 tagcatgaca aaaatcttag agccttttag aaaacaaaat ccagacatag ttatctatca
2641 atacatggat gatttgtatg taggatctga cttagaaata gggcagcata gaacaaaaat
2701 agaggagctg agacaacatc tgttgaggtg gggacttacc acaccagaca aaaaacatca
2761 gaaagaacct ccattccttt ggatgggtta tgaactccat cctgataaat ggacagtaca
2821 gcctatagtg ctgccagaaa aagacagctg gactgtcaat gacatacaga agttagtggg
2881 gaaattgaat tgggcaagtc agatttaccc agggattaaa gtaaggcaat tatgtaaact
2941 ccttagagga accaaagcac taacagaagt aataccacta acagaagaag cagagctaga
```

Exon 1

SD1


SA2
SD2
SA3
SD3
SA4c
SA5
SD4
Exon 2
Exon 3
SA4
SA4a
SA4b

6421 ttgtgccccg gctggttttg cgattctaaa atgtaataat aagacgttca atggaacagg
6481 accatgtaca aatgtcagca cagtacaatg tacacatgga attaggccag tagtatcaac
6541 tcaactgctg ttaaatggca gtctagcaga agaagaggta gtaattagat ctgtcaattt
6601 cacggacaat gctaaaacca taatagtaca gctgaacaca tctgtagaaa ttaattgtac
6661 aagacccaac aacaatacaa gaaaaagaat ccgtatccag agaggaccag ggagagcatt
6721 tgttacaata ggaaaaatag gaaatatgag acaagcacat tgtaacatta gtagagcaaa
6781 atggaataac actttaaaac agatagctag caaattaaga gaacaatttg gaaataataa
6841 aacaataatc tttaagcaat cctcaggagg ggacccagaa attgtaacgc acagttttaa
6901 ttgtggaggg gaatttttct actgtaattc aacacaactg tttaatagta cttggtttaa
6961 tagtacttgg agtactgaag ggtcaaataa cactgaagga agtgacacaa tcaccctccc
7021 atgcagaata aaacaaatta taaacatgtg gcagaaagta ggaaaagcaa tgtatgcccc
7081 tcccatcagt ggacaaatta gatgttcatc aaatattaca gggctgctat taacaagaga
7141 tggtggtaat agcaacaatg agtccgagat cttcagacct ggaggaggag atatgaggga
7201 caattggaga agtgaattat ataaatataa agtagtaaaa attgaaccat taggagtagc
7261 acccaccaag gcaaagagaa gagtggtgca gagagaaaaa agagcagtgg gaataggagc
7321 tttgttcctt gggttcttgg gagcagcagg aagcactatg ggcgcagcct caatgacgct
7381 gacggtacag gccagacaat tattgtctgg tatagtgcag cagcagaaca atttgctgag
7441 ggctattgag gcgcaacagc atctgttgca actcacagtc tggggcatca agcagctcca
7501 ggcaagaatc ctggctgtgg aaagatacct aaaggatcaa cagctcctgg ggatttgggg
7561 ttgctctgga aaactcattt gcaccactgc tgtgccttgg aatgctagtt ggagtaataa
7621 atctctggaa cagatttgga atcacacgac ctggatggag tgggacagag aaattaacaa
7681 ttacacaagc ttaatacact ccttaattga agaatcgcaa aaccagcaag aaaagaatga
7741 acaagaatta ttggaattag ataaatgggc aagtttgtgg aattggttta acataacaaa
7801 ttggctgtgg tatataaaat tattcataat gatagtagga ggcttggtag gtttaagaat
7861 agtttttgct gtactttcta tagtgaatag agttaggcag ggatattcac cattatcgtt
7921 tcagacccac ctcccaaccc cgaggggacc cgacaggccc gaaggaa[redacted] aagaagaagg **Exon 7**
7981 tggagagaga gacagagaca gatccattcg attagtgaac ggatccttgg cacttatctg
8041 ggacgatctg cggagcctgt gcctcttcag ctaccaccgc ttgagagact tactcttgat
8101 tgtaacgagg attgtggaac ttctgggacg caggggtgg gaagccctca aatattggtg
8161 gaatctccta cagtattgga gtcaggaact aaagaa[redacted]t gctgttagct tgctcaatgc
8221 cacagccata gcagtagctg aggggacaga tagggttata gaagtagtac aaggagcttg
8281 tagagctatt cgccacatac ctagaagaat aagacagggc ttggaaagga ttttgcta[redacted]
8341 [redacted]g[redacted]ggtgg caagtggtca aaaagtagtg tgattggatg gcctactgta agggaaagaa [redacted]
8401 tgagacgagc tgagccagca gcagataggg tgggagcagc atctcgagac ctggaaaaac
8461 atggagcaat cacaagtagc aatacagcag ctaccaatgc tgcttgtgcc tggctagaag
8521 cacaagagga ggaggaggtg ggttttccag tcacacctca ggtaccttta agaccaatga
8581 cttacaaggc agctgtagat cttagccact ttttaaaaga aaaggggga ctggaagggc
8641 taattcactc ccaaagaaga caagatatcc ttgatctgtg gatctaccac acacaaggct
8701 acttccctga ttagcagaac tacacaccag ggccaggggt cagatatcca ctgacctttg
8761 gatggtgcta caagctagta ccagttgagc cagataagat agaagaggcc aataaaggag
8821 agaacaccag cttgttacac cctgtgagcc tgcatgggat ggatgacccg gagagagaag
8881 tgttagagtg gaggtttgac agccgcctag catttcatca cgtggcccga gagctgcatc
8941 cggagtactt caagaactgc [redacted]catcgag cttgctacaa gggactttcc gctggggact
9001 ttccagggag gcgtggcctg ggcgggactg gggagtggcg agccctcaga tcctgcatat
9061 aagcagctgc tttttgcctg tactgggtct ctctggttag accagatctg agcctgggag
9121 ctctctggct aactagggaa cccactgctt aagcctcaat aaagcttgcc ttgagtgctt
9181 c

1.3.4 HIV-1 Particle

The HIV-1 particle is approximately 100nm in diameter (Figure 1.6). The outermost membrane of the particle originates from the host lipid bilayer membrane which surrounds the particle in one of the final steps of the virus replication cycle (section 1.3.5.6). The two glycoproteins, gp120 and gp41, are non-covalently bound to each other on the surface of the particle. Gp120, also known as the surface (SU) protein, comprises the extracellular portion of the viral envelope, and the gp41, also known as the transmembrane (TM) protein, spans the lipid bilayer. Under the lipid bilayer, is a membrane formed by approximately 1200 matrix proteins (Layne et al., 1992). In the mature form of the viral particle (section 1.3.5.7), an approximately equal number of capsid proteins form a cone-shaped capsid envelope, which is a characteristic of retroviruses.

Each particle contains two, in most cases identical, genomic RNAs packaged in the capsid structure. These single-stranded positive-sense RNAs are covered by NC proteins (Figure 1.6), which have been shown to act as chaperones by refolding nucleic acid molecules into the most energetically favorable conformation (Rein et al., 1998). The capsid structure also contains multiple copies of viral enzymes including reverse transcriptase, protease, and integrase. These enzymes have important roles in different stages of the virus life cycle (section 1.3.5). Additional molecules are either actively or passively packaged in the virion at the site of particle formation. These include, among other, host proteins (Arthur et al., 1992), $tRNA^{Lys3}$ molecules used as primers for reverse transcription of viral RNAs (section 1.3.5.2), and other HIV-1 proteins such as Vpr , p6, and Nef (Lu et al., 1993; Welker et al., 1996).

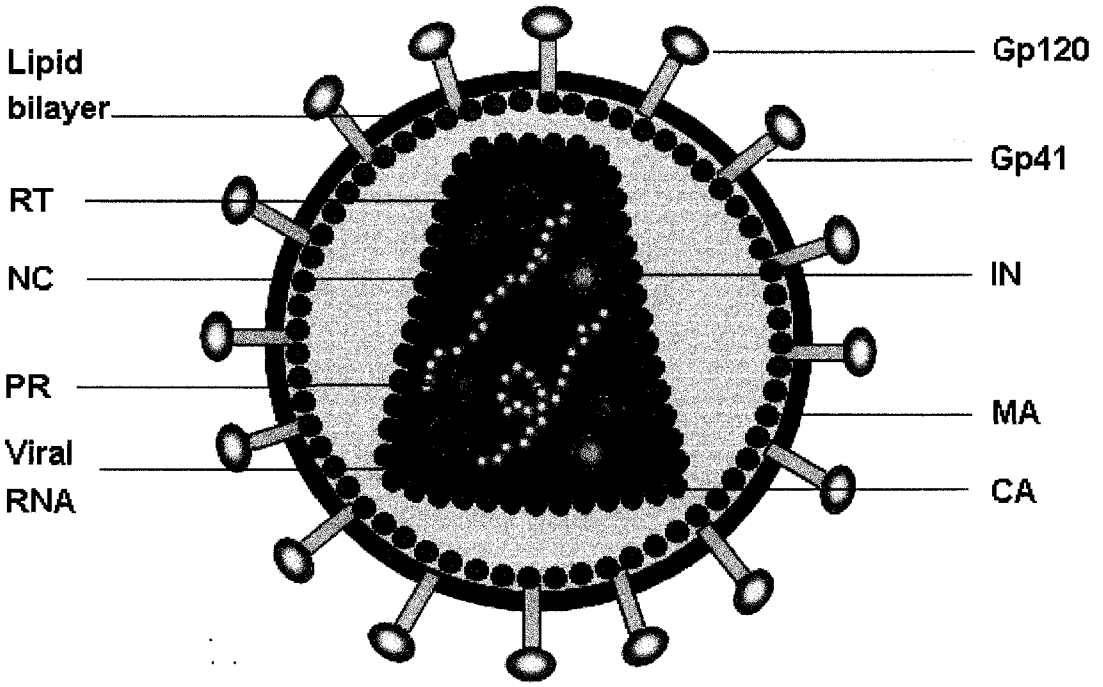


Figure 1.6 A Schematic diagram of mature HIV-1 particle. RT: reverse transcriptase, NC: nucleocapsid, PR: protease, IN: integrase, MA: matrix protein, CA: capsid protein.

Measuring the relative abundance of proteins in HIV-1 particles is beneficial to the studies of translation efficiency of different HIV-1 genes, which is one of the subjects of this thesis. The quantification of these proteins within a single virion particle, however, is difficult and the results obtained by different laboratories are sometimes contradictory (Liu et al., 1995; Zombek et al., 1994). It is estimated that the HIV-1 virus particle contains about 1,200 molecules of CA, MA, and NC, 80 molecules of RT and IN, 5-10 molecules of Nef, and about 280 molecules of gp120 and gp41 (Layne et al., 1992; Liu et al., 1995; Welker et al., 1996; Zombek et al., 1994).

1.3.5 HIV-1 Life Cycle

The life cycle of HIV-1 virus can be divided into the early and late phases (Cann and Karn, 1989). The early phase starts with the entry of the virus into the host cell followed by the reverse transcription of the viral RNA into a double stranded DNA, and the integration of

this proviral DNA into the host chromosome. The late phase includes transcription and translation of the viral genes, assembly of the viral components at the cell membrane, and the final release of the newly-formed virions which will go on to infect other target cells.

Even though HIV-1 viruses are able to enter both activated and resting CD4+ T cells, they replicate preferentially in the former group (Margolick et al., 1987). The latter group, as previously mentioned, is mainly used as latent reservoirs for HIV-1 (Blankson et al., 2002; Chun et al., 1998; Chun and Fauci, 1999). Most of the CD4+ T cells in an individual are in the resting quiescent state, which do not permit efficient virus replication at different stages including cell entry (section 1.3.5.1) (Pierson et al., 2000), reverse transcription (section 1.3.5.2) (Pierson et al., 2002), and integration of proviral DNA (section 1.3.5.3) (Bukrinsky et al., 1992).

1.3.5.1 Cell Entry

Through the sexual transmission of the HIV-1 virus, the immature dendritic cells (iDC) are the most frequent molecules that are infected in the mucosal tissue. The viral particles are then transmitted to the CD4+ T cells and macrophages (Ho et al., 1986; Steinman et al., 2003). These cells contain the CD4 receptors on their surface, which is required for the virus entry (Dalgleish et al., 1984; Maddon et al., 1986). The physiological function of the CD4 receptors is to interact with the major histocompatibility (MHC) class II molecules on the surface of antigen-representing cells, which is one of the steps in the onset of the immune response (Nag et al., 1993; Robey and Axel, 1990).

The HIV-1 cell entry starts with the binding of the gp120 to the N-terminal domain of the CD4 receptors present on the surface of the host target cells (Figure 1.7) (Dalgleish et al., 1984; McDougal et al., 1986). After this binding, the membrane of the virus and host cell

fuse through conformational changes in gp120 and gp41 (Sakai et al., 1993; Stein et al., 1987). In addition to CD4, the presence of specific chemokine coreceptors (a family of seven transmembrane G-coupled proteins), CCR5 and CXCR4, is required for the completion of membrane fusion (Berger et al., 1999; Doms and Peiper, 1997). The physiological function of these coreceptors is to mediate the immune cell response through interaction with a subset of soluble chemo-attractant molecules, called chemokines (Horuk, 1994).

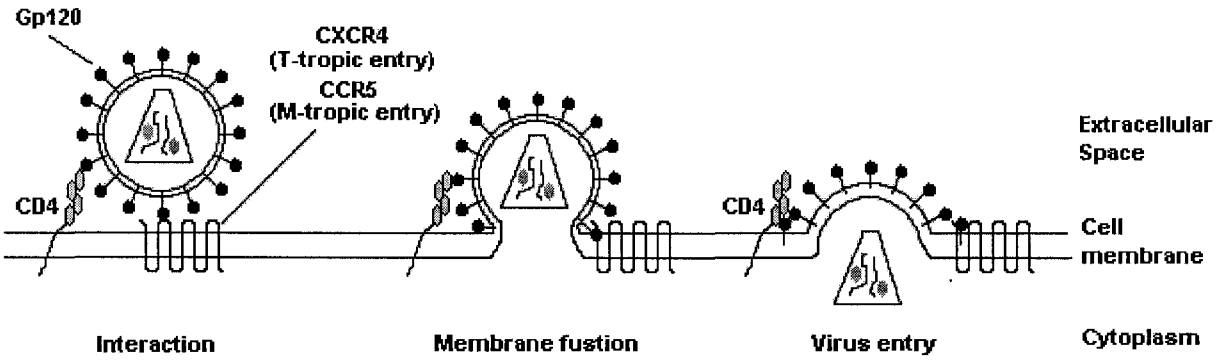


Figure 1.7 HIV-1 Virus entry into CD4+ T-cells. Interaction of gp120 on the surface of the virus with CD4 receptor in the presence of a coreceptor, CXCR4 or CCR5, induces a membrane fusion followed by the injection of viral particle contents into the cell cytoplasm.

1.3.5.2 Reverse Transcription

Upon cell entry, HIV-1 genomic RNA has to undergo reverse transcription to generate a double-stranded DNA, also known as the provirus, before it can be integrated into the host chromosome. The reverse transcription process in retroviruses, including the HIV-1 virus, is a detailed and complex mechanism including two unique strand-transfers (Gilboa et al., 1979; Panganiban and Fiore, 1988; Peliska and Benkovic, 1992). The end product of reverse transcription is a complete double-stranded DNA having identical U3-R-U5 elements (LTRs) at the two termini.

1.3.5.3 Integration

Following the reverse transcription of the viral genomic RNA, the resulting double-stranded DNA is transferred to the nucleus for integration into the chromosome. The viral integrase protein, packaged in the virus particle entering the cell, mediates this integration process. The complex consisting of the viral and cellular proteins that are transported to the nucleus and take part in the integration process, is called the pre-integration complex (PIC). Even though the integration sites are widespread in the human genome, gene-containing regions are strongly favored as integration acceptor sites (Schroder et al., 2002), and in general, nucleosome-free genomic regions which lack DNA binding proteins, such as histones, are preferred (Pryciak and Varmus, 1992).

1.3.5.4 Transcription

After the integration of the proviral DNA into the human chromosome, it functions like any other human gene and undergoes transcription and translation using cellular polymerase and ribosomes. The 5' LTR of the integrated provirus, specifically the U3 region (Figure 1.8), functions as the promoter for the transcription of viral genes. It contains the binding sites for at least 12 different *cis*-acting cellular transcription factors including three activator proteins (AP)-1, nuclear factor (NF)-AT, upstream stimulatory factor-1 (USF-1), two nuclear factors (NF)-κB, three SP-1, and a TATA box (Figure 1.8) (Gaynor, 1991; Gaynor, 1992; Harrich et al., 1989; Jones and Peterlin, 1994; Peterlin, 1991; Waterman et al., 1991).

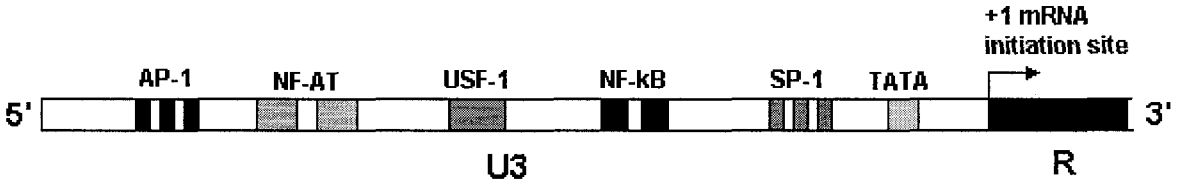


Figure 1.8 U3 region of HIV-1 LTR. Various transcriptional promoter elements are shown; activator protein-1 (AP-1); nuclear factor of activated T cells (NF-AT); upstream stimulatory factor-1 (USF-1); nuclear factor kappa B (NF-kB); SP-1, and TATA box. Figure adapted and modified from (Morrow et al., 1994).

More than 30 different mRNAs are produced from alternative splicing of the full-length HIV-1 transcript which are divided into three different classes; unspliced 9kb mRNAs, intermediate 4kb mRNAs, and short 2kb mRNAs (Table 1-2) (Purcell and Martin, 1993; Schwartz et al., 1990). This alternative splicing of HIV-1 transcripts has been suggested to modulate viral protein expression, replication, and infectivity (Purcell and Martin, 1993). In the early stages after integration, low basal transcription levels (Figure 1.9), modulated by the transcription factor binding sites (TFBS) in the 5' LTR, give rise to a family of more than 12 alternatively spliced small 2-kb mRNAs coding for Tat, Rev, and Nef (Arya and Gallo, 1986; Guatelli et al., 1990; Schwartz et al., 1990).

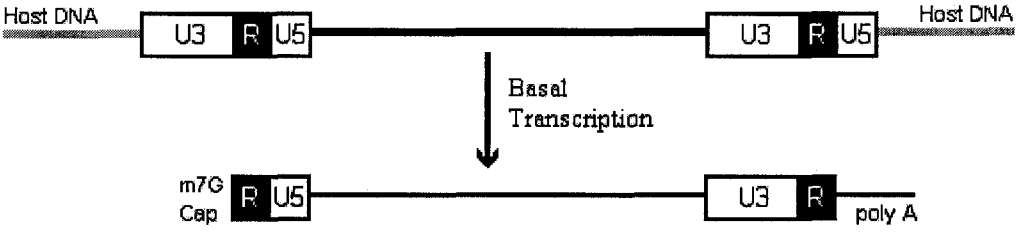


Figure 1.9 Basal transcription of HIV-1 genome from integrated provirus.

In addition to the TFBSs, a stable secondary structure called transactivating response region (TAR) located in the R segment of the 5' LTR, plays an important role in the expression of HIV-1 mRNAs. After translation of its mRNA in cytoplasm, the HIV-1 Tat

protein is transported back to the nucleus where it binds to TAR structure and increases the steady-state levels of all HIV-1 mRNAs (Berkhout et al., 1989; Cullen, 1986; Hauber et al., 1987; Rice and Mathews, 1988). While the absence of Tat results in the production of prematurely terminated transcripts, its presence results in a 100-fold increase in the production of full-length primary transcripts (Dayton et al., 1986).

The regulator of viral expression (Rev) is another viral protein which is synthesized in the early stages after provirus integration and is localized in the nuclei of infected cells. Rev binds to rev-responsive element (RRE), located within the Env ORF, and promotes the transport of RRE-containing mRNAs (i.e. unspliced and 4kb mRNAs) out of the nucleus via the CRM-1 export pathway (Dayton et al., 1988; Fischer et al., 1994; Hadzopoulou-Cladaras et al., 1989). Rev contains a leucine-rich nuclear export signal (NES) which allows it to shuttle between the nucleus and the cytoplasm (Meyer and Malim, 1994). As the level of Rev in the nucleus increases, it increases the amount of unspliced and 4kb mRNAs since it inhibits the splicing of the mRNAs containing the RRE element by interrupting the spliceosome assembly (Kjems and Sharp, 1993). In the early stages of infection, however, when Rev levels are low, a high number of multiply-spliced 2kb mRNAs lacking RRE are produced (Hadzopoulou-Cladaras et al., 1989).

Splicing of the HIV-1 mRNAs is a complex mechanism due to the presence of constitutive and alternative 5' splice donor (SD) and 3' splice acceptor (SA) sites. There are several weak competing splice sites in the middle of the HIV-1 genome that are alternatively selected to produce the mature HIV-1 mRNA (Figure 1.10B and 1.5) (Arrigo et al., 1990; Furtado et al., 1991; Guatelli et al., 1990).

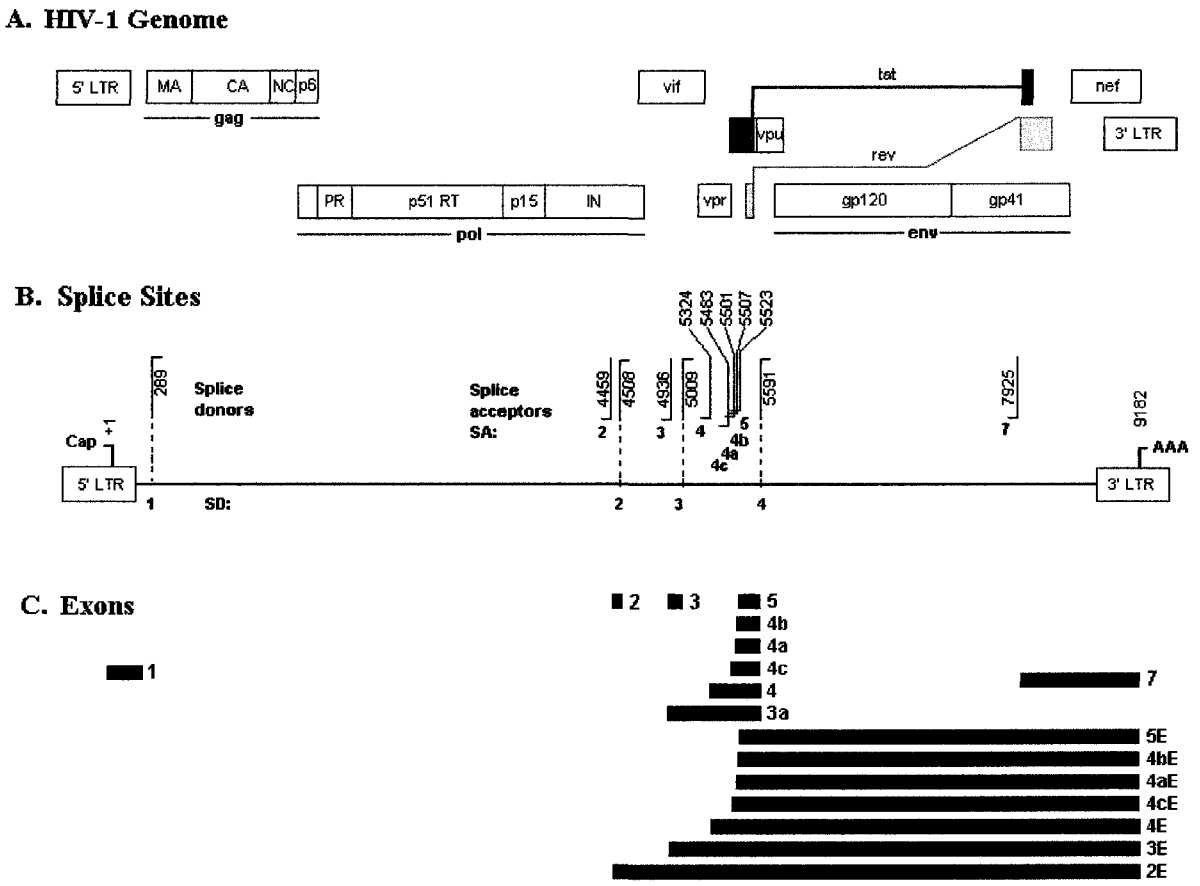


Figure 1.10 HIV-1 exons and splice sites. (A) HIV-1 genome and different open reading frames (ORF). The vertical lines inside *gag/pol* and *env* ORFs indicate the site of viral and cellular protease cleavage, respectively. *tat* and *rev* ORFs, made up of two coding exons, are shown in darker shades. Adapted and modified from Leitner *et al.* (Leitner et al., 2005). (B) Location of the splice donor (SD, with a L shape flipped vertically) and acceptor (SA, with a L shape flipped horizontally) sites in the NCBI HIV-1 type genome (NCBI Genome: NC_001802), corresponding to the numbers done by Schwartz *et al.* (Schwartz et al., 1990). The first position of the 5'-R region is designated as +1 (see text for more details). (C) Various HIV-1 exons generated by alternative use of splice sites found in Purcell and Martin report (Purcell and Martin, 1993), and numbered according to Muesing *et al.* (Muesing et al., 1985). Figures B and C, adapted and modified from Purcell and Martin (Purcell and Martin, 1993).

There are four SD and eight SA sites in the HIV-1 genome, which result in the production of 17 different exons (Figure 1.5, 1.10B and C). Alternative combinations of these exons, which produce different HIV-1 transcripts, are shown in table 1-2. Exon 1, which contains the TAR element, and exon 7 are shared among all transcripts. Also the non-coding exons 2 and 3 are shared among the majority of transcripts.

The alternative use of splice sites is observed in many HIV-1 strains of diverse origins (Robert-Guroff et al., 1990; Smith et al., 1992), which suggests a regulatory role for the resulting RNA isoforms in different stages of virus life cycle, rather than an inherent redundancy in viral RNA processing (Purcell and Martin, 1993). For example, induced alteration in the use of 4, 4a, 4b, and 4c splice acceptor sites resulted in dramatic changes in the level of gp160/gp120 production (Purcell and Martin, 1993).

Table 1-2 Different HIV-1 mRNAs generated from different combinations of exons. Exon numbers and combinations are according to figure 1(C) and adapted from Purcell and Martin, 1993.

2 kb class of RNA		4 kb class of RNA	
mRNA	**Exons**	**mRNA**	**Exons**
Tat 1	1 / 4 / 7	Vpr 3	1 / 3E
Tat 2	1 / 2 / 4 / 7	Vpr 4	1 / 2 / 3E
Tat 3	1 / 3 / 4 / 7	Tat 5	1 / 4E
Tat 4	1 / 2 / 3 / 4 / 7	Tat 6	1 / 2 / 4E
Rev 1	1 / 4b / 7	Tat 7	1 / 3 / 4E
Rev 2	1 / 4a / 7	Tat 8	1 / 2 / 3 / 4E
Rev 3	1 / 4c / 7	Env 1	1 / 5E
Rev 4	1 / 2 / 4b / 7	Env 2	1 / 4bE
Rev 5	1 / 2 / 4a / 7	Env 3	1 / 4aE
Rev 6	1 / 2 / 4c / 7	Env 4	1 / 4cE
Rev 7	1 / 3 / 4b / 7	Env 5	1 / 2 / 5E
Rev 8	1 / 3 / 4a / 7	Env 6	1 / 2 / 4bE
Rev 9	1 / 3 / 4c / 7	Env 7	1 / 2 / 4aE
Rev 10	1 / 2 / 3 / 4b / 7	Env 8	1 / 3 / 5E
Rev 11	1 / 2 / 3 / 4a / 7	Env 9	1 / 2 / 4cE
Rev 12	1 / 2 / 3 / 4c / 7	Env 10	1 / 3 / 4bE
Nef 1	1 / 7	Env 11	1 / 3 / 4aE
Nef 2	1 / 5 / 7	Env 12	1 / 3 / 4cE
Nef 3	1 / 2 / 5 / 7	Env 13	1 / 2 / 3 / 5E
Nef 4	1 / 3 / 5 / 7	Env 14	1 / 2 / 3 / 4bE
Nef 5	1 / 2 / 3 / 5 / 7	Env 15	1 / 2 / 3 / 4aE
Vpr 1	1 / 3a / 7	Env 16	1 / 2 / 3 / 4cE
Vpr 2	1 / 2 / 3a / 7	Vif 2	1 / 2E

1.3.5.5 Translation

The various forms of the HIV-1 mRNAs are transferred to the cytoplasm where they undergo translation by the host ribosomes. As opposed to cellular mRNAs, which are mostly

monocistronic, HIV-1 mRNAs are usually poly-cistronic. For example, *vpu* and *env* ORFs always appear together on the same mRNA. The HIV-1 structural and enzymatic proteins are translated in the form of Gag and Pol polyproteins (Figure 1.4), respectively, which are later cleaved into individual components by the action of HIV-1 protease in the virion maturation process (section 1.3.5.7). These polyproteins are generated from an unspliced 9kb mRNA. Translation of Pol polyprotein occurs via a -1 ribosomal frameshifting at a "slippery" hepta-uridine stretch in the 3' end of NC (Figure 1.4), thereby avoiding the translation termination signal at the end of the Gag gene (Jacks et al., 1988). This frameshift occurs at a frequency of 5-10%, which explains the relative abundance of Gag and Gag-Pol polyproteins (Cassan et al., 1994). The gp160 polyprotein is produced by the translation of a singly-spliced 4kb Env mRNA (Table 1-2), from which gp120 and gp41 are generated through the action of a cellular protease, furin (Allan et al., 1985).

Different hypotheses have been postulated regarding the mechanism of translation initiation in HIV-1 (for a review see Yilmaz *et al.* (Yilmaz et al., 2006)). Schwartz *et al.* (Schwartz et al., 1992) used a panel of cDNA clones corresponding to alternatively spliced HIV-1 mRNAs and investigated the translation mechanism of five HIV-1 genes; *tat*, *rev*, *vpu*, *env*, and *nef* in transfected human cells . They showed that *rev* and *nef* ORFs, which are positioned downstream of *tat* ORF, are not efficiently translated which is attributed to the strong *tat* Kozak (Kozak, 1986b) sequence. This was confirmed by mutating the *tat* Kozak sequence into a weak signal which resulted in the efficient translation of downstream ORFs. On the other hand, weak *rev* and *vpu* Kozak sequences allowed the efficient translation of downstream *nef* and *env* ORFs, respectively, through a mechanism referred to as "leaky" scanning (Kozak, 1986a; Kozak, 1989). When the region surrounding the translation initiation AUG of *rev* and *vpu* were mutated to resemble *tat*, the translation of downstream

ORFs were significantly inhibited. Based on these observations the authors concluded that HIV-1 mRNAs are translated via conventional cap-dependent ribosomal scanning mechanism (CDRSM) (see section 2.2) (Figure 1.11).

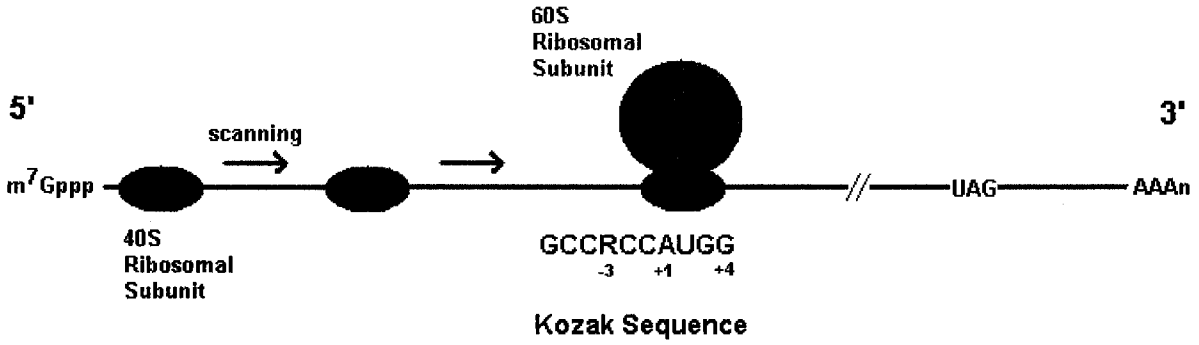


Figure 1.11 Eukaryotic cap-dependent ribosomal scanning mechanism (CDRSM) of translation initiation. The 40S ribosomal subunit, along with the eIF4F translation initiation complex, binds at the 5'cap and starts scanning the mRNA molecule until it reaches an AUG in optimal context (Kozak consensus). The 60S ribosomal subunit then binds to the complex and the translation of the polypeptide starts.

In a different study, Miele and colleagues (Miele et al., 1996), conducted an experiment in which they investigated the role of the HIV-1 packaging signal (Ψ), located in the 5'-UTR of unspliced mRNAs, in the translation initiation of *gag* ORF. They showed that the packaging signal greatly inhibits translation initiation and that this inhibition is overcome by heat denaturation that melts its secondary structure. This result was also consistent with translation initiation via CDRSM. Another line of evidence for CDRSM in HIV-1 comes from a recent study where a mutational analysis revealed that leaky scanning at the *rev* Kozak sequence was responsible for efficient translation of the downstream *vpu* ORF (Anderson et al., 2007).

Despite these observations, there are some obstacles associated with the CDRSM hypothesis in the HIV-1 virus. All HIV-1 transcripts share a 289-nucleotide sequence as part

of their 5'-UTR, which forms into stable stem-loop structures (Baudin et al., 1993; Berkhout and van Wamel, 2000) (Figure 1.13). The 5'-UTR is the most conserved region of the HIV-1 genome (Berkhout, 1996) and contains several functional domains which are important for virus replication. These include transactivation response region (TAR) (Selby et al., 1989), polyadenylation signal (polyA) (Klasens et al., 1999), primer binding site (PBS) (Beerens and Berkhout, 2002), RNA dimmer initiation signal (DIS) (Laughrea and Jette, 1994), the major splice donor site (SD1) (Purcell and Martin, 1993), and the packaging signal (Ψ), which is only present in unspliced mRNAs (Lever et al., 1989). The stable stem-loop structures formed by these domains have been shown to inhibit translation initiation by the scanning ribosome (Geballe and Gray, 1992; Miele et al., 1996; Parkin et al., 1988; SenGupta et al., 1990; Svitkin et al., 1994).

The possibility of alternative cap-independent translation initiation mechanisms in HIV-1 and other viruses was first supported by the observation of internal ribosome entry site (IRES)-mediated translation initiation in picornaviruses (Jang et al., 1988; Pelletier and Sonenberg, 1988) (Figure 1.12). Through this mechanism, the ribosome directly enters at IRES to initiate the translation without having to bind at the 5'cap and scan the viral mRNA. Accumulating evidence point to the presence of IRES-dependent translation initiation in *retroviridae* family as well (Berlioz and Darlix, 1995; Deffaud and Darlix, 2000; Lopez-Lastra et al., 1997; Ohlmann et al., 2000), which suggests that this form of translation initiation might be an important part of the retroviral life cycle. This mechanism would allow bypassing the stable 5' secondary structures of the HIV-1 5'-UTR, alleviating the barriers imposed by CDRSM on translation initiation.

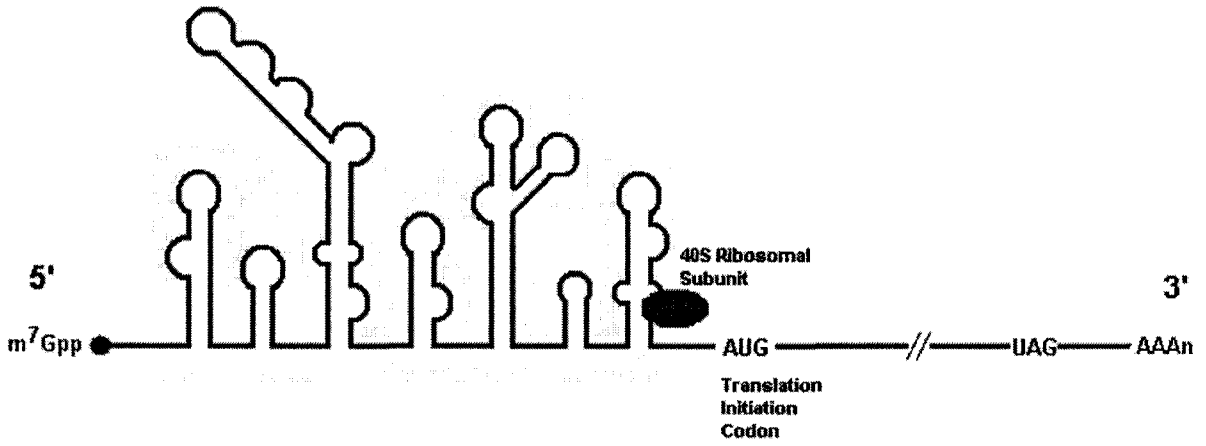


Figure 1.12 Internal Ribosome Entry Site (IRES)-dependent translation initiation mechanism. The 40S ribosomal subunit binds at IRES with the help of IRES Transacting Factors (ITAF). This mechanism is independent of the 5'cap and the 5'-UTR stable secondary structures are bypassed. The depicted secondary structure corresponds to the structure prediction of *vif* shown in (Yilmaz et al., 2006). The IRES location is hypothetical.

1.3.5.6 Viral Particle Formation and Release

In the final stages of the viral life cycle, sufficient amount of enzymes and structural proteins, along with regulatory proteins that are essential for the efficient formation and release of a new infectious particle, are produced. These proteins along with two full-length genomic RNAs accumulate at the cell membrane to form a particle which eventually "buds off' from the cell.

Contrary to the B and D retroviruses, where the viral particle is formed inside the cytoplasm and then released at the cell membrane, in type C retroviruses and lentiviruses including the HIV-1, the viral particle is formed at the cell membrane through the interaction of Gag, Gag-Pol, and Env polyproteins (Dickson et al., 1984). As opposed to simple retroviruses such as murine leukemia virus, where there are separate pools of mRNAs for translation and packaging, in HIV-1 a single pool of full-length mRNA is used for both translation and packaging (Dorman and Lever, 2000; Poon et al., 2002). At high

concentrations Gag binds to its own precursor and impedes its translation, and through this negative feedback mechanism controls the equilibrium between the translation and packaging of the full-length mRNA (Anderson and Lever, 2006). The MA part of the Gag and Gag-pol polyproteins is involved in their accumulation at the site of virion assembly (Yu et al., 1992). This is believed to be induced by the myristoylation (i.e. addition of a myristate fatty acid side chain to a glycine) (Farazi et al., 2001) of the N-terminal glycine of MA (Bouamr et al., 2003; Provitera et al., 2006). The MA protein also mediates the incorporation of gp120 and gp41 into the assembling viral particle (Mammano et al., 1995).

As different components of the virion particle accumulate, the lipid membrane of the host packages these components by forming into a circular configuration (Aloia et al., 1988; Gelderblom, 1991). The p6 part of the Gag polyprotein plays an important role at this stage by both mediating the incorporation of the Vpr protein into the assembling particle, and promoting efficient particle release from the cell (Huang et al., 1995).

1.3.5.7 Viral Particle Maturation

The newly-formed virions released from the cell have to undergo maturation prior to infecting other target cells. The maturation process refers to the cleavage of the Gag and Gag-pol polyproteins by the viral protease (Cann and Karn, 1989). This morphological maturation is essential for the production of infectious particles (De Clercq, 1998; Debouck, 1992; Kohl et al., 1988). If the activity of the protease is inhibited, non-infectious particles unable to replicate are formed (Loeb et al., 1989; Manchester et al., 1994; Park and Morrow, 1993). After being cleaved by the protease, the MA and CA components of the Gag precursors multimerize to form the outer membrane and the cone-shaped capsid structure,

respectively (Gelderblom et al., 1987), that enclose the two viral RNA molecules covered by the NC proteins (Aldovini and Young, 1990).

1.3.6 HIV-1 Mutation

The HIV-1 virus has one of the fastest mutating genomes, 10^{-4} to 10^{-5} mutations per site per replication cycle (Preston and Dougherty, 1996), due to the error-prone mechanism of reverse transcription. Over half of the DNA transcripts synthesized by reverse transcriptase contain at least one mistake in their nucleotide sequence (Hubner et al., 1992). These mutations can potentially attenuate the growth rate of virus by having an adverse effect on the function of viral proteins. The virus, however, partially compensates for this by having a high recombination rate whereby the reverse transcriptase switches back and forth between the two template RNA strands (Jetzt et al., 2000). This can result in a proviral DNA strand being synthesized which is more selectively fit than either of the parent RNA strands. High rates of mutation and recombination in HIV-1 result in a dynamic viral population with great genetic diversity both intra- and inter-individual. This high mutation rate is the key to the success of the virus in evading the antiviral drugs and immune system surveillance (Hu and Temin, 1990).

There are three steps in a retrovirus life cycle that can give rise to the observed high mutation rates; (1) synthesis of the minus-strand DNA based on the RNA template by RT, (2) synthesis of the plus-strand DNA based on the minus-strand DNA template by RT (section 1.3.5.2), and (3) synthesis of the viral mRNA by host-encoded RNA polymerase II (Smith et al., 2005). The reverse transcription appears to have, by far, the highest contribution to the genetic variability of HIV-1 strains due to the lack of a 3' → 5'

exonucleolytic proofreading activity in RT (Battula and Loeb, 1976; Roberts et al., 1988). The host-encoded RNA polymerase II plays a much smaller role in the rate of mutation in the viral genome (O'Neil et al., 2002).

Furthermore, the selective pressure imposed by the host immune system and antiviral drugs, result in the accumulation of escape variants through the course of HIV-1 infection. This has been demonstrated by the higher genetic variation in viral sequences in control patients compared to those with suppressed immune system (Booth et al., 1998). Host-to-host transmission, diverse cellular environments in different tissues, and anatomical restrictions, such as the blood-brain barrier, are some of the other selective pressures that contribute to the mutation profile of the virus (Vignuzzi et al., 2006).

1.4 HIV-1 resources used in this project

1.4.1 Public HIV Databases

One of the advantages in the HIV-1 research is the vast amount of publicly available resources with efficient search interfaces and useful sequence and structure annotations. The "HIV-1, Human Protein Interaction Database" (Table 1-3), hosted at NCBI, contains a search interface for the interactions between the HIV-1 and human proteins that are documented in the literature. The "HIV Drug Resistance Database" at Stanford University and the "HIV Resistance Response Database Initiative", contain information and resources on the drug resistance mutations, with the latter database dedicated mainly to clinical applications. The "HIV Structural Database" is divided into 2D and 3D structure databases. The former database mainly holds information on 2D structures of HIV inhibitors and a search interface to find these inhibitors or their specific fragments. The "HIV Infectious Diseases Integrated Database" is hosted in Japan and holds a record of virus isolation from HIV-infected

individuals in Japan and Thailand, their sequences, and other information such as chronological order, etc. The "HIV Protease Database" is dedicated to HIV-1, HIV-2, and SIV protease structures and related analytical tools such as the calculation of structural atomic distances, torsional angles, etc. The "HIV Positive Selection Mutation Database" is hosted by the bioinformatics group at the University of California, Los Angeles (UCLA), and enables the calculation of statistically significant Ka/Ks values for each individual amino acid mutation in protease and RT. The "bioafrica HIV database" is a comprehensive database on different aspects of HIV virus, from molecular epidemiology and characterization of subtypes to various proteomics tools. Finally, the Los Alamos National Library (LANL) Database, hosted at Los Alamos, USA, is the most widely used HIV database which is explained in the next section.

Table 1-3 HIV-1 molecular and structural resources.

Database	URL
HIV-1, Human Protein Interaction Database	http://www.ncbi.nlm.nih.gov/RefSeq/HIVInteractions/
Stanford University HIV Drug Resistance Database	http://hivdb.stanford.edu/
HIV Resistance Response Database Initiative	http://www.hivrdi.org/
HIV Structural Database	http://xpdb.nist.gov/hivsdb/hivsdb.html
HIV Infectious Diseases Integrated Database	https://aids.nih.go.jp/
HIV Protease Database	http://mcl1.ncifcrf.gov/hivdb/index.html
HIV Positive Selection Mutation Database	http://www.bioinformatics.ucla.edu/HIV/
bioafrica HIV database	http://www.bioafrica.net/
Los Alamos National Library Database	http://www.hiv.lanl.gov

1.4.2 LANL Database

The LANL database is comprised of four internal databases; sequence database, resistance database, immunology database, and vaccine trial database. The sequence database is periodically updated from the GenBank files at NCBI. It contains a search interface that allows the user to specify different attributes of sequences such as subtype, country, sequencing year, genomic region and more specific features such as days from

seroconversion, CD4 count, viral load, infection year, etc. As of June, 2007, there are 182,192 HIV-1 sequences and 2,286 HIV-2 sequences on this database. Table 1-4 shows the distribution of available HIV-1 complete genomes based on their geographical region. There are a total of 1598 complete HIV-1 genome sequences worldwide, with 732 (45%) from Africa. Asia and North America have the second and third highest number of complete genomes in this database, respectively.

Table 1-4 Available Number of HIV-1 Sequences on LANL based on Geographic Region. Table compiled on June 25th, 2007.

Geographical Region	Complete Genome
Any	1598
North America	215
South America	102
Central America	0
Caribbean	21
Oceania	41
Middle East	16
Asia	314
Africa	732
Europe	118
Former USSR	38

The distribution of HIV-1 sequences based on subtype and circulating recombinant forms (section 1.3.1), are shown in table 1-5. Even though subtype C accounts for almost 50% of all infections worldwide and subtype B for only 10% (Hemelaar et al., 2006), there are about 10 times as much subtype B sequences as subtype C on the LANL databases. This is due to the more funding and resources available for sequencing subtype B which is prevalent in North America. Each year the database curators gather a set of "subtype of reference sequences" that are considered to be representative complete genomes of all of the subtypes and circulating recombinants forms in the HIV-1 M, N, and O groups. The CRFs are numbered sequentially as they become known, starting from CRF01, CRF02, and so on.

The CRF names indicate the parental subtypes. For example, CRF01AE is a recombinant form of subtypes A and E. The "cpx" annotation in table 1-5, stands for "complex", meaning that this CRF is a mixture of more than two different subtypes. For example, CRF18_cpx is a recombination of A, F, G, H, K, and U subtypes.

Table 1-5 Available number of HIV-1 subtypes on LANL. Table compiled on June 25th, 2007

Sub-type	Qnt.	CRF	Qnt.	CRF	Qnt.
A	10006	CRF01_AE	6671	CRF13_cpx	98
B	104758	CRF02_AG	4012	CRF14_BG	91
C	10967	CRF03_AB	253	CRF15_01B	18
D	2979	CRF04_cpx	23	CRF16_A2D	5
F	1162	CRF05_DF	29	CRF18_cpx	61
G	2021	CRF06_cpx	867	CRF19_cpx	89
H	194	CRF07_BC	195		
J	104	CRF08_BC	187		
K	44	CRF09_cpx	45		
N	22	CRF10_CD	199		
O	980	CRF11_cpx	594		
U	439	CRF12_BF	346		

The distribution of individual components of HIV-1 genome in LANL database, as of June 25th, 2007, is shown in table 1-6. Due to the importance of the viral PR in antiviral drug design, this protein has by far the highest number of sequences (42,623) among different HIV-1 products. The search interface allows the user to download specific regions of the genome in nucleotide format or as amino acid sequences, translated in three different frames.

Table 1-6 Available Number of HIV-1 sequences on LANL. Table compiled on June 25th, 2007

Region	Qnt.	Region	Qnt.	Region	Qnt.
Complete Genome	1598	PR	42623	Env CDS	3760
Gag-Pol	1270	P31 (Integrase)	1944	gp41	3921
Gag	1748	P6	4949	gp120	4423
P17 (Matrix)	4519	Vif CDS	2612	TAR	1580
P24 (Capsid)	4260	Vpr CDS	2426	RRE	4659
P7 (Nucleocapsid)	4476	Vpu CDS	2887	5' LTR	428
Pol CDS	1674	Nef CDS	5098	5' LTR U3	760
P51 (RT)	2550	Tat CDS (plus intron)	1688	5' LTR R	1417
P15 (RNaseH)	2268	Rev CDS (plus intron)	1705	5' LTR U5	1250

For some of the patients in LANL database, multiple HIV-1 sequences have been extracted in different years (for a representative dataset see section 4.3). This occurs when the patient undergoes a longitudinal study, in which the patient is monitored over several years to observe different aspects of the infection, in particular dynamics of the virus evolution such as development of drug resistant mutations, cell tropism, etc.

Studies of translation initiation in HIV-1 could benefit from inclusion of human immunodeficiency virus type 2 (HIV-2) sequences. These HIV-2 sequences and in particular complete genomes, however, are quite rare in public databases, including the LANL. For example, as opposed to HIV-1, the search interface of the LANL database does not provide genomic region search abilities for HIV-2. This lack of sequence availability for HIV-2 can be attributed to both the confinement of HIV-2 to Western Africa and its relatively low prevalence and virulence.

1.4.3 HIV Structures in Protein Data Bank (PDB)

The RCSB protein data bank (PDB) is the database of protein and nucleic acid sequence structures (Berman et al., 2000). This database, which was originally established in 1971 with 7 structures, currently contains a total of 44,200 structures obtained by X-ray crystallography, NMR, and electron microscopy. It offers a variety of tools for browsing, searching, reporting, and visualizing the structures.

As of June 25th, 2007, there are about 830 HIV-1 related structures deposited in PDB, which range from protease and reverse transcriptase to small parts of regulatory proteins and inhibitors. Because of the aforementioned importance of HIV-1 protease in antiviral drug design, there are approximately 250 HIV-1 protease structures in this database as of June

2007. Some of these structures, however, are redundant corresponding to identical sequences. Also the majority of these structures are synthetic molecules, identified by a "Synthetic Construct" flag in the corresponding PDB file. These are usually obtained by inducing mutations of interest in common HIV-1 isolates (e.g. isolates BI21 and D19) and expressing them in *E. coli* expression system. Commonly induced mutations include Q7K, L33I, and L63I for minimizing autoproteolysis, and C67A and C95A for preventing cysteine-thiol oxidation.

1.5 Overview of subsequent chapters

This chapter was a general overview of AIDS, the HIV-1 virus and various stages of its replication cycle, and the available HIV-1 resources and sequence and structure databases. The next five chapters focus on the mechanism of translation initiation. HIV-1 is exclusively dependent on human translation machinery for the production of its proteins. Chapter 2 is devoted to the translation initiation in human, in particular, the two hypotheses regarding the role of +4G of Kozak consensus in translation initiation of human mRNAs are discussed; translation initiation hypothesis and penultimate amino acid constraint hypothesis. This chapter also establishes the grounds for the introduction of a bioinformatics approach in identifying penultimate amino acids that are efficient for N-terminal methionine excision, which is presented in chapter 6.

The CDRSM and IRES-dependent translation initiation hypotheses of translation initiation mechanism in HIV-1 were discussed in section 1.3.5.5. The Kozak consensus is important in the detection of the true AUG initiation codon by the scanning ribosome. Therefore, one of the predictions of the CDRSM hypothesis is a high conservation of Kozak

consensus sequences among different HIV-1 subtypes, which is assessed in chapter 3. In chapter 4, the upstream regions of HIV-1 CDSs are investigated for strong site conservations which could potentially be involved in cap-independent translation initiation mechanisms. This chapter also covers the separation of HIV-1 and human CDSs based on the 5'-UTR 50-mers upstream of their translation initiation codons. In chapter 5 we test the predictions of both CDRSM and IRES-dependent translation initiation hypotheses with regard to a selective pressure against ATG usage in optimal contexts in the HIV-1 5'-UTR. Finally, chapter 7 includes a summary, conclusion remarks, and suggested future approaches, based on the result of the analyses presented in previous chapters.

Chapter 2 Translation initiation in human: the role of +4G of the Kozak consensus

2.1 Abstract

The optimal sequence for translation initiation in mammalian species, known as the Kozak consensus, is RCCaugG, where aug is the start codon. The -3R and +4G, relative to the "a" in aug start codon (designated as +1), are the most important sites in the Kozak consensus. Experimental evidence shows that the -3R plays an important role in translation initiation by augmenting the start codon recognition signal for the scanning ribosome. The role of +4G in translation initiation, however, has been controversial. One hypothesis states that the +4G is important in translation initiation, especially in the absence of -3R. The other hypothesis explains the prevalence of +4G by invoking the observation that efficient N-terminal methionine excision (NME) of nascent protein sequences requires small amino acids such as Ala and Gly, which are coded by G-starting codons, at the penultimate (second) position. Here we use a bioinformatics approach to evaluate these alternative hypotheses in human coding sequences (CDS). Considering the claim that the +4G is involved in translation initiation especially in the absence of -3R, we expect to observe a higher frequency of +4G in CDSs with -3Y compared to those with -3R. Our result shows that this is not the case for either the entire set of genes or for the highly expressed genes. Our results contradict the predictions of the hypothesis which claims that the +4G is important in the translation initiation of human coding sequences.

2.2 Introduction

The prevailing hypothesis for translation initiation in multicellular eukaryotes is the cap-dependent ribosomal scanning mechanism (CDRSM) (Kozak, 1980; Kozak, 1989). According to this model, an initiation complex comprised of the small 40S ribosomal subunit, $tRNA_i^{Met}$ and multiple translation initiation factors, binds at the 5' end of the mRNA and scans linearly in the 5' to 3' direction until it reaches the first AUG in an optimal context. The 60S ribosomal subunit then binds to the complex and the translation of a polypeptide chain starts. The optimal context for translation initiation in mammalian species, known as the Kozak consensus, is RCCaugG (Kozak, 1986b). Using mutagenesis experiments it has been shown that changing -3R to -3Y significantly reduces the translation initiation efficiency, as does mutating +4G to any other nucleotide, especially in the absence of -3R (Kozak, 1986b; Kozak, 1997).

Both the experimental data and the high level of conservation of -3R in a wide variety of species including fungi, protists, plants, invertebrates, and mammals (Cavener, 1987; Cigan and Donahue, 1987; Hamilton et al., 1987; Joshi et al., 1997; Yamauchi, 1991), point to its involvement in the translation initiation mechanism. This is, however, not the case for the +4G. As far as the experimental data is concerned, the involvement of this site in translation initiation has not been clearly demonstrated. For example, in the presence of a +5U, no augmentation of translation initiation was observed in a mRNA containing +4G relative to the control mRNA with no +4G (Kozak, 1997). Also the +4G is not as conserved as the -3R among different taxonomic groups; in protozoa the consensus nucleotide at the +4 site is A (Yamauchi, 1991), and in yeast (*Saccaromyces cerevisiae*) the consensus nucleotide at this site is U (Cigan and Donahue, 1987; Hamilton et al., 1987). If the +4 site was

important for translation initiation, it would mean that the translation machinery uses +4G in multicellular eukaryotes, but +4A in protozoa and +4U in yeast for translation initiation, which is a questionable argument.

An alternative hypothesis, which explains the prevalence of +4G by invoking the amino acid constraint at the penultimate position (second position after initial methionine), was first proposed by Flinta et al. (Flinta et al., 1986). According to this hypothesis, hereafter referred to as the amino acid constraint hypothesis (Xia, 2007a), the prevalence of +4G is explained by the observation that the most frequent amino acids at the penultimate site, are small amino acids such as Ala, Gly, and Val. These amino acids are coded by G-starting codons and therefore their high frequency at the penultimate site results in the prevalence of G at the +4 site.

The amino acid constraint hypothesis is further supported by studies on N-terminal methionine excision (NME) and myristoylation. NME which occurs in more than half of all proteins in both prokaryotes and eukaryotes (Giglione et al., 2004; Giglione et al., 2003; Meinnel et al., 1993; Serero et al., 2003), refers to the removal of the initial methionine soon after the emergence of the amino terminus of the nascent polypeptide chain from ribosome. NME is not only an important amino-terminal modification in itself, but also required for further amino-terminal modifications such as myristoylation, which involves the attachment of a myristoyl ($C_{14}H_{28}O_2$) fatty acid side chain to a Gly at the penultimate position (Farazi et al., 2001). Myristoylation may involve many proteins and are implicated in protein subcellular relocalization (Farazi et al., 2001), apoptosis (Sakurai and Utsumi, 2006; Vilas et al., 2006), signal transduction (de Vries et al., 2006; Rowe et al., 2006), and the virulence and colonization of pathogens (Bentham et al., 2006; Breuer et al., 2006; Harkins et al., 2005; Provitera et al., 2006; Robert-Seilaniantz et al., 2006). Hence, myristoylation, which

only occurs on a Gly (coded by GGN) residue (Farazi et al., 2001), also contributes to the prevalence of +4G.

There are five amino acids coded by G-starting codons; Ala (GCN), Gly (GGN), Val (GTN), Asp (GAY), and Glu (GAR). According to the translation initiation hypothesis, which postulates that the +4G increases the translation initiation efficiency in multicellular eukaryotes, we should expect to see an even overrepresentation of these five amino acids at the penultimate position. The amino acid constraint hypothesis, however, predicts that only small amino acids such as Ala and Gly, which facilitate efficient NME and myristoylation, should be overrepresented at the penultimate site. Using this rationale, a recent extensive bioinformatics approach strongly supports the amino acid constraint hypothesis (Xia, 2007a).

Table 2-1 shows the molecular weight and Gyration radius (Levitt, 1992) of the 20 amino acids. The average of the molecular weights for amino acids coded by G-starting codons and amino acids coded by NonG-starting codons is 112.31 and 143.78, respectively. The Student's *t* test revealed that the average of molecular weights of the amino acids coded by the G-starting codons is significantly less than that of amino acids coded by NonG-starting codons ($p<0.05$). This is in support of the hypothesis of small amino acid coded by G-starting codons being present at the penultimate position and contributing to the prevalence of +4G. Even though the average of the Gyration radius of the amino acids coded by G-starting codons is less than that of the amino acids coded by NonG-starting codons (1.05 vs. 11.708), the Student's *t* test showed that this difference is not significant ($p>0.05$).

Table 2-1 The molecular weight (MW) and Gyration radius of the amino acids categorized into being coded by G-starting codons and NonG-starting codons

Coded by G-starting codons			Coded by NonG-starting codons								
AA	MW	Gyr. Rad.	AA	MW	Gyr. Rad.	AA	MW	Gyr. Rad.	AA	MW	Gyr. Rad.
Ala	89.09	0.77	Cys	121.16	1.22	Leu	131.18	1.54	Arg	174.2	2.38
Gly	75.07	0	Phe	165.19	1.9	Met	149.21	1.8	Ser	105.09	1.08
Val	117.15	1.29	His	155.16	1.78	Asn	132.12	1.45	The	119.12	1.24
Asp	133.1	1.43	Ile	131.18	1.56	Pro	115.13	1.25	Trp	204.23	2.21
Glu	147.13	1.77	Lys	146.19	2.08	Gln	146.15	1.75	Tyr	181.19	2.13

One might argue in the defense of the translation initiation hypothesis, that the experimental results point to the involvement of the +4G in translation initiation in those protein coding genes where the -3R is missing. A clear prediction of this hypothesis, therefore, would be that the +4G should be overrepresented in protein-coding genes with -3Y compared to those with -3R, especially in highly expressed genes where an efficient translation initiation is required.

In this study, we have tested this prediction on human protein-coding sequences. Our results are yet another support of amino acid constraint hypothesis and contradict the translation initiation hypothesis.

2.3 Materials and Methods

The entire set of human CDSs were downloaded from Ensembl with the BioMart search engine (Birney et al., 2004); http://www.ensembl.org/biomart/martview. "Ensembl 45" and "Homo sapiens genes (NCBI36)" were selected as the database and the dataset, respectively. The known protein-coding genes were selected in the "Filters" panel (i.e. the predicted and novel genes by the database were excluded), and coding sequences (CDS) were specified in the "Attributes" panel. Since we are interested in the frequency of the -3 and +4 site of the Kozak consensus, we also specified the three upstream flanking nucleotides for each CDS to

be included in the download. Ensembl gene ID and Ensembl transcript ID were both included in the fasta headers so that each sequence could be uniquely identified. The search interface allows the download of all transcripts of a gene. As of July 2007, there were a total of 31,484 human genes in the Ensembl database, 21,658 of which are known protein-coding genes. These 21,658 genes are associated with 39,684 transcripts (i.e. some genes generate more than one transcript), which were downloaded from the database. Excluding the transcripts with non-ATG start codons and those belonging to mitochondrial genes, there were a total of 37,315 known protein-coding transcripts. The transcripts with non-ATG start codons were excluded since this would add an extra filter in the exclusion of artificial genes from the dataset.

The downloaded dataset also included some redundant sequences, having identical CDS and the three upstream flanking bases. These identical sequences emanate from two different sources. One is the alternative splicing of the upstream non-coding exons from the primary transcript of a gene, which does not affect the sequence similarity of the coding regions and the three upstream bases in different transcripts of the same gene (e.g. ENST00000346432 and ENST00000355162, both of which belong to the TBL1Y gene: ENSG00000092377). The second source is the annotation of identical transcripts from two different groups (e.g. ENST0000215479 annotated by Ensembl and ENST00000383037 annotated by Havana group, both corresponding to the AMELY gene: ENSG00000099721). Identical transcripts are usually merged together to avoid redundancy, but in rare cases both transcripts are included in the downloaded dataset (personal communication with Ensembl helpdesk). Using a script written by Sam Khalouei in our lab, we excluded the identical sequences belonging to the same gene. The final dataset with unique human sequences used in this study contained 33,573 sequences.

The DAMBE program (Xia, 2001; Xia and Xie, 2001), available at http://dambe.bio.uottawa.ca, was used to calculate the codon adaptation index (CAI) (Sharp and Li, 1987) value for each CDS. The CAI algorithm included in DAMBE avoids implementation problems associated with several other publicly available programs, such as exclusion of codon families with a single codon in computing CAI (Xia, 2007c). We specified the Ehuman.cut, which contains a set of highly expressed human genes, distributed with EMBOSS (Rice et al., 2000), as the human reference set and the minimum length of sequences was set to 300 in order to avoid the bias associated with short sequences in CAI computation (Xia, 2007c). The human CDSs were split into low-CAI (CAI<= 0.65) and high-CAI (CAI>= 0.8) groups and the -3 and +4 site frequencies were calculated for each group, as well as for the entire set of sequences.

2.4 Results and Discussion

Table 2.2 shows the nucleotide frequency of the -3 and +4 sites of the Kozak consensus in the entire set of 33,573 human CDSs used in this analysis. The -3R is highly conserved (79.9%) in human CDSs, with -3A and -3G comprising 43.2% and 36.7% of the cases, respectively. Even though G is the most frequent nucleotide at the +4 site, it is not as highly conserved (48.3%) as the -3R.

The translation initiation hypothesis claims that the +4G is important in translation initiation, especially in the absence of -3R (Kozak, 1986b; Kozak, 1997). Therefore, according to this hypothesis, +4G should be overrepresented in human CDSs with -3Y to compensate for an otherwise weak translation initiation signal. Table 2.2 clearly shows that this is not the case. Indeed in the sequences where the -3R is absent, a significantly (X^2=40.33, d.f.=1, $p<0.0005$) lower percentage (44.9%) of the +4G is observed, compared to

those with the -3R (49.2%). This is in contradiction with the prediction of the translation initiation hypothesis. Albeit, the +4G is still the most frequent nucleotide in the presence of -3Y, but compared to the sequences with -3R, a higher percentage of the sequences with -3Y are expected to have a +4G according to the prediction of translation initiation hypothesis.

Table 2-2 The nucleotide frequencies of the -3 and +4 sites of the human CDSs. The percentages in the brackets, represent the proportion of +4 sites in each -3 site category so that the sum of the percentages in each row is equal to 100.

-3	+4A	+4G	+4C	+4U	Sum
A	3517 (24.2)	6593 (45.4)	2184 (15.0)	2233 (15.4)	14527
G	2341 (19.0)	6606 (53.7)	1811 (14.7)	1553 (12.6)	12311
C	810 (20.1)	1930 (47.9)	797 (19.8)	491 (12.2)	4028
U	659 (24.3)	1091 (40.3)	518 (19.1)	439 (16.2)	2707
R	5858 (21.8)	13199 (49.2)	3995 (14.9)	3786 (14.1)	26838
Y	1469 (21.8)	3021 (44.9)	1315 (19.5)	930 (13.8)	6735
Sum	7327 (21.8)	16220 (48.3)	5310 (15.8)	4716 (14.0)	33573

It may be argued that the +4G should only be overrepresented in highly expressed genes (Kozak, 1999), to promote efficient translation initiation in these genes. In other words, the results observed in table 2.2, may simply be due to the inclusion of many lowly-expressed genes, upon which little, if any, selective pressure would act to compensate for -3Y in translation initiation. It is therefore imperative to divide the dataset into highly-expressed and lowly-expressed genes and perform the analysis separately for each group.

In order to categorize genes into lowly and highly-expressed groups, in the absence of high-throughput human gene expression data, we used the CAI values calculated using the DAMBE program (Xia, 2001; Xia and Xie, 2001). CAI is a measure of the degree of bias in synonymous codon usage (Sharp and Li, 1987). It is a powerful tool which has been widely used in identifying highly expressed genes (Goetz and Fuglsang, 2005; Popescu et al., 2006; Wu et al., 2005), predicting protein abundance (Futcher et al., 1999; Jansen et al., 2003; Lithwick and Margalit, 2005), detecting horizontal gene transfer (Bodilis and Barray, 2006),

and improving the codon usage effects of DNA vaccines (Ruiz et al., 2006). The conventional method of calculating this index is to use a set of highly expressed genes as a reference set to construct a table of codon weights, *w*. These weights are obtained in each codon family by dividing each codon frequency by the maximum codon frequency (i.e. the codon with the highest frequency in each codon family is assigned a weight of 1). The codon weights are then used in the following equation to calculate the CAI value for a gene:

$$CAI = e^{\left(\frac{\sum_{i=1}^{n} [CodFreq_i \ln(w_i)]}{\sum_{i=1}^{n} CodFreq_i} \right)}$$

where n stands for the number of sense codons and *CodFreq* refers to the frequency of the codons in the gene of interest. The CAI values obtained from this equation range from 0 to 1, with a CAI value greater than 0.7 generally considered to represent a highly expressed gene (Sharp and Li, 1987).

Even though the use of CAI has been controversial in assessing the gene expression levels in humans (Duret, 2002; Semon et al., 2006), recent studies point to their reliability for this purpose (Xia, 2007a). Even in the presence of gene expression data for a species, using CAI as the index of gene expression can be advantageous. In higher eukaryotes such as human, many genes are highly expressed only at specific times and in specific tissues. Therefore, low expression of a gene in a particular study, which could have involved few time points and few tissues, might result in the misinterpretation that it is a lowely-expressed gene. Highly expressed genes are expected to be efficient at translation elongation. The CAI index which evaluates the codon usage bias of the gene in a time- and tissue-independent manner is a reliable index of translation elongation efficiency of a gene.

The CAI values for the set of lowly and highly-expressed genes in our dataset are less than 0.65 and bigger than 0.8, respectively. Table 2-3 shows the nucleotide frequencies for the -3 and +4 site for each group. As previously mentioned, +4G should be overrepresented in human CDSs with -3Y to compensate for an otherwise weak translation initiation signal. Within both the low-CAI (X^2=17.9, d.f.=1, $p<0.0005$) and high-CAI (X^2=6.13, d.f.=1, $p<0.02$) groups, however, in the presence of -3Y a significantly lower percentage of the +4G is observed, compared to those with the -3R . This is in contradiction with the prediction of the translation initiation hypothesis.

Another prediction of the translation initiation hypothesis is that in the absence of -3R, +4G should be overrepresented in the highly-expressed genes compared to the lowely-expressed genes. Even though, the percentage of -3Y+4G is 40.0% in the low-CAI group and 42.6% in the high-CAI group, this difference is not significant (X^2=0.608, d.f.=1, $p>0.25$), which contradicts the prediction of translation initiation hypothesis.

Table 2-3 The nucleotide frequencies of the -3 and +4 sites of the human CDSs in the low-CAI and high-CAI groups. The percentages in the brackets, represent the proportion of +4 sites in each -3 site category so that the sum of the percentages in each row is equal to 100.

	-3	+4A	+4G	+4C	+4U	Sum
Low CAI	A	593 (25.6)	1064 (45.9)	285 (12.3)	376 (16.2)	2318
	G	351 (21.7)	812 (50.3)	196 (12.1)	255 (15.8)	1614
	C	131 (27.0)	205 (42.2)	96 (19.8)	54 (11.1)	486
	U	119 (27.5)	162 (37.5)	72 (16.7)	79 (18.3)	432
	R	944 (24.0)	1876 (47.7)	481 (12.2)	631 (16.0)	3932
	Y	250 (27.2)	367 (40.0)	168 (18.3)	133 (14.5)	918
	Sum	1194 (24.6)	2243 (46.2)	649 (13.4)	764 (15.8)	4850
High CAI	A	199 (23.5)	372 (43.9)	145 (17.1)	132 (15.6)	848
	G	113 (14.6)	446 (57.7)	123 (15.9)	91 (11.8)	773
	C	49 (24.5)	90 (45.0)	41 (20.5)	20 (10.0)	200
	U	23 (25.8)	33 (37.1)	22 (24.7)	11 (12.4)	89
	R	312 (19.2)	818 (50.5)	268 (16.5)	223 (13.8)	1621
	Y	72 (24.9)	123 (42.6)	63 (21.8)	31 (10.7)	289
	Sum	384 (20.1)	941 (49.3)	331 (17.3)	254 (13.3)	1910

A recent extensive bioinformatics study (Xia, 2007a) provided strong support for the alternative amino acid constraint hypothesis, which explains the prevalence of +4G as a result of the overuse of small amino acids at the penultimate position to promote efficient NME and myristoylation (Khalouei et al., 2007; Moerschell et al., 1990). Xia (Xia, 2007a) performed a critical test on the alternative hypotheses concerning the +4G site. The prediction of translation initiation hypothesis is that all five amino acids coded by G-starting codons (Ala coded by GCN, Asp by GAY, Glu by GAR, Gly by GGN, and Val by GUN) should be evenly distributed at the penultimate position, whereas the prediction of the penultimate constraint hypothesis is that only Ala and Gly, which are efficient for NME, should be overrepresented at this position. Using 34,169 human protein-coding genes, Xia showed that only Ala and Gly to a smaller extent are overrepresented at the penultimate position. The results presented in Xia's study (Xia, 2007a) were inconsistent with +4G being needed for efficient translation initiation, but consistent with the proposal of amino acid constraint hypothesis. A broader analysis from our lab, similar to the study presented here, with additional model organisms, and a bioinformatic approach to identify penultimate amino acids efficient for NME (chapter 6), both support the amino acid constraint hypothesis as well.

In this study the lower and upper limits of CAI for categorizing genes into lowely- and highly-expressed sets were arbitrarily chosen as 0.65 and 0.8, respectively. An alternative option would be to select one third of the dataset with the lowest CAI values and one third of the dataset with the highest CAI values as the lowely- and highly-expressed sets, respectively. Also, given that +4G can compensate -3Y, it is likely that other nucleotides at other positions of the Kozak consensus may also play compensatory roles in the absence of -3R. As table 2-2 shows, many genes have -3Y and do not have +4G. A future approach

would be to look at this set of genes with -3Y and +4H (H= nonG nucleotides), and investigate the similarities in other positions (e.g. sites -1 and -2), which could indicate alternative compensating mechanisms for translation initiation signal in the absence of -3R.

Chapter 3 Kozak Consensus in HIV-1 Genes

3.1 Abstract

Human Immunodeficiency Virus Type 1 (HIV-1) is an intracellular obligatory parasite that depends on the human transcription and translation machinery to complete its infection cycle. Different hypotheses have been postulated regarding the mechanism of translation initiation in HIV-1 mRNAs. The data so far support both the cap-dependent ribosomal scanning mechanism (CDRSM) and internal ribosome entry site (IRES)-dependent hypotheses of translation initiation in HIV-1, as described in section 1.3.5.5. Here, we have analyzed the Kozak consensus sequence in the five most abundant HIV-1 subtypes worldwide (A, B, C, CRF01_AE, and CRF02_AG). The -3 and +4 sites of the Kozak sequences appear to be highly conserved among these subtypes. The high level of conservation of the Kozak -3 site in a large number of HIV-1 strains from the most abundant subtypes, highlights the importance of this region in translation initiation and supports the CDRSM hypothesis. The strong conservation of the +4 site among these sequences supports the amino acid constraint hypothesis.

3.2 Introduction

Translation of eukaryotic mRNAs and the importance of -3 and +4 sites of the Kozak consensus were explained in the opening introduction of chapter 2. HIV-1 viruses use the human translation machinery for the expression of their genes. The CDRSM and IRES-dependent hypotheses of translation initiation in HIV-1 were described in section 1.3.5.5. The strength of the Kozak consensus sequence has been shown to play an important role in the regulation of translation in HIV-1 mRNAs (Luukkonen et al., 1995; Schwartz et al.,

1992). There have, however, been some contradictory results regarding the translation initiation in HIV-1 through CDRSM. For example, whereas Schwartz and colleagues (Schwartz et al., 1992) stated that the leaky scanning at the *vpu* AUG is necessary for the production of downstream Env proteins through ribosomal scanning, a recent report by Anderson *et al.* (Anderson et al., 2007), provided evidence that *env* is translated through a discontinuous ribosomal scanning such as internal ribosome entry (IRE).

One of the predictions of the CDRSM hypothesis is the conservation of Kozak sequences in the genes of phylogenetically different HIV-1 strains. Regardless of their relative strength, if these regions play a role in CDRSM, they should be conserved within and among different strains. To our knowledge, there has not been any extensive analysis of the Kozak consensus regions in HIV-1 related studies. Here we have analyzed the conservation of the -3 and +4 sites of the Kozak sequences among the five most abundant HIV-1 subtypes and CRFs (Hemelaar et al., 2006), (i.e. subtypes A, B, C, CRF01_AE, and CRF02_AG, described in section 1.3.1). The high level of conservation of the -3 site among most virulent and prevalent HIV-1 strains in different regions of the world supports the CDRSM hypothesis as a universal mechanism of translation initiation in HIV-1 viruses.

3.3 Materials and Methods

In order to analyze the level of conservation of -3 and +4 sites in Kozak sequences among the five most abundant HIV-1 subtypes, we used the Los Alamos National Library (LANL) database (described in section 1.4.2), to extract the region around the translation initiation codons in HIV-1 CDSs. We used the HIV Sequence Compendium 2005 (Leitner et al., 2005) to find the location of these regions in all the eight HIV-1 genes that have independent translation initiation codons; *gag*, *vif*, *vpr*, *tat*, *rev*, *vpu*, *env*, and *nef*. All the

HIV-1 sequences in LANL database are aligned against the HXB2 reference sequence and by selecting a specific region in the search interface at (http://www.hiv.lanl.gov/components/hiv-db/combined_search_s_tree/search.html), all the sequences which contain that region can be retrieved.

The location of the translation initiation codons in the HXB2 reference sequence are specified in the LANL database (the numbers in bracket specify the location of A in ATG start codon, except in *vpu* where the translation initiation codon is ACG in HXB2): *gag* (790), *vif* (5041), *vpr* (5559), *tat* (5831), *rev* (5970), *vpu* (6062), *env* (6225), *nef* (8797). In the search interface we specified a region of seven nucleotides around the translation initiation codon of each gene, with three nucleotides upstream and one nucleotides downstream of A in ATG translation initiation codon. For each gene all the sequences available for the aforementioned five HIV-1 subtypes were downloaded by specifying the subtype and HIV-1 strains in the "Subtype" and "Organism" menus, respectively. In the "Download options" of the "Results" page, the sequences are downloaded by default, as being aligned to the HIV-1 HXB2 reference sequence and also clipped to the region of interest. This makes it possible to download only the 7 nucleotides surrounding the start codon. We also checked the option "include the HXB2 Reference Sequence (K03455)" to help us locate the true translation initiation codon in the aligned downloaded sequences. The problematic sequences are omitted from download by default. According to the LANL database, these sequences belong to five different classes: (A) sequences with high content of non-ACTG characters, (B) sequences likely contaminated with a laboratory strain, (C) Hypermutated sequences, (D) Synthetic sequences, and (E) sequences containing an artificial deletion of >100 nucleotides. As an extra step, we also filtered out those sequences containing ambiguous nucleotides in their Kozak sequence and also the ones with non-ATG

start codons. Five sequences with accession IDs AY818644, U12055, AF443106, AF075719, and M37898 were also omitted since they contained a high number of insertions around the *gag* (subtype B), *vpr* (subtype B), *tat* (subtype C), *nef* (subtype B), and *vpu* (subtype B) Kozak sequences, respectively.

This resulted in a dataset of 40 different categories consisting of eight genes in five different subtypes. Using a script, we then compiled the nucleotide frequencies of the -3 and +4 sites of the Kozak sequences in each category.

3.4 Results and Discussion

Table 3-1 shows the Kozak sequences in the NCBI HIV-1 type sequence (NC_001802). *Gag*, *vif*, *tat*, and *nef* contain both -3R and the +4G of the Kozak consensus, with other genes missing either but not both of these. Except in *vpu* and *env*, a G is present at the +4 site of all genes. *vpr* and *rev* are the only genes where the -3R is missing. There is no "CC" conservation observed in either -1 and -2 (G**CC**) or -4 and -5 (R**CC**) positions of the Kozak consensus sequences. In fact, C is the least frequent nucleotide in these positions.

Table 3-1 Kozak sequences in NCBI HIV-1 type sequence (NC_001802) and their comparison with Kozak consensus. Positions -3 and +4 of the Kozak consensus are shown in bold. Individual nucleotides that resemble the corresponding Kozak consensus positions are shown in capital letters.

Gene Name	Kozak consensus GCC **R**CC ATG **G**
gag	aga **G**ag ATG **G**
vif	Ggg **A**tt ATG **G**
vpr	Gga **t**ag ATG **G**
tat	caa **G**aa ATG **G**
rev	tCt **c**Ct ATG **G**
vpu	cat **G**ta ATG **c**
env	Gtg **G**Ca ATG **a**
nef	tat **A**ag ATG **G**

There were a total of 15789 HIV-1 sequences included in this analysis with 1,211 *gag*, 1,512 *vif*, 2,229 *vpr*, 2,704 *tat*, 2,566 *rev*, 1,701 *vpu*, 2,000 *env*, and 1,866 *nef* sequences. Of these 15,789 sequences, 1,062, 8,939, 3,724, 1,614, and 449 sequences belonged to subtypes A, B, C, CRF01_AE, and CRF02_AG, respectively (Table 3-2). The detailed output of the -3 and +4 site frequencies of the Kozak sequences of these subtypes is shown in table 3-2. In the case of *gag*, 95.13% of all sequences contain the -3G+4G combination (i.e. both -3G and +4G are present in the same sequence), and almost all of them (99.84%) have the -3R+4G combination, conforming to the Kozak consensus. Also all five subtypes almost exclusively contain the -3A+4G combination (99.8%) in the *vif* gene. The *vpr* gene shows a lower level of conservation at the -3 position, with -3T+4G being the most abundant combination (91.11%), followed by -3C+4G (8.7%). The -3G+4G combination is also highly conserved in *tat* gene (99.26%) among all five subtypes.

In the case of *rev*, -3C+4G is the most abundant combination (95.28%), with subtype B containing about 2% of each -3A+4G and -3T+4G, and subtype C containing about 9% of -3A+4G combinations. *vpu* shows the highest level of diversity among the HIV-1 genes at the +4 position where nucleotides A, C, and T are frequent. As opposed to these nucleotides, however, G almost never appears at this position except in subtype C. The -3G site, however, is highly conserved in *vpu*, observed in 99% of the sequences.

All *env* ORFs appear on the same mRNA with the upstream *vpu* ORF. The weak translation initiation signal of this upstream *vpu* has been proposed to be necessary for efficient translation of the downstream *env* ORF by the scanning ribosome (Schwartz et al., 1992). The conserved -3G of *vpu* is part of a minimal upstream ORF consisting only of a start and stop codon which forms the AU**G**UAaug motif, where "aug" is the *vpu* start codon and the conserved G is shown in bold. It has been shown that this minimal upstream ORF

allows an efficient translation initiation at the downstream *env* AUG (Krummheuer et al., 2007). All subtypes show a high conservation for the -3G+4A combination in the *env* gene (99.15%). Finally, -3A+4G is highly conserved in *nef* among all five subtypes (98.6%).

Table 3-2 Nucleotide frequencies of the -3 and +4 sites in Kozak consensus of HIV-1 sequences downloaded from Los Alamos National Library Database (LANL).

			+4															
			gag				*vif*				*vpr*				*tat*			
			A	G	C	T	A	G	C	T	A	G	C	T	A	G	C	T
-3	A	A	0	2	0	0	0	124	0	0	0	0	0	0	0	1	0	0
		G	1	149	0	0	0	0	0	0	0	2	0	0	0	107	0	0
		C	0	0	0	0	0	0	0	0	0	32	0	0	0	0	0	0
		T	0	0	0	0	0	0	0	0	0	93	0	0	0	1	0	0
	B	A	0	6	0	0	0	653	0	0	0	0	0	0	0	4	0	0
		G	0	496	0	0	0	1	0	0	0	0	0	0	2	1861	9	0
		C	0	0	0	0	0	0	0	0	0	77	0	0	0	1	0	0
		T	0	0	0	0	0	0	0	0	2	1298	0	0	0	0	0	0
	C	A	0	1	0	0	0	461	0	0	0	0	0	0	0	0	0	0
		G	0	396	0	0	0	2	0	0	0	0	0	0	0	456	0	0
		C	0	0	0	0	0	0	0	0	0	23	0	0	0	1	0	0
		T	0	0	0	0	0	0	0	0	0	434	0	0	0	0	0	0
	AE	A	0	48	0	0	0	212	0	0	0	0	0	0	0	1	0	0
		G	1	89	0	0	0	0	0	0	0	0	0	0	0	202	0	0
		C	0	0	0	0	0	0	0	0	0	7	0	0	0	0	0	0
		T	0	0	0	0	0	0	0	0	0	196	0	0	0	0	0	0
	AG	A	0	0	0	0	0	59	0	0	0	0	0	0	0	0	0	0
		G	0	22	0	0	0	0	0	0	0	0	0	0	0	58	0	0
		C	0	0	0	0	0	0	0	0	0	55	0	0	0	0	0	0
		T	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0
			rev				*vpu*				*env*				*nef*			
			A	G	C	T	A	G	C	T	A	G	C	T	A	G	C	T
-3	A	A	0	0	0	0	0	0	0	0	0	0	0	0	1	190	0	0
		G	0	0	0	0	53	1	21	37	113	1	0	0	0	0	0	0
		C	0	132	0	0	0	0	0	0	1	0	0	0	0	0	0	0
		T	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	B	A	0	31	0	0	0	0	0	0	1	0	0	0	5	918	0	2
		G	0	0	0	0	218	1	610	30	1079	3	2	1	0	3	0	0
		C	1	1584	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		T	0	36	0	0	0	0	2	0	0	0	0	0	0	2	0	0
	C	A	1	47	0	0	0	0	0	0	1	0	0	0	0	466	0	0
		G	0	4	0	0	68	53	47	298	489	1	0	0	0	0	0	0
		C	0	466	0	0	0	1	0	1	1	0	0	0	0	0	0	0
		T	0	0	0	0	0	0	1	1	0	0	0	0	0	4	0	0
	AE	A	0	0	0	0	0	0	0	0	0	0	0	0	0	201	0	0
		G	0	0	0	0	78	0	1	118	241	3	0	0	0	0	0	0
		C	0	203	0	0	0	0	0	0	1	0	0	0	0	0	0	0
		T	0	0	0	0	2	0	0	1	0	0	0	0	0	9	0	0
	AG	A	0	0	0	0	0	0	1	0	0	0	0	0	0	65	0	0
		G	0	0	0	0	5	0	51	1	61	1	0	0	0	0	0	0
		C	0	60	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		T	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 3-3 shows the frequency of the +4 sites in each -3 site category for the eight genes over all five subtypes used in this study. There is a conserved purine (R) at position -3 in all genes, except in *vpr* and *rev*, where there are only 0.09% and 3.2% -3R, respectively. Also nucleotide G is highly conserved at position +4 in all genes, except *vpu* and *env*, where there are only 3.3% and 0.5% +4G, respectively. *env* instead has a conserved +4A (99.4%).

The four genes *gag*, *vif*, *tat*, and *nef* resemble the Kozak consensus in -3 and +4 sites with more than 99% of their sequences having an R (A/G) in the -3 position and a G in the +4 position. The other four genes, *vpr*, *rev*, *vpu*, and *env*, on the other hand, differ from the Kozak consensus, with only 3.2% of *rev* and *vpu* sequences and less than 1% of *vpr* and *env* sequences resembling the Kozak consensus.

Table 3-3 Pooled results of Kozak consensus analysis in HIV-1 genes of the five most abundant subtypes obtained from LANL database. The most abundant combination of nucleotides at positions -3 and +4, and the percentage of sequences with -3R and +4G (last column) are shown. N/A in *vpu* refers to the fact that the most abundant combination of *vpu*, as opposed to other genes, is not the same among the five subtypes.

HIV-1 Genes	Total	-3R	+4G	Most Abundant Combination	Kozak Consensus
gag	1211	100%	99.8%	-3G +4G 95.1%	-3R +4G 99.8%
vif	1512	100%	100%	-3A +4G 99.8%	-3R +4G 100%
vpr	2229	0.09%	99.9%	-3T +4G 91.1%	-3R +4G 0.09%
tat	2704	99.9%	99.6%	-3G +4G 99.3%	-3R +4G 99.5%
rev	2566	3.2%	99.9%	-3C +4G 95.3%	-3R +4G 3.2%
vpu	1701	99.5%	3.3%	N/A	-3R +4G 3.2%
env	2000	99.9%	0.5%	-3G +4A 99.2%	-3R +4G 0.5%
nef	1866	99.2%	99.6%	-3A +4G 98.6%	-3R +4G 98.8%

Overall, these results show a high level of conservation of -3 and +4 sites of the Kozak consensus in sequences of the most prevalent HIV-1 subtypes. While the conservation at the -3 site is in support of the CDRSM hypothesis of translation initiation in HIV-1, the conservation at the +4 site is most likely attributed to the amino acid constraint hypothesis mentioned in the previous chapter. For example, if the +4 site was involved in the translation initiation of HIV-1 genes, we would expect this site to be also conserved in *vpu*, regardless of the specific type of the nucleotide. Also the high conservation of +4G in *gag* and *nef* can be explained by the myristoylation of their proteins (Bentham et al., 2006; Provitera et al., 2006), which occurs only on a Gly residue (coded by GGN) (Farazi et al., 2001) and contributes to the prevalence of +4G. The *vif*, *vpr*, and *tat* proteins have a glutamic acid in the penultimate position, and the *rev* protein has an alanine in the penultimate position. These conserved amino acids could contribute to the conservation of +4G in these four proteins.

Chapter 4 Strong site conservations in HIV-1 5'-UTR point to functional role in translation initiation

4.1 Abstract

Both the cap-dependent ribosomal scanning mechanism (CDRSM) and internal ribosome entry (IRE)-dependent hypotheses of translation initiation have been shown to operate in HIV-1, with occasional contradictory evidences. Here we have analyzed 331 unique HIV-1 5'-UTR 50-mers immediately upstream of the AUG translation initiation codons, in a search for conserved sites that could potentially be involved in cap-independent translation initiation mechanisms. Our results indicate that these 50-mers and in particular the -17 and -25 sites, relative to AUG initiation codon, are highly conserved. Furthermore, we were able to separate the human and HIV-1 sequences based on these 50-mers which indicates the possibility of designing antiviral drugs to specifically target translation initiation in HIV-1 mRNAs.

4.2 Introduction

The two hypotheses regarding the translation initiation mechanism in HIV-1 were described in section 1.3.5.5. In at least two reports, there have been evidence for the initiation of translation in HIV-1 virus through cap-independent internal ribosome entry (IRE) (Brasey et al., 2003; Buck et al., 2001). We decided to search common features shared in the 5'-UTR among HIV-1 protein-coding genes which could be potentially involved in IRE-dependent translation initiation. A preliminary compilation of the HIV-1 type genome in NCBI (NC_001802) shows two strongly conserved sites (Fig. 4.1), one at the -17 site with all protein-coding genes having A and the other at the -25 site with all protein-coding genes

having R (A or G). Given the nucleotide frequencies of A in the HIV-1 genome (NC_001802) being 0.3564, the number of sites with an A in all eight genes from the 50 nucleotides (50-mers) 5'-UTR is 0.013 (=50×0.3564^8). This suggests that the site conservation at the -17 site (Fig. 4.1) may not be due to chance alone.

```
          -41       -31       -21       -11       -1
    ----|----|----|----|----|----|----|----|----|----|
gag ACTGGTGAGTACGCCAAAAATTTTGACTAGCGGAGGCTAGAAGGAGAGAGATGG
vif TAGTGACATAAAAGTAGTGCCAAGAAGAAAAGCAAAGATCATTAGGGATTATGG
vpr AAAAAGATAAAGCCACCTTTGCCTAGTGTTACGAAACTGACAGAGGATAGATGG
tat TGGGTGTCGACATAGCAGAATAGGCGTTACTCGACAGAGGAGAGCAAGAAATGG
rev TTGCTTTCATTGCCAAGTTTGTTTCATAACAAAAGCCTTAGGCATCTCCTATGG
vpu ACAGTCAGACTCATCAAGCTTCTCTATCAAAGCAGTAAGTAGTACATGTAATGC
env AATAGACAGGTTAATTGATAGACTAATAGAAAGAGCAGAAGACAGTGGCAATGA
nef CCACATACCTAGAAGAATAAGACAGGGCTTGGAAAGGATTTTGCTATAAGATGG
```

Figure 4.1 The 50 nucleotides upstream of the translation initiation sites in NCBI HIV-1 type genome (NC_001802). The conserved -17 and -25 sites are shown in bold.

While protein-coding sequences and their 5'-UTR are known to be conserved among homologous sequences of HIV-1 sequences, the site conservation in the 5'-UTR among different genes is unexpected. One possible explanation is as follows. Due to the overlapping nature of HIV-1 protein-coding genes, the -17 and -25 sites of HIV-1 genes (except for *gag*) are in the coding region of an upstream gene. If they happen to be at the third codon position, then the -17 site being A may be the result of codon usage bias in favor of A-ending codons. This explanation, however, is not valid in this case. The proportion of A at the third codon position in HIV-1 coding sequences (NC_001802) is 0.409. Even if all -17 sites are at the third codon position, the chance of observing a site with all A's is 0.039 (=50×0.409^8). The -17 site of the *gag* gene, however, is not at the third codon position, and the -17 site of the *vpr* gene, which lies in the coding region of *vif*, is at the first codon position of a lysine codon

(AAA). In any case, it is important to validate the site conservation because it may lead to a conserved drug target.

In this study we compiled data from 86 HIV-1 genomes belonging to 25 AIDS patients and extracted 50 nucleotides upstream of the initiation AUG codon. Also compiled are 24,850 human protein-coding genes. Our objectives are not only to find support for the site conservation at the -17 and the -25 sites among HIV-1 genes with independent translation initiation sites, but also to check whether the 5'-UTR of the HIV-1 genes can be separated from the human genes. A clear separation between the HIV-1 5'-UTR and human 5'-UTR would indicate that HIV-1 transcripts could be theoretically targeted specifically.

4.3 Materials and Methods

Position Weight Matrix (PWM)

The HIV-1 dataset in this study comprises of 86 complete HIV-1 genomes belonging to 25 different patients obtained from the LANL database. These patients have complete HIV-1 genomes in at least two different years. The patient IDs and the corresponding GenBank accession IDs are as follows (patient IDs are in bold): **1** (M17449, AF075719); **7082** (AY970946-AY970950); **57665** (AF224507, DQ295194); **138040** (AJ320484, AF484504); **10139674** (AF539404-AF539406); **10139902** (AY423381-AY423387); **10139968** (U69584-U69593); **10142058** (AY322184-AY322185); **10144867** (AY779550-AY779552, AY779564); **10144868** (AY779553-AY779556); **10144869** (AY779557-AY779563); **10146536** (DQ164105, DQ056408); **10150009** (AY835748-AY835752); **10150010** (AY835759-AY835761); **10150011** (AY835754, AY835765, AY835775); **10150012** (AY835755-AY835756, AY835781); **10150014** (AY835757-AY835758); **10150015** (AY835762-AY835764); **10150016** (AY835766-AY835768); **10150017** (AY835770-AY835773); **10150018** (AY835774, AY835776); **10150019**

(AY835777-AY835778); **10150020** (AY835779-AY835780); **10153344** (DQ487188-DQ487189); **10153345** (DQ487190-DQ487191).

The genomes in patient 10142058 belong to subtype A1, those in patient 10139674 belong to subtype A1 and A1c subtypes, and the ones in patient 10146536 belong to subtype C. All the other genomes are of subtype B, which is the most abundant subtype in public databases. The downloaded GenBank files were opened and the 50 bases upstream of the eight CDSs (*gag*, *vif*, *vpr*, *tat*, *rev*, *vpu*, *env*, and *nef*) were extracted using the DAMBE program (Xia, 2001; Xia and Xie, 2001).

With 86 HIV-1 genomes and eight protein-coding genes with independent translation initiation in each genome, there are 688 (= 8×86) 50-mer upstream sequences. A small number of these genomes lack specific genes or proper annotation of CDSs (e.g. AF075719 lacking *nef*, U69587 lacking *vpu*). A total of 17 such sequences are missing from our dataset resulting in 671 (= 688-17) 50-mers. A final modification was performed on these sequences, which is explained here. Other than the full-length HIV-1 transcript, coding for Gag and Gag-pol polyproteins, all the other transcripts are composed of different combinations of exons (Figure 1.10C and Table 1-2). Even though there are many different transcripts for genes such as *env* and *rev* with 16 and 12 alternatively spliced transcripts, respectively, the predominant transcript in these genes is the one with the fewest number of upstream non-coding exons 2, 3, and 5 (Purcell and Martin, 1993). The only exception to this is *nef* where Nef 2 transcript (49% relative abundance) is much more abundant than Nef 1 (5% relative abundance) (Table 1-2). This is despite the fact that Nef 2 contains the upstream non-coding exon 5 which is absent in Nef 1 (Purcell and Martin, 1993). All the predominant forms of the eight genes, except *rev*, contain the 50-mer immediately upstream of the start codon. As Table 1-2 shows, the 12 *rev* mRNAs all share exons 1 and 7. Rev 1, 2, and 3 mRNAs lack

both upstream non-coding exons 2 and 3 and Rev 4 to 12 mRNAs lack one or both of these exons. In all cases, alternative forms of exon 4 (4c, 4a, and 4b) are attached to exon 7 which is located at the 3' end. These alternative forms of exon 4 are generated by different competing splicing acceptor sites in the middle of the HIV-1 genome (Figure 1.10B). Figure 4.2 shows the location of these alternative splicing acceptor sites in relation to the *rev* start codon.

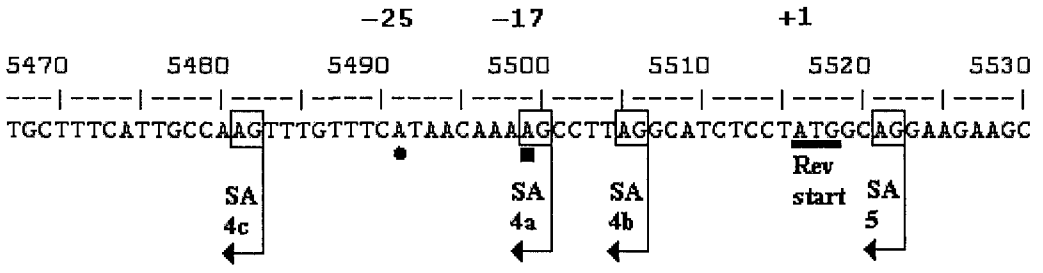


Figure 4.2 Location of splicing acceptor sites (SA) 4c, 4a, 4b, 5, and the Rev start codon on the NCBI HIV-1 type sequence (NC_001802). The region is from nucleotides 5456 to 5532 of the complete genome. The small rectangle and circle show the position of the -17 and -25, respectively, relative to the A in Rev start codon, designated as +1.

When SD3 is used with SA4a or SA4b, the 5'-UTR region of *rev* will contain only 15 or 9 nucleotides, respectively, and does not cover the 50-mer with the -17 and -25. When SD3 is used with SA4c, the 5'-UTR region of *rev* will contain 33 nucleotides upstream of the start codon. We extracted the sequences of all the exons shown in figure 1.10C for the NCBI HIV-1 type (NC_001802). We assembled all these sequences according to the different combinations of the exons in different *rev* transcripts. When exon 3 is present, regardless of which isoform of exon 4 follows after it, both the -17A and -25R are preserved (Rev 7 to 12 in Table 1-2). Also, when exon 2 precedes exons 4c and 4a, the -17A and -25R are preserved (Rev 5 and 6). Overall, in 9 out of 12 Rev transcripts, the -17A and -25R are preserved. Albeit, the two most predominant forms of these transcripts, Rev 1 and Rev 2 (Purcell and

Martin, 1993), do not show a conservation of these two sites. The relative abundance of these transcripts, however, was measured in a condition which does not represent the cellular stress under apoptosis or cell-cycle arrest at which mechanisms other than cap-dependent ribosomal scanning are believed to come into play. It has been proposed that alternative 5'-UTRs of HIV-1 transcripts may contain structures that facilitate the synthesis of the corresponding gene product in a cell cycle-dependent manner (Yilmaz et al., 2006). Indeed, the conservation of -17A and -25R in 9 of 12 *rev* transcripts points to a potential regulatory role for these two sites. The alternatively spliced transcripts are postulated to play regulatory roles instead of being merely redundant (Purcell and Martin, 1993).

Since we are interested in the 50-mer upstream of the HIV-1 CDSs and in particular the -17 and -25 sites discussed above, we used Rev 3 as the representative of these transcripts. This transcript is the third most predominant transcript and contains exon 4c which includes the -17A and -25R conserved sites. As mentioned above, when exon 4c is present, there are 33 nucleotides upstream of the *rev* start codon. We therefore concatenated 17 nucleotides from the 3' end of exon 1, which precedes exon 4c in this transcript, to the 33 nucleotides to form the 50-mers of *rev*, which is used in our dataset.

Out of the 671 sequences in our dataset, there are 331 unique sequences which constitute the HIV-1 dataset in the position weight matrix (Xia, 2007b, pp. 83-92) and perceptron (Xia, 2007b, pp. 93-100) analysis. We used the position weight matrix or PWM, to characterize the sequence feature of the 50-mers. Two hypotheses are explicit in a PWM; one being the null hypothesis assuming no structure in the 50-mers, i.e., the sequences are not different from random assembly of nucleotides based on the nucleotide frequencies, p_A, p_C, p_G and p_T. For example, the likelihood of observing a 50-mer of "ACTG.....AG" is simply

$$L_0 = p_A p_C p_T p_G \cdots p_A p_G \tag{4.1}$$

In contrast, the alternative assumes site-specific nucleotide frequencies so that the likelihood of observing the same 50-mer is

$$L_1 = p_{A1} p_{C2} p_{T3} p_{G4} \cdots p_{A49} p_{G50} \tag{4.2}$$

where p_{A1} is the frequency of A at site 1, p_{C2} the frequency of C at site 2, and so on. The PWM score (PWMS) is the log-odds of the two likelihoods that measures the support of the alternative hypothesis relative to the null, give a particular sequence, e.g., "ACTG.....AG":

$$PWMS = \log_2\left(\frac{L_1}{L_0}\right) = \log_2 \frac{p_{A1}}{p_A} + \log_2 \frac{p_{C2}}{p_C} + \ldots + \log_2 \frac{p_{G50}}{p_G} \tag{4.3}$$

where the individual terms on the right-hand side are the entries in a PWM, designated as PWM_{ij}. An entry favors the alternative when its value is greater than 0, favors the null when its value is smaller than 0, and uninformative when its value is 0. In practice, because p_{ij} values may be zero and because no logarithm is defined for 0, *PWMij* is often computed with pseudocounts such that

$$\begin{aligned} PWM_{ij} &= \log_2 L + \log_2\left(\frac{f_{ij} + f_{i.pseudo}}{f_i + f_{pseudo}}\right) \\ f_{i.pseudo} &= \frac{\alpha f_i}{L} \\ f_{pseudo} &= \sum_{i=1}^{N} f_{i.pseudo} \end{aligned} \tag{4.4}$$

where f_{ij} is the number of nucleotides i at site j, f_i is the number of nucleotides i at all sites, L is the sequence length (50 in our case), α is a small value (set to 0.01 in our study) and N (= 4) is the number of different nucleotides.

Perceptron: Separation of HIV-1 and Human CDSs

Using the same procedure as described in the materials and methods of chapter 2, we downloaded all the unique upstream 50-mers flanking the ATG start codons of human CDSs. As of August 2007, there are a total of 31,484 human genes in Ensembl database, 21,658 of which are known protein-coding genes. Excluding the CDSs with non-ATG start codons, mitochondrial genes, and redundant 5'-UTRs belonging to the same gene, there are a total of 24,850 unique 5'-UTR 50-mers, which constitute the human dataset in the perceptron analysis of this study.

The perceptron (Xia, 2007b, pp. 93-100) is a binary classifier that has been used in bioinformatics to identify the translational initiation sites in *E. coli* (Stormo et al., 1982). The input to perceptron consists of two groups of sequences of the same length and our objective is to find a scoring function to maximize the difference between the two groups. The scoring function is in the form of a weight matrix, derived from training the perceptron with the two groups of sequences. The output from training a perceptron is a weight matrix and a score for each input sequence. A perceptron that has achieved convergence will have sequence scores positive in the POS group and negative in the NEG group. Here the POS group corresponds to the 331 unique HIV-1 sequences and the NEG group corresponds to the set of 24,850 human sequences. Both perceptron and PWM analyses were carried out by using DAMBE (Xia, 2001; Xia and Xie, 2001). Sequence logo is produced by WebLogo (Crooks et al., 2004).

Perceptron: cross-validation

We performed three different approaches of cross-validation as an assessment of the validity of our perceptron results. In the first part we used the delete-half jackknifing method by using the random integer generator function of a Python script to randomly assign half (=166) of HIV-1 sequences into the POS group and half (=12,425) of the human sequences

into the NEG group. We conducted this random assignment of sequences twenty times and each time used the corresponding groups in training the perceptron. In order to see if smaller subsets of the original datasets are sufficient for the convergence of perceptron, we randomly assigned consecutively smaller portions of HIV-1 and human sequences into POS and NEG groups, by repeatedly dividing the number of randomly selected sequences in half (i.e. 25%, 12.5%, 6.25%, and 3.125%). For each portion category, we performed the random assignment of sequences five times. For example, for the 25% category, we randomly distributed 25% of the HIV-1 sequences into the POS group and 25% of the human sequences into the NEG group for five times and each time performed the perceptron on the resulting groups. Finally, to account for any bias in the results due to the disproportionate number of sequences in the HIV-1 (331) and human (24850) datasets, we used the entire set of 331 HIV-1 sequences as the POS group and randomly assigned 331 human sequences into the NEG group. This was performed ten times and each time we performed the perceptron on the resulting groups.

4.4 Results and Discussion

The nucleotide frequencies and the position weight matrix of the 331 unique HIV-1 50-mers are shown in Table 4-1. The most conspicuous site conservation within the 50-mer is the -17 site, with 95.8% being A (Table 4-1 and Figure 4.3). The -25 and -7 sites are also conserved, with 96.4% and 84.3% being purine (Table 4-1 and Figure 4.3). This confirms and generalizes our finding of site conservation among non-homologous HIV-1 genes based on the HIV-1 type genome (NC_001802, Fig. 4.2). One may note that, although nucleotide A has the highest frequency in the HIV-1 genome (= 0.3564 in the type genome NC_001802),

it is not the most frequent nucleotide for most sites in the 50-mer (Fig. 4.3). This background A-usage contrasts dramatically to its overuse at the -17 and -25 sites (Fig. 4.3).

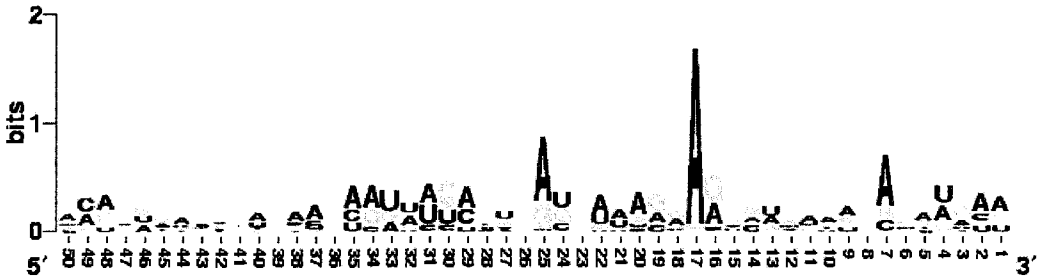


Figure 4.3 Sequence logo of the 331 unique HIV-1 sequences in this study produced by WebLogo (Crooks et al., 2004).

Table 4-1 Site-specific frequency distribution and position weight matrix (computed with the logarithm of base 2) for the 50 nucleotides upstream of the coding genes in 86 HIV-1 genomes. Highest scores for sites -17 and -25 are in bold.

Site	A	C	G	U	A	C	G	U
-50	153	45	51	82	0.3067	-0.2204	-0.7012	0.2564
-49	132	135	48	16	0.0956	1.3533	-0.7872	-2.0522
-48	139	9	125	58	0.1694	-2.4767	0.5782	-0.2382
-47	126	57	94	54	0.0291	0.1171	0.1702	-0.34
-46	83	13	142	93	-0.5658	-1.9711	0.7611	0.4366
-45	92	34	134	71	-0.4194	-0.6194	0.6779	0.0505
-44	151	74	58	48	0.2879	0.4906	-0.5184	-0.5076
-43	99	107	85	40	-0.315	1.0194	0.0264	-0.7666
-42	118	86	92	35	-0.0646	0.7059	0.1395	-0.9557
-41	122	45	79	85	-0.017	-0.2204	-0.0781	0.3078
-40	156	47	42	86	0.3345	-0.1584	-0.9763	0.3246
-39	78	109	94	50	-0.6541	1.046	0.1702	-0.4496
-38	160	71	60	40	0.3707	0.4313	-0.4702	-0.7666
-37	171	76	35	49	0.4659	0.5288	-1.2336	-0.4783
-36	56	94	98	83	-1.1233	0.8335	0.2298	0.2738
-35	163	90	3	75	0.3973	0.7711	-4.4553	0.1288
-34	160	49	116	6	0.3707	-0.099	0.4712	-3.3709
-33	76	6	80	169	-0.691	-3.0221	-0.0602	1.2935
-32	127	24	42	138	0.0404	-1.1127	-0.9763	1.0025
-31	150	37	8	136	0.2784	-0.4992	-3.2528	0.9815
-30	9	40	183	99	-3.6053	-0.3883	1.1251	0.5262
-29	160	116	8	47	0.3707	1.1354	-3.2528	-0.5376
-28	88	94	38	111	-0.4827	0.8335	-1.1177	0.6901
-27	53	41	73	164	-1.2009	-0.3531	-0.1909	1.2504
-26	122	74	79	56	-0.017	0.4906	-0.0781	-0.2882
-25	223	4	96	8	**0.8466**	-3.5497	**0.2003**	-2.9936
-24	4	69	113	145	-4.5711	0.3904	0.4337	1.0735
-23	98	88	58	87	-0.3294	0.7389	-0.5184	0.3411
-22	177	19	46	89	0.5153	-1.4416	-0.8475	0.3737
-21	140	31	46	114	0.1797	-0.7505	-0.8475	0.7284
-20	196	41	50	44	0.6615	-0.3531	-0.7293	-0.6313
-19	100	64	158	9	-0.3007	0.2827	0.9142	-2.8365
-18	120	70	121	20	-0.0406	0.411	0.5316	-1.7423
-17	317	1	10	3	**1.3516**	-5.1158	-2.9586	-4.229
-16	135	35	157	4	0.1277	-0.5782	0.9051	-3.8832
-15	104	52	130	45	-0.2448	-0.0141	0.6345	-0.5994
-14	107	99	112	13	-0.2042	0.9079	0.4209	-2.338
-13	111	13	81	126	-0.1519	-1.9711	-0.0425	0.8719
-12	83	33	127	88	-0.5658	-0.6618	0.601	0.3575
-11	137	30	68	96	0.1487	-0.797	-0.292	0.4821
-10	143	44	94	50	0.21	-0.2525	0.1702	-0.4496
-9	118	14	117	82	-0.0646	-1.8682	0.4835	0.2564
-8	53	88	115	75	-1.2009	0.7389	0.4588	0.1288
-7	225	50	54	2	0.8594	-0.0701	-0.62	-4.6848
-6	44	61	124	102	-1.4627	0.214	0.5667	0.5689
-5	147	38	102	44	0.2495	-0.4613	0.2871	-0.6313
-4	118	0	79	134	-0.0646	-6.6457	-0.0781	0.9603
-3	83	42	166	40	-0.5658	-0.3188	0.9851	-0.7666
-2	162	81	9	79	0.3885	0.6201	-3.0982	0.2031
-1	143	9	111	68	0.21	-2.4767	0.4081	-0.0112

The position weight matrix scores (PWMS) for HIV-1 50-mers (Table 4-2) suggest that the upstream sequences are highly structured. The average PWMS of all 50-mers is 11.99 (Table 4-2). Note that the position weight matrix is for testing two alternative hypotheses, the null being no structure (i.e., all sequences are randomly assembled given the nucleotide frequencies) and the alternative incorporating site-specific nucleotide frequencies. A PWMS of 11.99 means the alternative hypothesis is 4068 ($= 2^{11.99}$) times more likely than the no-structure null hypothesis.

Table 4-2 Summary statistics of position weight matrix scores by HIV-1 genes

Gene	N	Mean	SE
env	44	10.11	0.30
gag	29	10.54	0.58
nef	48	13.84	0.59
rev	38	18.99	0.48
tat	47	5.45	0.43
vif	33	4.46	0.24
vpr	43	15.47	0.51
vpu	49	15.58	1.52
All	331	11.99	0.37

Site conservation is important in two ways. First, understanding why the sites are conserved may lead to better understanding of the molecular biology of HIV-1. In particular, it has been proposed that HIV-1 may have translation mechanisms different from the host, which uses mainly the cap-dependent ribosomal scanning (for a review see Yilmaz et al 2006). Conserved features upstream of the initiation codon in HIV-1 genes may help in the recognition of the translation initiation sites. Second, if the site conservation leads to clear separation of the HIV-1 sequences from the corresponding human sequences, then it is theoretically possible to inhibit HIV-1 translation with little effect on the translation of

human genes. This prompted us to use perceptron to classify the two groups of 50-mer upstream of the translation initiation codon.

Perceptron

A simple perceptron carried out on the two groups of 50-mers (331 sequences from HIV-1 genomes and 24,850 sequences from human) reached convergence after 266 iterations. The resulting weighting matrix (Table 4-3) can separate all HIV-1 50-mers from the 24,850 human 50-mers. The perceptron score is computed as

$$PS = \sum_{j=1}^{L} W_{S_j,j} \tag{4.5}$$

where $W_{Sj,j}$ represent entries in the weight matrix. For example, a 50-mer with the sequence "ACUGGUGAGUACGCCAAAAUUCUUGACUAGCGGAGGCUAGAAGGAGAGAG" would, according to Table 4-3, have its *PS* equal to

$$\begin{aligned} &W_{A1} + W_{C2} + W_{U3} + \ldots + W_{G48} + W_{A49} + W_{G50} \\ &= 53 + 14 + 4 + \ldots + 26 - 22 - 17 = 355 \end{aligned} \tag{4.6}$$

The *PS* values range from 50 to 584 for the 331 HIV-1 50-mers, and -2645 to -51 for the 24,850 human 50-mers. The clear separation of the two groups implies the theoretical possibility of targeting the HIV-1 50-mers to inhibit HIV-1 gene translation initiation with little effect on the translation of human genes. The maximum value in the perceptron weight matrix (PTWM) (Table 4-3) is at the -17 site (118), followed by the -4, -22, -5/-15, and -25 sites with 97, 77, 74, and 71 PTWM values, respectively. This implies that these sites, in particular the -17 and -4, are conserved in HIV-1 but not human genes.

Table 4-3 Weighting matrix from perceptron with the maximum value in each row shown in the last column.

Site	A	C	G	U	Max.
-50	53	-40	-56	-34	53
-49	13	14	15	-119	15
-48	12	-132	39	4	39
-47	-19	-12	-37	-9	-9
-46	-23	-75	44	-23	44
-45	-4	2	8	-83	8
-44	27	-8	-23	-73	27
-43	-6	35	-106	0	35
-42	47	-14	-38	-72	47
-41	5	-55	-29	2	5
-40	39	20	-74	-62	39
-39	-80	-24	45	-18	45
-38	-28	31	-66	-14	31
-37	22	-32	-51	-16	22
-36	-95	13	-17	22	22
-35	-7	56	-55	-71	56
-34	34	-57	41	-95	41
-33	-41	-40	-57	61	61
-32	39	-67	-70	21	39
-31	62	16	-159	4	62
-30	-61	-29	31	-18	31
-29	44	39	-82	-78	44
-28	15	-5	-71	-16	15
-27	-37	-70	-23	53	53
-26	-15	25	-22	-65	25
-25	**71**	-112	**21**	-57	71
-24	-97	30	-37	27	30
-23	-19	-44	-37	23	23
-22	77	-56	-44	-54	77
-21	18	-67	-87	59	59
-20	9	-4	-59	-23	9
-19	-27	-15	48	-83	48
-18	21	-61	36	-73	36
-17	**118**	-121	-60	-14	118
-16	-8	-6	27	-90	27
-15	-73	-11	74	-67	74
-14	-28	6	22	-77	22
-13	66	-174	-2	33	66
-12	14	-86	-29	24	24
-11	-14	-87	-11	35	35
-10	54	-64	-5	-62	54
-9	16	-56	-12	-25	16
-8	-63	21	-13	-22	21
-7	58	-8	12	-139	58
-6	-87	-53	13	50	50
-5	74	-63	17	-105	74
-4	-11	-184	21	**97**	97
-3	-73	29	26	-59	29
-2	-22	-42	-80	67	67
-1	34	-106	-17	12	34

Perceptron Cross-validation

Because of the disproportionate magnitude of HIV-1 (331) and human (24,850) datasets and the relatively small number of sequences in the HIV-1 dataset, there will be some site combinations that, by chance, happen to differ between the two groups. For example, it is expected that some of the sites in table 4-3 have reached high ranking by chance without any biological significance. A random sampling can avoid this issue by nearly eliminating the possibility that such site combinations will always be frequent in one group and rare in the other group.

Table 4-4 shows the results of the delete-half jackknifing. The PTWM values of the six sites with highest values in the original dataset (Table 4-3) are also shown. In all 20 runs of perceptron, convergence was reached after a number of iterations ranging from 57 to 110 iterations. Site -17 is always among the six highest PTWM values and in 85% (17 out of 20) of the cases, it has the maximum PTWM value. Site -25 is among the six highest PS values in 90% (18 out of 20) of the cases. Sites -4 and -22 and especially sites -5 and -15 are not among the six highest PS values in most cases. This implies that sites -17 and -25, mentioned in the introduction of this chapter, are highly conserved in HIV-1 sequences compared to human sequences.

Table 4-4 Delete-half jackknifing. Number of perceptron iterations before the convergence was reached are shown for each of the 20 runs. Light-gray boxes indicate that the site was among the six highest perceptron weight matrix (PTWM) values and the dark-gray boxes indicate that the site had the highest PTWM value.

Perceptron	Number of Iterations before convergence	Nucleotide positions relative to "A" in ATG start codon					
		-17	-4	-22	-15	-5	-25
1st	57						
2nd	80						
3rd	68						
4th	78						
5th	80						
6th	77						
7th	74						
8th	74						
9th	106						
10th	72						
11th	96						
12th	110						
13th	85						
14th	78						
15th	80						
16th	73						
17th	63						
18th	72						
19th	80						
20th	62						

The perceptron also converged on smaller subsets of the original HIV-1 and human datasets; 25% (82 HIV-1 and 6212 human 50-mers), 12.5% (41 HIV-1 and 3106 human 50-mers), 6.25% (20HIV-1 and 1553 human 50-mers), and 3.125% (10 HIV-1 and 776 human 50-mers) (Table 4-5). In all cases site -17 is among the six highest PTWM values and in most cases (17 out of 20), it contains the highest value. With 6.25% and 3.125% of the sequences, the ability of perceptron to discriminate between the POS and NEG groups based on sites -17 and especially -25 diminishes (Table 4-5).

Table 4-5 Perceptron performed on smaller subsets of HIV-1 and human dataset. Perceptron was run five times for each subset category. Light-gray boxes indicate that the site was among the six highest perceptron weight matrix (PTWM) values and the dark-gray boxes indicate that the site had the highest PTWM value.

Dataset Proportion	Perceptron	Number of Iterations before convergence	Nucleotide positions relative to "A" in ATG start codon					
			-17	-4	-22	-15	-5	-25
25%	1st	37						
	2nd	54						
	3rd	45						
	4th	50						
	5th	43						
12.5%	1st	30						
	2nd	31						
	3rd	30						
	4th	31						
	5th	29						
6.25%	1st	24						
	2nd	24						
	3rd	21						
	4th	24						
	5th	25						
3.125%	1st	14						
	2nd	12						
	3rd	12						
	4th	11						
	5th	13						

Since the original human dataset (24,850 sequences) is 75 times bigger than the original HIV-1 dataset (331 sequences), we performed the perceptron using the original HIV-1 dataset as the POS group, and 331 of randomly selected human sequences as the NEG group (Table 4-6). The perceptron converged in all cases after a number of iterations ranging from 18 to 25. Site -17 had the highest PTWM value in all cases and site -25 was among the six highest PTWM values in 80% (8 out of 10) of the cases. Site -22 was also among the six highest PTWM values in 90% of he cases (Table 4-6) which is attributed to the higher proportion of A at site -22 in HIV-1 (0.535), compared to human sequences (0.186).

Table 4-6 Perceptron performed on identical amount of HIV-1 and human sequences. All 331 unique HIV-1 50-mers were used as the POS group, and 331 randomly selected human 50-mers were used as the NEG group in each run. Light-gray boxes indicate that the site was among the six highest perceptron weight matrix (PTWM) values and the dark-gray boxes indicate that the site had the highest PTWM value.

Perceptron	Number of Iterations before convergence	Nucleotide positions relative to "A" in ATG start codon					
		-17	-4	-22	-15	-5	-25
1st	24						
2nd	21						
3rd	20						
4th	25						
5th	23						
6th	24						
7th	23						
8th	18						
9th	23						
10th	24						

Some of the sites in our results obtain high PTWM values because of the aforementioned disproportionate magnitudes of the original datasets or specific random distribution in POS and NEG groups. For example site -15 had the fourth highest value in the original perceptron weight matrix (Table 4-3), but was almost never among the six highest PTWM values in the subsequent perceptron cross-validations (Tables 4-4 to 4-6). Some other sites, however, such as -17 and -25, are not only highly conserved in HIV-1 5'-UTRs but also consistently obtained high PTWM values in all our perceptron weight matrices. The reason that site -25 does not always fall among the six highest PTWM values, could be attributed to the fact that it is a purine (R) that is highly conserved at this site and not nucleotide A alone. The distribution of PTWM scores between A and G, lowers the maximum value that site -25 obtains in the perceptron weight matrix, contributed only by the more frequent A nucleotide (Table 4-3).

As a final evaluation of perceptron performance, we trained it using the first half of the HIV-1 50-mers as the POS group and the first half of human 50-mers as the NEG group to generate a PTWM. We then used this PTWM on the other halves to determine if it can successfully separate them. In a second approach, we used the second half of the HIV-1 50-mers as the POS group and the second half of the human 50-mers as the NEG group to generate a PTWM and then used it to separate the other halves. In both cases, more than half of HIV-1 50-mers, and none of the human 50-mers were categorized into the POS group by the PTWM. This is another validation of the ability of perceptron to separate the HIV-1 and human sequences based on the 5'-UTR 50-mers.

Overall, the results presented here indicate strong site conservation in HIV-1 5'-UTRs that are different from human 5'-UTRs. Experimental verification of the importance of these sites in HIV-1 life cycle, and in particular translation initiation, could pave the path to designing HIV-1 specific antiviral drugs with minimal side effects for human.

Chapter 5 HIV-1 translation initiation and selective pressure against ATG usage in 5'-UTR

5.1 Abstract

Recent reports have shown evidence for mechanism of translation initiation in HIV-1 by cap-independent mechanisms, such as the direct binding of ribosome at internal ribosome entry sites (IRES), distant from the 5' cap. The CDRSM hypothesis of translation initiation in HIV-1 predicts that there should be a selective pressure against ATG usage in optimal context (i.e. containing either or both of -3R and +4G) in the HIV-1 5'-UTR to avoid their erroneous detection by the scanning ribosome which hampers the detection of the true downstream translation initiation codon. It has been proposed that the IRES-dependent mechanism of translation initiation in HIV-1 mRNAs, allows the ribosome to bypass the stable secondary structures in the 5'-UTR. We refer to this IRES-dependent mechanism in the presence of 5'-UTR Stable Secondary Structures (SSS) as SSS/IRES hypothesis. This hypothesis predicts no selective pressure against ATG usage in optimal context in the HIV-1 5'-UTR since any such ATGs would be embedded in the stable secondary structures anyways. Our results show that there is indeed a selective pressure against ATG usage in optimal context in the HIV-1 5'-UTR. This finding supports the cap-dependent translation initiation hypothesis but does not support the prediction of the SSS/IRES hypothesis.

5.2 Introduction

The CDRSM and SSS/IRES-dependent translation initiation hypotheses, described in section 1.3.5.5, have different predictions in regard to selective pressure against ATG usage in the HIV-1 5'-UTR. The CDRSM hypothesis predicts that there should be a selective

pressure against ATG usage in optimal context in the 5'-UTR region. This is to avoid their erroneous detection by the scanning ribosome which hampers the detection of the true downstream translation initiation codon. The SSS/IRES-dependent translation initiation hypothesis, on the other hand predicts no selective pressure against these ATGs because they would be embedded in the secondary structures and not exposed to the ribosome.

In this study we have assessed the predictions of the CDRSM and SSS/IRES-dependent translation initiation hypotheses in HIV-1, with regard to selective pressure against ATG usage in optimal context in the HIV-1 5'-UTR. The +4G of the Kozak consensus has been shown to be important in translation initiation especially in the absence of the -3R (Kozak, 1986b; Kozak, 1987; Kozak, 1997), therefore the optimal contexts of ATG in this study are considered as the presence of either or both -3R and +4G; Rnnatg (ATG_{-3R}), RnnatgG (ATG_{-3R+4G}), and YnnatgG (ATG_{-3Y+4G}), where "atg" is the translation initiation codon and "n" represents any of the four unambiguous nucleotides.

5.3 Materials and Methods

Calculating the observed and expected number of ATGs in optimal context in the 5'-UTR

The NCBI HIV-1 type genome (NCBI Genome: NC_001802), hereafter referred to as $HIV\text{-}1_{NC1802}$, was used to analyze the selective pressure against ATG usage in optimal context in the 5'-UTR region. $HIV\text{-}1_{NC1802}$ is a viral genome (i.e. packaged in the virus particle) and consists of 5'-R-U5-3' in the 5' end and 5'-U3-R-3' in the 3' end. The splice sites position numbering by Purcell and Martin (Purcell and Martin, 1993) corresponds to the HIV-1 vector pNL4-3 (NCBI Genome: AF324493), which is a provirus sequence (i.e. integrated into host genome), with identical repeats, 5'-U3-R-U5-3', at both ends. Their

numbering starts from +1 at the beginning of the U3 region. In order to annotate the splice site locations in HIV-1_{NC1802}, which lacks the initial U3 and terminal U5 regions, we aligned the two sequences using Blast2 program at NCBI (Altschul et al., 1990), and identified the corresponding splice site positions (Figure 1.10B). The position of start and stop codons for the eight HIV-1 ORFs; *gag-pol*, *vif*, *vpr*, *tat*, *rev*, *vpu*, *env*, and *nef*, along with their sequence lengths are shown in Table 5-1.

Table 5-1 The HIV-1 coding sequences. The position of start and stop codons are indicated according to the NCBI HIV-1 type sequence (NCBI Genome: NC_001802). Start position corresponds to the position of A in ATG start codon and stop position corresponds to the third base position of the stop codon.

Gene Name	Start Codon	Stop Codon	Length
gag-pol	336	4642	4307
vif	4587	5165	579
vpr	5105	5396	291
tat	5377	7970	261 (excluding the intron from 5592 to 7924)
rev	5516	8199	351 (excluding the intron from 5592 to 7924)
vpu	5608	5856	249
env	5771	8341	2571
nef	8343	8963	621

There are more than 30 alternatively spliced HIV-1 transcripts (Purcell and Martin, 1993), each with a different combination of exons. We performed our analysis on the 5'-UTR of the most abundant HIV-1 transcript of each gene according to figure 4 of the report published by Purcell and Martin (Purcell and Martin, 1993). Gag has only one transcript, which is the same as the full-length genomic RNA with a 335-nucleotide 5'-UTR. Nef has five different transcripts with Nef2, consisting of exons 1, 5, and 7 (Figure 1.10C), being the most abundant transcript. Rev has twelve different transcripts, with Rev2 made up of exon 1, 4a, and 7, being the most abundant transcript. Tat has four different transcripts in the 2kb mRNA family and four transcripts in the 4kb mRNA family, with Tat1 and Tat5 being the most abundant in the former and latter groups, respectively. These two transcripts, however,

have identical 5'-UTRs, consisting of exon 1 and exon 4 up to the *tat* start codon, which was used in this analysis. Vpr also has two transcripts in the 2kb mRNA family and two in the 4kb mRNA family. The most abundant ones in the two groups, Vpr1 and Vpr3, respectively, share the same 5'-UTR, consisting of exon 1 and exon 3a up to *vpr* start codon, which was used in this study. Env has 16 different transcripts in the 4kb family of mRNAs with Env1, made up of exons 1 and 5E, being by far the most abundant transcript. Vpu always appears with Env on bicistronic mRNAs, therefore Env1 is considered as the most abundant transcript for Vpu as well. Finally, Vif2, consisting of exon 1 and exon 2E up to *vif* start codon, shown as the only transcript for Vif in Purcell and Martin (Purcell and Martin, 1993) report , was used in this analysis. The 5'-UTR of these major transcripts were extracted and their nucleotide frequencies were calculated using the DAMBE program (Xia, 2001; Xia and Xie, 2001). The expected number of ATGs for each sequence, NE_{ATG}, is obtained by the following formula (Xia, 2007b, pp. 4-6):

$$NE_{ATG} = pA * pT * pG * (L - 3 + 1)$$

where L stands for the sequence length and pA, pT, and pG are the frequencies of A, T, and G nucleotides, respectively. The expected number of ATG_{-3R}, ATG_{-3R+4G}, and ATG_{-3Y+4G} were calculated as $pR * NE_{ATG}$ (pR= pA + pG), $pR * NE_{ATG} * pG$, and $pY * NE_{ATG} * pG$ (pY= pC + pT), respectively. The observed numbers of ATGs in the three different optimal contexts were also recorded for each 5'-UTR.

Calculating the observed and expected number of ATGs in optimal context in the HIV-1 concatenated 5'-UTR

As shown in figure 1.10A, except parts of the terminal LTR regions, the rest of the viral genome is made up almost entirely of overlapping coding regions. The only exception is a short 16-nucleotide sequence between the major splice donor site 4 (SD4) and the *vpu*

start codon. It should be noted that due to the complex and overlapping nature of the HIV-1 genome, all non-coding segments, except exon1, overlap with a coding sequence (e.g. non-coding exons 2 and 3 are in the *gag* and *vif* ORFs, respectively). They are, however, part of a 5'-UTR in the major HIV-1 transcripts (see figures 1D, 1E, and 4 in Purcell and Martin 1993).

The concatenated 5'-UTR region was assembled by connecting the following seven segments; exon1 (1:289), beginning of exon2E up to *vif* start codon (4459:4586), beginning of exon3a up to *vpr* start codon (4936:5104), beginning of exon4 up to *tat* start codon (5324:5376), beginning of exon4a up to *rev* start codon (5501:5515), beginning of exon5E up to *env* start codon (5523:5770), and beginning of exon7 up to *nef* start codon (7925:8342). The resulting sequence is 1320 nucleotides long. In addition to the aforementioned analysis, the concatenated 5'-UTR was divided into nucleotide triplets. In order to test the significance of deviation of the observed number of ATG_{-3R} from the expected, two-way tables were constructed which contained the observed and expected numbers of ATG_{-3R} and Non-ATG_{-3R} triplets in the concatenated 5'-UTR, and the following chi-square test was conducted:

$$X^2 = \frac{(NO_{ATG_{-3R}} - NE_{ATG_{-3R}})^2}{NE_{ATG_{-3R}}} + \frac{(NO_{NonATG_{-3R}} - NE_{NonATG_{-3R}})^2}{NE_{NonATG_{-3R}}}$$

where "*NO*" refers to the observed number and "*NE*" refers to the expected number of ATG_{-3R} and Non-ATG_{-3R} triplets in the 5'-UTR.

5.4 Results and Discussion

Selective pressure against ATG usage in optimal context in HIV-1 5'-UTRs

Table 5-2 shows the nucleotide frequencies of the 5'-UTRs in the major HIV-1 transcripts (see methods). With G being the most frequent nucleotide in most cases, these frequencies do not conform to the HIV-1_{NC1802} A-rich genome with 36% A, 18% C, 24% G, and 22% T. The observed and expected numbers of ATGs in optimal context in the 5'-UTR of HIV-1 major transcripts are shown in table 5-3. In all cases, the observed number is less than the expected value. In fact, other than the ATG_{-3R} located in the 5'-UTR of *env*, which corresponds to the upstream *vpu* start codon (GtaATGc), there are no other ATGs in optimal context in the 5'-UTR regions.

Table 5-2 Nucleotide frequencies of the 5' UTR of major HIV-1 mRNAs

Name	A	C	G	T	Sum	pA	pC	pG	pT
gag/pol	82	79	104	70	335	0.2448	0.2358	0.3104	0.209
vif	119	92	122	84	417	0.2854	0.2206	0.2926	0.2014
vpr	129	107	122	100	458	0.2817	0.2336	0.2664	0.2183
tat	85	81	105	71	342	0.2485	0.2368	0.307	0.2076
rev	68	79	90	67	304	0.2237	0.2599	0.2961	0.2204
vpu	98	91	108	77	374	0.262	0.2433	0.2888	0.2059
env	172	109	141	115	537	0.3203	0.203	0.2626	0.2142
nef	214	177	224	161	776	0.2758	0.2281	0.2887	0.2075

Table 5-3 The expected and observed number of ATGs in HIV-1 5'-UTRs of major transcripts in different optimal contexts

Name	NExp. ATG_{-3R}	NObs. ATG_{-3R}	NExp. ATG_{-3R+4G}	NObs. ATG_{-3R+4G}	NExp. ATG_{-3Y+4G}	NObs. ATG_{-3Y+4G}
gag/pol	2.94	0	0.91	0	0.73	0
vif	4.03	0	1.18	0	0.86	0
vpr	4.09	0	1.09	0	0.90	0
tat	2.99	0	0.92	0	0.73	0
rev	2.29	0	0.68	0	0.63	0
vpu	3.19	0	0.92	0	0.75	0
env	5.62	1	1.48	0	1.06	0
nef	7.22	0	2.08	0	1.61	0

Selective pressure against ATG usage in optimal context in HIV-1 concatenated 5'-UTRs

As previously mentioned, the first 289 nucleotides are shared among all HIV-1 5'-UTR sequences. We, therefore, decided to assess the selective pressure against ATG usage in

optimal context in the concatenated non-overlapping 5'-UTR sequence comprised of 1320 nucleotides. This sequence contains 32.6% A, 19.9% C, 26.4% G, and 21.1% T, which match the A-richness and C-poorness of the HIV-1_{NC1802} genome. There are no ATG_{-3R+4G} and ATG_{-3Y+4G} observed in the concatenated 5'-UTR despite the expected numbers of 3.72 and 2.58, respectively, which implies a selective pressure against these ATGs.

In the case of ATG_{-3R}, while the expected number is 14.1, there is only one observed ATG_{-3R} in the concatenated 5'-UTR. The significance of the difference between the observed and expected number in this case was measured by analyzing the sequence as consisting of 440 (1320/3) ATG_{-3R} and Non-ATG_{-3R} triplets (Table 5-4). A χ^2 test revealed that the difference is significant (X^2= 12.57, d.f.= 1, $p<0.0005$), indicating a selective pressure against ATG_{-3R} in the 5'-UTR.

Table 5-4 The expected and observed number of ATG_{-3R} and Non-ATG_{-3R} triplets in the concatenated HIV-1 5'-UTR

	Concatenated 5'-UTR		
	ATG_{-3R} triplets	Non-ATG_{-3R} triplets	Sum
Observed	1	439	440
Expected	14.1	425.9	440

The CDRSM and SSS/IRES-dependent translation initiation hypotheses of HIV-1 mRNAs, have different predictions regarding the selective pressure against ATG usage in optimal contexts in the 5'-UTR. The CDRSM hypothesis predicts a selective pressure against the use of these ATGs in the 5'-UTR to avoid their erroneous detection by the scanning ribosome, whereas the SSS/IRES-dependent hypothesis predicts an absence of such selective pressure because such ATGs would be embedded in the stable secondary structures. Our results imply a selective pressure against the ATG_{-3R} and suggest possible selection against

ATG_{-3R+4G} and ATG_{-3Y+4G} in the 5'-UTR region. This is in agreement with the CDRSM hypothesis of translation initiation in HIV-1.

While there are some experimental results indicating the presence of SSS/IRES in HIV-1 (Brasey et al., 2003; Buck et al., 2001), currently used methodology for validating IRES remains problematic (Kozak, 2001). Our results suggest that SSS/IRES-dependent translation initiation must either be a weak mechanism or a new mechanism that has not yet produced evolutionary consequences on the distribution of ATG in optimal context in 5'-UTR.

Our results presented in this study clearly indicate a selective pressure against the ATG usage in optimal context in the HIV-1 5'-UTR. This finding supports the cap-dependent ribosomal scanning translation initiation hypothesis and challenges the prediction of SSS/IRES-dependent translation initiation hypothesis in HIV-1.

Chapter 6 Bioinformatic approach to identify penultimate amino acids efficient for N-terminal methionine excision

Published in the 1st International Conference on Bioinformatics and Biomedical Engineering (ICBBE 2007) (Khalouei et al., 2007). This chapter is identical to the published version.

6.1 Abstract

More than half of proteins in prokaryotes and eukaryotes undergo N-terminal methionine excision (NME). While it is known that the penultimate amino acid affects the efficiency of NME in several bacterial and eukaryotic species, it is experimentally difficult and tedious to verify which amino acid at the penultimate site (the site after initiator Met) is the most efficient for NME in different species. Here we present a new bioinformatic approach to identify penultimate amino acids that are efficient for NME. Amino acids most efficient for NME are alanine, serine and proline in human, and alanine, glycine, valine, proline and serine in the yeast *Saccharomyces cerevisiae*. This finding also helps resolve the two hypotheses that have been proposed to explain the presence of +4G site in the Kozak consensus for translation initiation.

6.2 Introduction

N-terminal modifications of nascent peptides occur in more than half of proteins in both prokaryotes and eukaryotes (Giglione et al., 2004; Giglione et al., 2003; Meinnel et al., 1993; Serero et al., 2003) as well as in mitochondria and plastids (Giglione et al., 2004). N-terminal methionine excision (NME), which occurs soon after the amino terminus of the growing polypeptide chain emerges from the ribosome, is not only an important amino-

terminal modification in itself, but also required for further N-terminal modifications. For example, it is required for myristoylation where glycine at the amino terminus, after the removal of the initiator methionine, is needed to attach to a myristoyl ($C_{14}H_{28}O_2$) fatty acid side chain (Farazi et al., 2001).

NME is carried out by methionine aminopeptidase (MAP). Eubacteria contain only one type of MAP whereas eukaryotes contain two (MAP1 and MAP2). The efficiency of NME depends heavily on the penultimate (the second) amino acid. In the yeast, *Saccharomyces cerevisiae*, NME occurs most efficiently when the penultimate amino acid is small (Moerschell et al., 1990). Such studies have contributed significantly to the understanding of not only NME itself, but also eukaryotic translation initiation.

The optimal context for translation initiation in mammalian species is GCC**R**CCaug**G** (where R = purine and "aug" is the initiation codon), with the -3R and +4G being particularly important (Kozak, 1997; Kozak, 1999). The presence of +4G has been interpreted as necessary for efficient translation initiation (Kozak, 1986b; Kozak, 1997), and this interpretation is featured in virtually all textbooks of molecular biology. The finding that a small amino acid is needed to facilitate NME leads to an alternative hypothesis invoking amino acid constraint (Xia, 2007a). Because alanine and glycine happen to be the smallest amino acids, we expect the initiator Met to be followed often by alanine or glycine. The resulting overuse of Ala and Gly codons (GCN and GGN) following the initiation codon AUG leads to the prevalence of +4G in protein-coding genes. An extensive bioinformatic study (Xia, 2007a) lends strong support for this alternative hypothesis.

In spite of the scientific importance of NME, characterizing the efficiency of NME conferred by different amino acids at the penultimate site in different species is experimentally difficult and tedious. For this reason, only a few studies have been carried out

in *Escherichia coli* (Frottin et al., 2006), *Saccharomyces cerevisiae* (Moerschell et al., 1990) or both (Flinta et al., 1986). In this chapter we propose a bioinformatic approach to characterize the efficiency of NME conferred by different amino acids at the penultimate site and apply the method to the study of human and yeast proteins.

6.3 Methods and Materials

For a genome with N protein-coding genes (representing nascent proteins before any N-terminal processing), there are N amino acids at the penultimate site, with its frequency distribution specified by N_i where i = 1, 2, ..., 20 corresponding to the 20 amino acids. Suppose we have M proteins known to undergo NME, with its frequency distribution specified by M_i. Define $p_i = N_i/N$ and $q_i = M_i/M$. If all amino acids at the penultimate site lead to equal efficiency in NME, then we expect $q_i = p_i$. If amino acid i is very conducive to NME, then $q_i > p_i$ and vice versa.

We use $E_{NME\cdot i}$ defined below as a quantitative measure of the NME efficiency for amino acid i

$$E_{NME\cdot i} = \log_2 \frac{q_i}{p_i} = \log_2 \frac{M_i}{N_i} + \log_2 \frac{N}{M} \qquad (6.1)$$

The interpretation of $E_{NME\cdot i}$ is straightforward: $E_{NME\cdot i} = 0$ when $q_i = p_i$, $E_{NME\cdot i} > 0$ when $q_i > p_i$; $E_{NME\cdot i} < 0$ when $q_i < p_i$.

To obtain N_i for human, we retrieved the rna.gbk.gz file at ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/RNA/, dated Sept. 3, 2006, and extracted all 34,169 annotated coding sequences (CDSs). For the yeast, we retrieved the orf_coding.fasta from NCBI which contained 5888 yeast CDSs. Translating these CDSs to amino acid

sequences and computing the N_i values were done with DAMBE (Xia, 2001; Xia and Xie, 2001).

To obtain M_i for human and yeast, we extracted all the N-terminal methionine-cleaved protein sequences from the UniProtKB/Swiss-Prot database (Wu et al., 2006), by using the search interface at (http://us.expasy.org/srs5). We limited the results to reviewed sequences in Swiss-Prot database by excluding the computationally annotated sequences in TrEMBL database. *Saccharomyces cerevisiae* and *Homo sapiens* were used respectively as the species name, and "INIT_MET" as the "FtKey (Feature)" in the info menu. According to the Swiss-Prot specifications, INIT_MET indicates evidence of NME. From a total of 6093 Swiss-Prot yeast protein sequences, 267 of them contained evidence for initiator methionine cleavage. These 267 sequences were manually inspected. Those proteins without direct experimental evidence and being flagged as "Potential", "Probable", or "By similarity" are not excluded. The remaining 232 proteins were experimentally verified to undergo NME. These 232 sequences were downloaded in FASTA format and the penultimate amino acid frequency was obtained using DAMBE (Xia, 2001; Xia and Xie, 2001). The same was done for human with 484 proteins having experimental evidence of NME.

6.4 Results and Discussion

For yeast proteins, $q_i > p_i$ for alanine (Ala), glycine (Gly), valine (Val), proline (Pro), serine (Ser) and threonine (Thr), i.e., these amino acids are overrepresented in the proteins that have undergone NME (Figure 6.1). The list of amino acids, ranked by their $E_{NME\cdot i}$ values, is shown in columns 2-4 in Table 6-1. The six amino acids with positive $E_{NME\cdot i}$ values (Ala, Gly, Val, Pro, Ser and Thr), with radii of gyration of 1.29 Å or less, were also

found experimentally to result in complete cleavage of initiator Met in yeast (Moerschell et al., 1990). Cys was also found previously to result in complete cleavage of the initiator Met (Moerschell et al., 1990), but no protein with Cys at the penultimate site is represented in our yeast proteins with known experimental evidence of NME. This finding, however, does not reject the hypothesis that Cys is as efficient as other six amino acids with positive $E_{NME\cdot i}$ values because of the rarity of Cys at the penultimate site in proteins encoded in the yeast genome (only 38 out of 5888 proteins encoded in yeast genome, Table 6-1). We need more data to evaluate whether Cys at the penultimate site can lead to efficient NME *in vivo* in yeast.

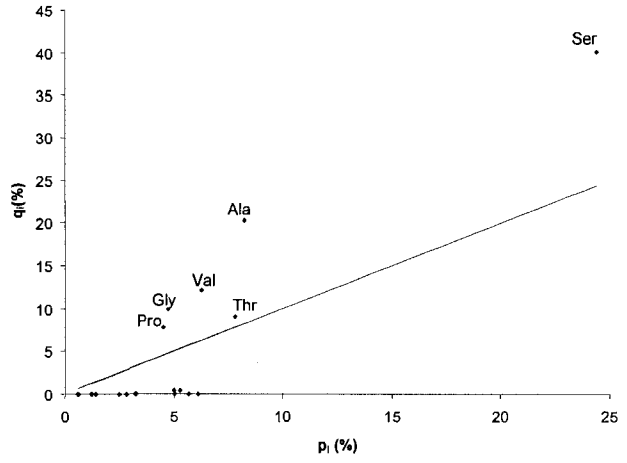


Figure 6.1 Different amino acids at the penultimate site result in dramatically different NME Efficiencies in yeast proteins. p_i and q_i are proportions of amino acid i at the penultimate site of unprocessed proteins and NME-processed proteins, respectively. The line indicates the position when $q_i = p_i$.

Cys is rare not only at the penultimate site but also at other sites in yeast proteins, accounting for only 0.645% of the amino acids at the penultimate and 1.263% at other sites excluding the first two sites. This may be related to the fact that the *S. cerevisiae* genome contains only 4 $tRNA^{Cys}$ genes, the fewest among all 20 amino acids. Because the number of tRNA genes in the genome is strongly correlated with the tRNA concentration in the cell

(Duret, 2000; Ikemura, 1992; Kanaya et al., 1999; Percudani et al., 1997), the few $tRNA^{Cys}$ genes implies relatively few $tRNA^{Cys}$ in the cytoplasm to translate Cys codons. This should result in selection against Cys usage in proteins because its usage would decrease translation efficiency. It has been theoretically suggested and empirically substantiated that tRNA availability is linearly correlated with the square root of amino acid usage (Xia, 1998). Because translation initiation is often the rate-limiting step during protein translation (Bulmer, 1991; Liljenstrom and von Heijne, 1987), it is disadvantageous to code for an amino acid that is slow to translate.

Ser is the most frequently found amino acid at the penultimate site (accounting for 24.4% of all amino acids at the penultimate site, Table 6-1). This is partially explained by the fact that it is also the most frequently used amino acids at other sites in yeast proteins among those amino acids with positive $E_{NME\cdot i}$ values, accounting for 8.97% of all amino acids excluding the first two amino acids (i.e., the initiator Met and the penultimate amino acid) in the proteins. In contrast, Ala, Gly, Val, Pro, and Thr account for only 5.50%, 4.98%, 5.57%, 4.38%, and 5.93%, respectively. It is likely that Ser is more abundant, and consequently translated faster, than other amino acids with positive $E_{NME\cdot i}$ values, which would account for its overuse at the penultimate site to accelerate the movement of the ribosome downstream.

Table 6-1 Details of computing NME efficiency ($E_{NME\cdot i}$) for amino acid i, based on yeast and human proteins. AA – amino acids in 3-letter code, N_i – number of amino acid i at the penultimate site of proteins before any N-terminal processing, M_i – number of amino acid i in the penultimate site of proteins known to undergo NME. $E_{NME\cdot i}$ is specified in Eq. (6.1).

	Yeast			*Human*		
AA	N_i	M_i	$E_{NME\cdot i}$	N_i	M_i	$E_{NME\cdot i}$
Ala	484	47	1.3013	6904	204	1.0548
Gly	278	23	1.0702	2983	59	0.4757
Val	370	28	0.9416	1576	25	0.1574
Pro	265	18	0.7857	1955	53	0.9306
Ser	1437	93	0.7159	3664	108	1.0513
Thr	459	21	0.2155	1674	32	0.4265
Asn	296	1	-3.5439			
Glu	311	1	-3.6152			
Arg	167	0		1788	1	-4.6685
Asp	297	0				
Cys	38	0		376	2	-1.4190
Gln	147	0				
His	71	0				
Ile	191	0				
Leu	360	0				
Lys	335	0				
Met	84	0		495	2	-1.8157
Phe	190	0				
Trp	35	0				
Tyr	73	0				

An alternative explanation for the overuse of Ser at the penultimate site invokes the recognition of translation initiation signal. The consensus sequence including the initiation codon aug is (A/U)A(A/C)A(A/C)AaugUC(U/C) for highly expressed yeast genes. If the +4U, +5C and/or +6Y sites are part of the recognition sequence for the ribosome-dependent scanning model of translation initiation (Kozak, 1989; Kozak, 1992; Kozak, 1997; Kozak, 1999), then the use of Ser at the penultimate site will be increased as a consequence because UCY codons code for Ser. However, this explanation appears unnecessary. After all, it is tenuous to argue that, while +4U is important for translation initiation in yeast, it is +4G that is important for translation initiation in mammals. Empirical evidence suggests that the +4 site is not really important for translation initiation (Xia, 2007a).

Given that most proteins undergo NME (Giglione et al., 2004; Giglione et al., 2003; Meinnel et al., 1993; Serero et al., 2003), one naturally would expect that most proteins should have a small amino acid with a positive $E_{NME\cdot i}$ at the penultimate site. This is true, the sum of N_i values for the six amino acids with $E_{NME\cdot i} > 0$ accounts for 55.93% of all amino acids at the penultimate site in yeast (Table 6-1). Given that Ala, Gly and Val are all coded by G-starting codons, the over-representation of these amino acids at the penultimate site is sufficient to explain the presence of +4G site in protein-coding genes.

For human proteins, $q_i > p_i$ for Ala, Gly, Ser, Pro, Thr and Val (Figure 6.2). This list of six amino acids is the same as that in the yeast (Figure 6.1). In contrast to the yeast proteins where Ser is used most often at the penultimate site, human proteins have Ala as the most frequently used amino acid at the penultimate site, with Ser being the second (Figure 6.2). $E_{NME\cdot i}$ values and the computational details are shown in the last three columns of Table 6-1.

The reason for Ala to be found more frequently at the penultimate site than Ser (Figure 6.2 and Table 6-1) may be related to the relative availability of $tRNA^{Ala}$ and $tRNA^{Ser}$. There are 43 $tRNA^{Ala}$ genes and only 28 $tRNA^{Ser}$ genes in the human genome (http://lowelab.ucsc.edu/GtRNAdb).

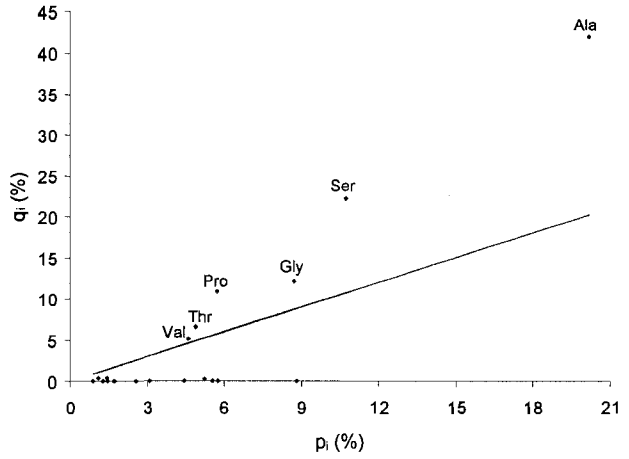


Figure 6.2 Different amino acids at the penultimate site result in dramatically different NME efficiencies in human proteins, with the meaning of symbols as in Figure 6.1.

There are several differences between the yeast and the human results. First, although eukaryotes have both MAP1 and MAP2 proteins, MAP1 appears to be the dominant isoform whose loss leads to a dramatic decrease in growth whereas loss of MAP2 decreases growth only slightly (Li and Chang, 1995). In contrast, MAP2 is more important than MAP1 in higher eukaryotes (Giglione et al., 2004). MAP2 is less efficient than MAP1 in NME when the penultimate site is occupied by Gly. A yeast MAP1 mutant that over-expressed MAP2 proteins can restore most of the NME activity observed in the wild-type except for the test peptide with Gly at the penultimate site (Table I in Li and Chang, 1995). This implies that the $E_{NME.i}$ value for Gly should be smaller in human (where MAP2 is more important) than in yeast (where MAP1 is more important). Our result confirms this prediction, with $E_{NME.i}$ = 1.0702 and 0.4757, respectively, for yeast and human (Table 6-1).

In conclusion, we have demonstrated the utility of a bioinformatic approach to characterize NME efficiencies of different amino acids at the penultimate site that can be adapted for any new species with empirical NME data.

Chapter 7 Concluding remarks

HIV-1 is a complex retrovirus, which has evolved various mechanisms to not only evade the host immune system and potent antiviral drugs, but also to compete for and utilize host cellular resources for the completion of its life cycle. Even though the HIV-1 genome codes for important structural, enzymatic, and regulatory proteins, it is not large enough to encode proteins necessary for transcription and translation of the viral genes.

An efficient combat against the AIDS epidemic requires a clear understanding of different stages of HIV-1 life cycle from cell entry to virion maturation. Transcription and translation are two components of the virus life cycle for which the virus is almost entirely dependent on the host cell machinery. Although this dependence provides various opportunities to interrupt the HIV-1 life cycle, the involvement of transcription and translation molecules in human protein production, necessitate a more cautious approach compared to other virus-specific processes such as reverse transcription and integration. Despite the more than 200,000 available HIV-related publications, and the importance of this stage of the viral life cycle, translation initiation mechanism has not received as much attention as other areas of HIV-1 research.

The main focus of this thesis has been the translation initiation mechanism in HIV-1 virus. Due to the translation of HIV-1 transcripts by the host ribosomes, one of the first steps in the elucidation of this mechanism is a clear understanding of the Kozak consensus and the role it plays in translation initiation. The +4G site of the Kozak consensus has been generally accepted to be involved in translation initiation and is presented as such in most published biology textbooks. This claim, however, has been controversial. Using bioinformatics approaches we provided strong support for the alternative penultimate amino acid constraint

hypothesis. This alternative hypothesis explains the prevalence of +4G by invoking the observation that small amino acids, coded by G-starting codons, which are efficient for NME, are preferred at the penultimate position.

We then assessed the prediction of the two prevailing hypotheses regarding the translation initiation mechanism in HIV-1, namely, CDRSM and SSS/IRES-dependent hypotheses. Two lines of evidence in our studies support the CDRSM hypothesis. First, a high level of conservation of Kozak consensus in a large number of HIV-1 sequences from the five most abundant subtypes worldwide, and second, a selective pressure against ATG usage in optimal context in the HIV-1 5'-UTR. The latter finding also implies that SSS/IRES-dependent translation initiation in HIV-1 is a weak mechanism or a new mechanism that has not yet produced evolutionary consequences on the distribution of ATGs in the HIV-1 5'-UTR.

In conclusion, more computational and experimental verification are required to demonstrate whether the +4G of Kozak consensus plays a role in translation initiation, both in human and HIV-1. Our Bioinformatic approach, presented in chapter 6, will help the researchers with the difficult task of experimental identification of penultimate amino acids efficient for N-terminal methionine excision. The HIV-1 related results presented in this thesis, followed by experimental verification, can help in elucidating the mechanisms of translation initiation in HIV-1. Repeating Brasey *et. al* (Brasey et al., 2003) and Buck *et. al* (Buck et al., 2001) experiments, where IRES activity have been reported in HIV-1, by using modified constructs that contain multiple ATGs in optimal context in the 5'-UTR can provide a better understanding of relative contribution of CDRSM and SSS/IRES-dependent translation initiation. Monitoring the highly conserved sites in mutant HIV-1 strains can also be useful in understanding their potential role in translation initiation. These findings can

help in suppressing the translation initiation and protein production in HIV-1 viruses, even in resting subset of CD4+ T-cells which is currently one of the major obstacles in controlling the AIDS progression in infected individuals.

It is also important to find out if T-cells have unique tRNA pools. One way to get an indirect answer is to compile proteins produced in large quantities in T-cells and study the codon usage bias of these highly expressed proteins. An efficient translation elongation in HIV-1 requires the adaptation of its genes to exploit the specific tRNA pool of the CD4+ T-cells.

Contribution of Collaborators

Dr. Xuhua Xia instructed and guided the experiments and helped with the revision and modification of all parts of this thesis. Chapters 2 and 6 were the result of collaboration among Dr. Xia's laboratory members and Jan Mennigen from Dr. Trudeau's laboratory at the University of Ottawa. In both chapters Sam Khalouei and Xiaoquan Yao downloaded the sequences, compiled and analyzed the data, and contributed equally to the formatting and revision of the reports. Malisa Carullo, Pinchao Ma, Ziyu Song, Huiling Xiong, and Jan Mennigen helped with different aspects of these studies including sequence retrieval and compilation and revision of the reports.

References

Aldovini, A., and R. A. Young. 1990. Mutations of RNA and protein sequences involved in human immunodeficiency virus type 1 packaging result in production of noninfectious virus. J Virol 64:1920-6.

Allan, J. S., J. E. Coligan, F. Barin, M. F. McLane, J. G. Sodroski, C. A. Rosen, W. A. Haseltine, T. H. Lee, and M. Essex. 1985. Major glycoprotein antigens that induce antibodies in AIDS patients are encoded by HTLV-III. Science 228:1091-4.

Aloia, R. C., F. C. Jensen, C. C. Curtain, P. W. Mobley, and L. M. Gordon. 1988. Lipid composition and fluidity of the human immunodeficiency virus. Proc Natl Acad Sci U S A 85:900-4.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. J Mol Biol 215:403-10.

Anderson, E. C., and A. M. Lever. 2006. Human immunodeficiency virus type 1 Gag polyprotein modulates its own translation. J Virol 80:10478-86.

Anderson, J. L., A. T. Johnson, J. L. Howard, and D. F. Purcell. 2007. Both linear and discontinuous ribosome scanning are used for translation initiation from bicistronic human immunodeficiency virus type 1 env mRNAs. J Virol 81:4664-76.

Arrigo, S. J., S. Weitsman, J. A. Zack, and I. S. Chen. 1990. Characterization and expression of novel singly spliced RNA species of human immunodeficiency virus type 1. J Virol 64:4585-8.

Arthur, L. O., J. W. Bess, Jr., R. C. Sowder, 2nd, R. E. Benveniste, D. L. Mann, J. C. Chermann, and L. E. Henderson. 1992. Cellular proteins bound to immunodeficiency viruses: implications for pathogenesis and vaccines. Science 258:1935-8.

Arya, S. K., and R. C. Gallo. 1986. Three novel genes of human T-lymphotropic virus type III: immune reactivity of their products with sera from acquired immune deficiency syndrome patients. Proc Natl Acad Sci U S A 83:2209-13.

Ayouba, A., S. Souquieres, B. Njinku, P. M. Martin, M. C. Muller-Trutwin, P. Roques, F. Barre-Sinoussi, P. Mauclere, F. Simon, and E. Nerrienet. 2000. HIV-1 group N among HIV-1-seropositive individuals in Cameroon. Aids 14:2623-5.

Baltimore, D. 1992. Viral RNA-dependent DNA polymerase. 1970. Biotechnology 24:3-5.

Barre-Sinoussi, F. 1983. Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). Science 220:868-871.

Battula, N., and L. A. Loeb. 1976. On the fidelity of DNA replication. Lack of exodeoxyribonuclease activity and error-correcting function in avian myeloblastosis virus DNA polymerase. J Biol Chem 251:982-6.

Baudin, F., R. Marquet, C. Isel, J. L. Darlix, B. Ehresmann, and C. Ehresmann. 1993. Functional sites in the 5' region of human immunodeficiency virus type 1 RNA form defined structural domains. J Mol Biol 229:382-97.

Beerens, N., and B. Berkhout. 2002. The tRNA primer activation signal in the human immunodeficiency virus type 1 genome is important for initiation and processive elongation of reverse transcription. J Virol 76:2329-39.

Bentham, M., S. Mazaleyrat, and M. Harris. 2006. Role of myristoylation and N-terminal basic residues in membrane association of the human immunodeficiency virus type 1 Nef protein. J Gen Virol 87:563-71.

Berger, E. A., P. M. Murphy, and J. M. Farber. 1999. Chemokine receptors as HIV-1 coreceptors: roles in viral entry, tropism, and disease. Annu Rev Immunol 17:657-700.

Berkhout, B. 1996. Structure and function of the human immunodeficiency virus leader RNA. Prog Nucleic Acid Res Mol Biol 54:1-34.

Berkhout, B., R. H. Silverman, and K. T. Jeang. 1989. Tat trans-activates the human immunodeficiency virus through a nascent RNA target. Cell 59:273-82.

Berkhout, B., and J. L. van Wamel. 2000. The leader of the HIV-1 RNA genome forms a compactly folded tertiary structure. Rna 6:282-95.

Berlioz, C., and J. L. Darlix. 1995. An internal ribosomal entry mechanism promotes translation of murine leukemia virus gag polyprotein precursors. J Virol 69:2214-22.

Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. 2000. The Protein Data Bank. Nucleic Acids Res 28:235-42.

Birney, E., T. D. Andrews, P. Bevan, M. Caccamo, Y. Chen, L. Clarke, G. Coates, J. Cuff, V. Curwen, T. Cutts, T. Down, E. Eyras, X. M. Fernandez-Suarez, P. Gane, B. Gibbins, J. Gilbert, M. Hammond, H. R. Hotz, V. Iyer, K. Jekosch, A. Kahari, A. Kasprzyk, D. Keefe, S. Keenan, H. Lehvaslaiho, G. McVicker, C. Melsopp, P. Meidl, E. Mongin, R. Pettett, S. Potter, G. Proctor, M. Rae, S. Searle, G. Slater, D. Smedley, J. Smith, W. Spooner, A. Stabenau, J. Stalker, R. Storey, A. Ureta-Vidal, K. C. Woodwark, G. Cameron, R. Durbin, A. Cox, T. Hubbard, and M. Clamp. 2004. An overview of Ensembl. Genome Res 14:925-8.

Blankson, J. N., D. Persaud, and R. F. Siliciano. 2002. The challenge of viral reservoirs in HIV-1 infection. Annu Rev Med 53:557-93.

Bodilis, J., and S. Barray. 2006. Molecular evolution of the major outer-membrane protein gene (oprF) of Pseudomonas. Microbiology 152:1075-88.

Booth, J. C., U. Kumar, D. Webster, J. Monjardino, and H. C. Thomas. 1998. Comparison of the rate of sequence variation in the hypervariable region of E2/NS1 region of hepatitis C virus in normal and hypogammaglobulinemic patients. Hepatology 27:223-7.

Bouamr, F., S. Scarlata, and C. Carter. 2003. Role of myristylation in HIV-1 Gag assembly. Biochemistry 42:6408-17.

Brasey, A., M. Lopez-Lastra, T. Ohlmann, N. Beerens, B. Berkhout, J. L. Darlix, and N. Sonenberg. 2003. The leader of human immunodeficiency virus type 1 genomic RNA harbors an internal ribosome entry segment that is active during the G2/M phase of the cell cycle. J Virol 77:3939-49.

Breuer, S., H. Gerlach, B. Kolaric, C. Urbanke, N. Opitz, and M. Geyer. 2006. Biochemical indication for myristoylation-dependent conformational changes in HIV-1 Nef. Biochemistry 45:2339-49.

Buck, C. B., X. Shen, M. A. Egan, T. C. Pierson, C. M. Walker, and R. F. Siliciano. 2001. The human immunodeficiency virus type 1 gag gene encodes an internal ribosome entry site. J Virol 75:181-91.

Bukrinsky, M. I., N. Sharova, M. P. Dempsey, T. L. Stanwick, A. G. Bukrinskaya, S. Haggerty, and M. Stevenson. 1992. Active nuclear import of human immunodeficiency virus type 1 preintegration complexes. Proc Natl Acad Sci U S A 89:6580-4.

Bulmer, M. 1991. The selection-mutation-drift theory of synonymous codon usage. Genetics 129:897-907.

Cann, A. J., and J. Karn. 1989. Molecular biology of HIV: new insights into the virus life-cycle. Aids 3 Suppl 1:S19-34.

Cassan, M., N. Delaunay, C. Vaquero, and J. P. Rousset. 1994. Translational frameshifting at the gag-pol junction of human immunodeficiency virus type 1 is not increased in infected T-lymphoid cells. J Virol 68:1501-8.

Cavener, D. R. 1987. Comparison of the consensus sequence flanking translational start sites in Drosophila and vertebrates. Nucleic Acids Res 15:1353-61.

Chermann, J. C., F. Barre-Sinoussi, C. Dauguet, F. Brun-Vezinet, C. Rouzioux, W. Rozenbaum, and L. Montagnier. 1983. Isolation of a new retrovirus in a patient at risk for acquired immunodeficiency syndrome. Antibiot Chemother 32:48-53.

Chohan, B., L. Lavreys, S. M. Rainwater, and J. Overbaugh. 2005. Evidence for frequent reinfection with human immunodeficiency virus type 1 of a different subtype. J Virol 79:10701-8.

Chun, T. W., D. Engel, M. M. Berrey, T. Shea, L. Corey, and A. S. Fauci. 1998. Early establishment of a pool of latently infected, resting CD4(+) T cells during primary HIV-1 infection. Proc Natl Acad Sci U S A 95:8869-73.

Chun, T. W., and A. S. Fauci. 1999. Latent reservoirs of HIV: obstacles to the eradication of virus. Proc Natl Acad Sci U S A 96:10958-61.

Cigan, A. M., and T. F. Donahue. 1987. Sequence and structural features associated with translational initiator regions in yeast--a review. Gene 59:1-18.

Clark, S., C. Calef, and J. Mellors. 2005. Mutations in Retroviral Genes Associated with Drug Resistance. Pages 80-175 *in* HIV Sequence Compendium 2005 (L. H. Thomas, B. Foley, B. Hahn, P. Marx, F. McCutchan, J. Mellors, S. Wolinsky, and B. Korber, eds.). Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM. LA-UR 06-0680.

Cleaveland, S., M. K. Laurenson, and L. H. Taylor. 2001. Diseases of humans and their domestic mammals: pathogen characteristics, host range and the risk of emergence. Philos Trans R Soc Lond B Biol Sci 356:991-9.

Crooks, G. E., G. Hon, J. M. Chandonia, and S. E. Brenner. 2004. WebLogo: a sequence logo generator. Genome Res 14:1188-90.

Cullen, B. R. 1986. Trans-activation of human immunodeficiency virus occurs via a bimodal mechanism. Cell 46:973-82.

Daar, E. S., T. Moudgil, R. D. Meyer, and D. D. Ho. 1991. Transient high levels of viremia in patients with primary human immunodeficiency virus type 1 infection. N Engl J Med 324:961-4.

Dalgleish, A. G., P. C. Beverley, P. R. Clapham, D. H. Crawford, M. F. Greaves, and R. A. Weiss. 1984. The CD4 (T4) antigen is an essential component of the receptor for the AIDS retrovirus. Nature 312:763-7.

Dayton, A. I., J. G. Sodroski, C. A. Rosen, W. C. Goh, and W. A. Haseltine. 1986. The trans-activator gene of the human T cell lymphotropic virus type III is required for replication. Cell 44:941-7.

Dayton, A. I., E. F. Terwilliger, J. Potz, M. Kowalski, J. G. Sodroski, and W. A. Haseltine. 1988. Cis-acting sequences responsive to the rev gene product of the human immunodeficiency virus. J Acquir Immune Defic Syndr 1:441-52.

De Clercq, E. 1998. New perspectives for the treatment of HIV infections. Verh K Acad Geneeskd Belg 60:13-41; discussion 41-5.
de Vries, J. S., V. M. Andriotis, A. J. Wu, and J. P. Rathjen. 2006. Tomato Pto encodes a functional N-myristoylation motif that is required for signal transduction in Nicotiana benthamiana. Plant J 45:31-45.
Debouck, C. 1992. The HIV-1 protease as a therapeutic target for AIDS. AIDS Res Hum Retroviruses 8:153-64.
Deffaud, C., and J. L. Darlix. 2000. Rous sarcoma virus translation revisited: characterization of an internal ribosome entry segment in the 5' leader of the genomic RNA. J Virol 74:11581-8.
Dickson, C., R. Eisenman, H. Fan, E. Hunter, and N. Teich. 1984. Protein biosynthesis and assembly. Pages 513-640 *in* Molecular Biology of Tumor Viruses-RNA Tumor Viruses (R. A. Weiss, N. Teich, H. E. Varmus, and J. M. Coffin, eds.). Cold Spring Harbor, NY.
Doms, R. W., and S. C. Peiper. 1997. Unwelcomed guests with master keys: how HIV uses chemokine receptors for cellular entry. Virology 235:179-90.
Dorman, N., and A. Lever. 2000. Comparison of viral genomic RNA sorting mechanisms in human immunodeficiency virus type 1 (HIV-1), HIV-2, and Moloney murine leukemia virus. J Virol 74:11413-7.
Duret, L. 2000. tRNA gene number and codon usage in the C. elegans genome are co-adapted for optimal translation of highly expressed genes. Trends Genet 16:287-9.
Duret, L. 2002. Evolution of synonymous codon usage in metazoans. Curr Opin Genet Dev 12:640-9.
Farazi, T. A., G. Waksman, and J. I. Gordon. 2001. The biology and enzymology of protein N-myristoylation. J Biol Chem 276:39501-4.
Fauci, A. S. 1993. Multifactorial nature of human immunodeficiency virus disease: implications for therapy. Science 262:1011-8.
Fischer, U., S. Meyer, M. Teufel, C. Heckel, R. Luhrmann, and G. Rautmann. 1994. Evidence that HIV-1 Rev directly promotes the nuclear export of unspliced RNA. Embo J 13:4105-12.
Flinta, C., B. Persson, H. Jornvall, and G. von Heijne. 1986. Sequence determinants of cytosolic N-terminal protein processing. Eur J Biochem 154:193-6.
Frankel, A. D., and J. A. Young. 1998. HIV-1: fifteen proteins and an RNA. Annu Rev Biochem 67:1-25.
Frottin, F., A. Martinez, P. Peynot, S. Mitra, R. C. Holz, C. Giglione, and T. Meinnel. 2006. The proteomics of N-terminal methionine cleavage. Mol Cell Proteomics 5:2336-49.
Furtado, M. R., R. Balachandran, P. Gupta, and S. M. Wolinsky. 1991. Analysis of alternatively spliced human immunodeficiency virus type-1 mRNA species, one of which encodes a novel tat-env fusion protein. Virology 185:258-70.
Futcher, B., G. I. Latter, P. Monardo, C. S. McLaughlin, and J. I. Garrels. 1999. A sampling of the yeast proteome. Mol Cell Biol 19:7357-68.
Gaynor, R. 1991. Cellular factors involved in regulating HIV gene expression. Pages 107-134 *in* Genetic Structure and Regulation of HIV (W. A. Haseltine, and F. Wong-Staal, eds.). Raven, New York.
Gaynor, R. 1992. Cellular transcription factors involved in the regulation of HIV-1 gene expression. Aids 6:347-63.

Geballe, A. P., and M. K. Gray. 1992. Variable inhibition of cell-free translation by HIV-1 transcript leader sequences. Nucleic Acids Res 20:4291-7.
Gelderblom, H. R. 1991. Assembly and morphology of HIV: potential effect of structure on viral function. Aids 5:617-37.
Gelderblom, H. R., E. H. Hausmann, M. Ozel, G. Pauli, and M. A. Koch. 1987. Fine structure of human immunodeficiency virus (HIV) and immunolocalization of structural proteins. Virology 156:171-6.
Giglione, C., A. Boularot, and T. Meinnel. 2004. Protein N-terminal methionine excision. Cell Mol Life Sci 61:1455-74.
Giglione, C., O. Vallon, and T. Meinnel. 2003. Control of protein life-span by N-terminal methionine excision. Embo J 22:13-23.
Gilboa, E., S. W. Mitra, S. Goff, and D. Baltimore. 1979. A detailed model of reverse transcription and tests of crucial aspects. Cell 18:93-100.
Goetz, R. M., and A. Fuglsang. 2005. Correlation of codon bias measures with mRNA levels: analysis of transcriptome data from Escherichia coli. Biochem Biophys Res Commun 327:4-7.
Guatelli, J. C., T. R. Gingeras, and D. D. Richman. 1990. Alternative splice acceptor utilization during human immunodeficiency virus type 1 infection of cultured cells. J Virol 64:4093-8.
Hadzopoulou-Cladaras, M., B. K. Felber, C. Cladaras, A. Athanassopoulos, A. Tse, and G. N. Pavlakis. 1989. The rev (trs/art) protein of human immunodeficiency virus type 1 affects viral mRNA and protein expression via a cis-acting sequence in the env region. J Virol 63:1265-74.
Hahn, B. H., G. M. Shaw, K. M. De Cock, and P. M. Sharp. 2000. AIDS as a zoonosis: scientific and public health implications. Science 287:607-14.
Hamilton, R., C. K. Watanabe, and H. A. de Boer. 1987. Compilation and comparison of the sequence context around the AUG startcodons in Saccharomyces cerevisiae mRNAs. Nucleic Acids Res 15:3581-93.
Harkins, S., C. T. Cornell, and J. L. Whitton. 2005. Analysis of translational initiation in coxsackievirus B3 suggests an alternative explanation for the high frequency of R+4 in the eukaryotic consensus motif. J Virol 79:987-96.
Harrich, D., J. Garcia, F. Wu, R. Mitsuyasu, J. Gonazalez, and R. Gaynor. 1989. Role of SP1-binding domains in in vivo transcriptional regulation of the human immunodeficiency virus type 1 long terminal repeat. J Virol 63:2585-91.
Hauber, J., A. Perkins, E. P. Heimer, and B. R. Cullen. 1987. Trans-activation of human immunodeficiency virus gene expression is mediated by nuclear events. Proc Natl Acad Sci U S A 84:6364-8.
Hemelaar, J., E. Gouws, P. D. Ghys, and S. Osmanov. 2006. Global and regional distribution of HIV-1 genetic subtypes and recombinants in 2004. Aids 20:W13-23.
Ho, D. D., A. U. Neumann, A. S. Perelson, W. Chen, J. M. Leonard, and M. Markowitz. 1995. Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection. Nature 373:123-6.
Ho, D. D., T. R. Rota, and M. S. Hirsch. 1986. Infection of monocyte/macrophages by human T lymphotropic virus type III. J Clin Invest 77:1712-5.
Holmes, E. C., and A. Rambaut. 2004. Viral evolution and the emergence of SARS coronavirus. Philos Trans R Soc Lond B Biol Sci 359:1059-65.

Horuk, R. 1994. Molecular properties of the chemokine receptor family. Trends Pharmacol Sci 15:159-65.
Hu, W. S., and H. M. Temin. 1990. Retroviral recombination and reverse transcription. Science 250:1227-33.
Huang, M., J. M. Orenstein, M. A. Martin, and E. O. Freed. 1995. p6Gag is required for particle production from full-length human immunodeficiency virus type 1 molecular clones expressing protease. J Virol 69:6810-8.
Hubner, A., M. Kruhoffer, F. Grosse, and G. Krauss. 1992. Fidelity of human immunodeficiency virus type I reverse transcriptase in copying natural RNA. J Mol Biol 223:595-600.
Ikemura, T. 1992. Correlation between codon usage and tRNA content in microorganisms. Pages 87-111 *in* Transfer RNA in protein synthesis. (D. L. Hatfield, B. Lee, and J. Pirtle, eds.). CRC Press, Boca Raton, Fla.
Jacks, T., M. D. Power, F. R. Masiarz, P. A. Luciw, P. J. Barr, and H. E. Varmus. 1988. Characterization of ribosomal frameshifting in HIV-1 gag-pol expression. Nature 331:280-3.
Jang, S. K., H. G. Krausslich, M. J. Nicklin, G. M. Duke, A. C. Palmenberg, and E. Wimmer. 1988. A segment of the 5' nontranslated region of encephalomyocarditis virus RNA directs internal entry of ribosomes during in vitro translation. J Virol 62:2636-43.
Jansen, R., H. J. Bussemaker, and M. Gerstein. 2003. Revisiting the codon adaptation index from a whole-genome perspective: analyzing the relationship between gene expression and codon occurrence in yeast using a variety of models. Nucleic Acids Res 31:2242-51.
Jetzt, A. E., H. Yu, and G. J. Klarmann. 2000. High rate of recombination throughout the human immunodeficiency virus type 1 genome. Journal of Virology:1234-1240.
Jones, K. A., and B. M. Peterlin. 1994. Control of RNA initiation and elongation at the HIV-1 promoter. Annu Rev Biochem 63:717-43.
Joshi, C. P., H. Zhou, X. Huang, and V. L. Chiang. 1997. Context sequences of translation initiation codon in plants. Plant Mol Biol 35:993-1001.
Kanaya, S., Y. Yamada, Y. Kudo, and T. Ikemura. 1999. Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of Bacillus subtilis tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. Gene 238:143-155.
Katz, R. A., and A. M. Skalka. 1990. Generation of diversity in retroviruses. Annu Rev Genet 24:409-45.
Khalouei, S., X. Yao, J. Mennigen, M. Carullo, P. Ma, Z. Song, H. Xiong, and X. Xia. 2007. Bioinformatic approach to identify penultimate amino acids efficient for N-terminal methionine excision. Proceedings of the First International Conference on Bioinformatics and Biomedical Engineering (ICBBE); 6-8 July 2007, Wuhan:386-389.
Kjems, J., and P. A. Sharp. 1993. The basic domain of Rev from human immunodeficiency virus type 1 specifically blocks the entry of U4/U6.U5 small nuclear ribonucleoprotein in spliceosome assembly. J Virol 67:4769-76.
Klasens, B. I., M. Thiesen, A. Virtanen, and B. Berkhout. 1999. The ability of the HIV-1 AAUAAA signal to bind polyadenylation factors is controlled by local RNA structure. Nucleic Acids Res 27:446-54.

Kohl, N. E., E. A. Emini, W. A. Schleif, L. J. Davis, J. C. Heimbach, R. A. Dixon, E. M. Scolnick, and I. S. Sigal. 1988. Active human immunodeficiency virus protease is required for viral infectivity. Proc Natl Acad Sci U S A 85:4686-90.

Kozak, M. 1980. Evaluation of the "scanning model" for initiation of protein synthesis in eucaryotes. Cell 22:7-8.

Kozak, M. 1986a. Bifunctional messenger RNAs in eukaryotes. Cell 47:481-3.

Kozak, M. 1986b. Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. Cell 44:283-92.

Kozak, M. 1987. An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. Nucleic Acids Res 15:8125-48.

Kozak, M. 1989. The scanning model for translation: an update. J Cell Biol 108:229-41.

Kozak, M. 1992. A consideration of alternative models for the initiation of translation in eukaryotes. Crit Rev Biochem Mol Biol 27:385-402.

Kozak, M. 1997. Recognition of AUG and alternative initiator codons is augmented by G in position +4 but is not generally affected by the nucleotides in positions +5 and +6. Embo J 16:2482-92.

Kozak, M. 1999. Initiation of translation in prokaryotes and eukaryotes. Gene 234:187-208.

Kozak, M. 2001. New ways of initiating translation in eukaryotes? Mol Cell Biol 21:1899-907.

Krummheuer, J., A. T. Johnson, I. Hauber, S. Kammler, J. L. Anderson, J. Hauber, D. F. Purcell, and H. Schaal. 2007. A minimal uORF within the HIV-1 vpu leader allows efficient translation initiation at the downstream env AUG. Virology.

Lai, M. M. 1992. RNA recombination in animal and plant viruses. Microbiol Rev 56:61-79.

Laughrea, M., and L. Jette. 1994. A 19-nucleotide sequence upstream of the 5' major splice donor is part of the dimerization domain of human immunodeficiency virus 1 genomic RNA. Biochemistry 33:13464-74.

Layne, S. P., M. J. Merges, M. Dembo, J. L. Spouge, S. R. Conley, J. P. Moore, J. L. Raina, H. Renz, H. R. Gelderblom, and P. L. Nara. 1992. Factors underlying spontaneous inactivation and susceptibility to neutralization of human immunodeficiency virus. Virology 189:695-714.

Leitner, T., B. Foley, B. Hahn, P. Marx, F. McCutchan, J. Mellors, S. Wolinsky, and B. Korber. 2005. HIV Sequence Compendium. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory.

Lever, A., H. Gottlinger, W. Haseltine, and J. Sodroski. 1989. Identification of a sequence required for efficient packaging of human immunodeficiency virus type 1 RNA into virions. J Virol 63:4085-7.

Levitt, M. 1992. Accurate modeling of protein conformation by automatic segment matching. J Mol Biol 226:507-33.

Levy, J. A. 1993. Pathogenesis of human immunodeficiency virus infection. Microbiol Rev 57:183-289.

Li, X., and Y. H. Chang. 1995. Amino-terminal protein processing in Saccharomyces cerevisiae is an essential function that requires two distinct methionine aminopeptidases. Proc Natl Acad Sci U S A 92:12357-61.

Liljenstrom, H., and G. von Heijne. 1987. Translation rate modification by preferential codon usage: intragenic position effects. J Theor Biol 124:43-55.

Lithwick, G., and H. Margalit. 2005. Relative predicted protein levels of functionally associated proteins are conserved across organisms. Nucleic Acids Res 33:1051-7.

Liu, H., X. Wu, M. Newman, G. M. Shaw, B. H. Hahn, and J. C. Kappes. 1995. The Vif protein of human and simian immunodeficiency viruses is packaged into virions and associates with viral core structures. J Virol 69:7630-8.

Loeb, D. D., R. Swanstrom, L. Everitt, M. Manchester, S. E. Stamper, and C. A. Hutchison, 3rd. 1989. Complete mutagenesis of the HIV-1 protease. Nature 340:397-400.

Lopez-Lastra, M., C. Gabus, and J. L. Darlix. 1997. Characterization of an internal ribosomal entry segment within the 5' leader of avian reticuloendotheliosis virus type A RNA and development of novel MLV-REV-based retroviral vectors. Hum Gene Ther 8:1855-65.

Lu, Y. L., P. Spearman, and L. Ratner. 1993. Human immunodeficiency virus type 1 viral protein R localization in infected cells and virions. J Virol 67:6542-50.

Luukkonen, B. G., W. Tan, and S. Schwartz. 1995. Efficiency of reinitiation of translation on human immunodeficiency virus type 1 mRNAs is determined by the length of the upstream open reading frame and by intercistronic distance. J Virol 69:4086-94.

MacDonell, K. B., J. S. Chmiel, L. Poggensee, S. Wu, and J. P. Phair. 1990. Predicting progression to AIDS: combined usefulness of CD4 lymphocyte counts and p24 antigenemia. Am J Med 89:706-12.

Maddon, P. J., A. G. Dalgleish, J. S. McDougal, P. R. Clapham, R. A. Weiss, and R. Axel. 1986. The T4 gene encodes the AIDS virus receptor and is expressed in the immune system and the brain. Cell 47:333-48.

Mammano, F., E. Kondo, J. Sodroski, A. Bukovsky, and H. G. Gottlinger. 1995. Rescue of human immunodeficiency virus type 1 matrix protein mutants by envelope glycoproteins with short cytoplasmic domains. J Virol 69:3824-30.

Manchester, M., L. Everitt, D. D. Loeb, C. A. Hutchison, 3rd, and R. Swanstrom. 1994. Identification of temperature-sensitive mutants of the human immunodeficiency virus type 1 protease through saturation mutagenesis. Amino acid side chain requirements for temperature sensitivity. J Biol Chem 269:7689-95.

Margolick, J. B., D. J. Volkman, T. M. Folks, and A. S. Fauci. 1987. Amplification of HTLV-III/LAV infection by antigen-induced activation of T cells and direct suppression by virus of lymphocyte blastogenic responses. J Immunol 138:1719-23.

McCutchan, F. E. 2006. Global epidemiology of HIV. J Med Virol 78 Suppl 1:S7-S12.

McDougal, J. S., M. S. Kennedy, J. M. Sligh, S. P. Cort, A. Mawle, and J. K. Nicholson. 1986. Binding of HTLV-III/LAV to T4+ T cells by a complex of the 110K viral protein and the T4 molecule. Science 231:382-5.

Meinnel, T., Y. Mechulam, and S. Blanquet. 1993. Methionine as translation start signal: a review of the enzymes of the pathway in *Escherichia coli*. Biochimie 75:1061-75.

Meyer, B. E., and M. H. Malim. 1994. The HIV-1 Rev trans-activator shuttles between the nucleus and the cytoplasm. Genes Dev 8:1538-47.

Miele, G., A. Mouland, G. P. Harrison, E. Cohen, and A. M. Lever. 1996. The human immunodeficiency virus type 1 5' packaging signal structure affects translation but does not function as an internal ribosome entry site structure. J Virol 70:944-51.

Moerschell, R. P., Y. Hosokawa, S. Tsunasawa, and F. Sherman. 1990. The specificities of yeast methionine aminopeptidase and acetylation of amino-terminal methionine in vivo. Processing of altered iso-1-cytochromes c created by oligonucleotide transformation. J Biol Chem 265:19638-43.

Morrow, C. D., J. Park, and J. K. Wakefield. 1994. Viral gene products and replication of the human immunodeficiency type 1 virus. Am J Physiol 266:C1135-56.

Moss, A. R., and P. Bacchetti. 1989. Natural history of HIV infection. Aids 3:55-61.
Muesing, M. A., D. H. Smith, C. D. Cabradilla, C. V. Benton, L. A. Lasky, and D. J. Capon. 1985. Nucleic acid structure and expression of the human AIDS/lymphadenopathy retrovirus. Nature 313:450-8.
Nag, B., H. G. Wada, D. Passmore, B. R. Clark, S. D. Sharma, and H. M. McConnell. 1993. Purified beta-chain of MHC class II binds to CD4 molecules on transfected HeLa cells. J Immunol 150:1358-64.
O'Neil, P. K., G. Sun, H. Yu, Y. Ron, J. P. Dougherty, and B. D. Preston. 2002. Mutational analysis of HIV-1 long terminal repeats to explore the relative contribution of reverse transcriptase and RNA polymerase II to viral mutagenesis. J Biol Chem 277:38053-61.
Ohlmann, T., M. Lopez-Lastra, and J. L. Darlix. 2000. An internal ribosome entry segment promotes translation of the simian immunodeficiency virus genomic RNA. J Biol Chem 275:11899-906.
Panganiban, A. T., and D. Fiore. 1988. Ordered interstrand and intrastrand DNA transfer during reverse transcription. Science 241:1064-9.
Pantaleo, G., C. Graziosi, J. F. Demarest, L. Butini, M. Montroni, C. H. Fox, J. M. Orenstein, D. P. Kotler, and A. S. Fauci. 1993. HIV infection is active and progressive in lymphoid tissue during the clinically latent stage of disease. Nature 362:355-8.
Park, J., and C. D. Morrow. 1993. Mutations in the protease gene of human immunodeficiency virus type 1 affect release and stability of virus particles. Virology 194:843-50.
Parkin, N. T., E. A. Cohen, A. Darveau, C. Rosen, W. Haseltine, and N. Sonenberg. 1988. Mutational analysis of the 5' non-coding region of human immunodeficiency virus type 1: effects of secondary structure on translation. Embo J 7:2831-7.
Peeters, M. 2001. Recombinant HIV sequences: their role in the global epidemic. Pages 54-72 *in HIV sequence compendium 2000* (C. Kuiken, B. Foley, B. Hahn, F. E. McCutchan, J. W. Mellors, J. I. Mullins, J. Sodroski, S. Wolinsky, and B. Korber, eds.). Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos.
Peliska, J. A., and S. J. Benkovic. 1992. Mechanism of DNA strand transfer reactions catalyzed by HIV-1 reverse transcriptase. Science 258:1112-8.
Pelletier, J., and N. Sonenberg. 1988. Internal initiation of translation of eukaryotic mRNA directed by a sequence derived from poliovirus RNA. Nature 334:320-5.
Percudani, R., A. Pavesi, and S. Ottonello. 1997. Transfer RNA gene redundancy and translational selection in Saccharomyces cerevisiae. J Mol Biol 268:322-30.
Peterlin, B. M. 1991. Transcriptional regulation of HIV. Pages 237-250 *in* Genetic Structure and Regulation of HIV (W. A. Haseltine, and F. Wong-Staal, eds.). Raven, New York.
Piantadosi, A., B. Chohan, V. Chohan, R. S. McClelland, and J. Overbaugh. 2007. Chronic HIV-1 Infection Frequently Fails to Protect against Superinfection. PLoS Pathog 3:e177.
Pierson, T., T. L. Hoffman, J. Blankson, D. Finzi, K. Chadwick, J. B. Margolick, C. Buck, J. D. Siliciano, R. W. Doms, and R. F. Siliciano. 2000. Characterization of chemokine receptor utilization of viruses in the latent reservoir for human immunodeficiency virus type 1. J Virol 74:7824-33.

Pierson, T. C., Y. Zhou, T. L. Kieffer, C. T. Ruff, C. Buck, and R. F. Siliciano. 2002. Molecular characterization of preintegration latency in human immunodeficiency virus type 1 infection. J Virol 76:8518-31.

Poon, D. T., E. N. Chertova, and D. E. Ott. 2002. Human immunodeficiency virus type 1 preferentially encapsidates genomic RNAs that encode Pr55(Gag): functional linkage between translation and RNA packaging. Virology 293:368-78.

Popescu, C. E., T. Borza, J. P. Bielawski, and R. W. Lee. 2006. Evolutionary rates and expression level in Chlamydomonas. Genetics 172:1567-76.

Preston, B. D., and J. P. Dougherty. 1996. Mechanisms of retroviral mutation. Trends Microbiol 4:16-21.

Provitera, P., R. El-Maghrabi, and S. Scarlata. 2006. The effect of HIV-1 Gag myristoylation on membrane binding. Biophys Chem 119:23-32.

Pryciak, P. M., and H. E. Varmus. 1992. Nucleosomes, DNA-binding proteins, and DNA sequence modulate retroviral integration target site selection. Cell 69:769-80.

Purcell, D. F., and M. A. Martin. 1993. Alternative splicing of human immunodeficiency virus type 1 mRNA modulates viral protein expression, replication, and infectivity. J Virol 67:6365-78.

Rein, A., L. E. Henderson, and J. G. Levin. 1998. Nucleic-acid-chaperone activity of retroviral nucleocapsid proteins: significance for viral replication. Trends Biochem Sci 23:297-301.

Rice, A. P., and M. B. Mathews. 1988. Transcriptional but not translational regulation of HIV-1 by the tat gene product. Nature 332:551-3.

Rice, P., I. Longden, and A. Bleasby. 2000. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 16:276-7.

Robert-Guroff, M., M. Popovic, S. Gartner, P. Markham, R. C. Gallo, and M. S. Reitz. 1990. Structure and expression of tat-, rev-, and nef-specific transcripts of human immunodeficiency virus type 1 in infected lymphocytes and macrophages. J Virol 64:3391-8.

Robert-Seilaniantz, A., L. Shan, J. M. Zhou, and X. Tang. 2006. The Pseudomonas syringae pv. tomato DC3000 type III effector HopF2 has a putative myristoylation site required for its avirulence and virulence functions. Mol Plant Microbe Interact 19:130-8.

Roberts, J. D., K. Bebenek, and T. A. Kunkel. 1988. The accuracy of reverse transcriptase from HIV-1. Science 242:1171-3.

Robertson, D. L., J. P. Anderson, J. A. Bradac, J. K. Carr, B. Foley, R. K. Funkhouser, F. Gao, B. H. Hahn, M. L. Kalish, C. Kuiken, G. H. Learn, T. Leitner, F. McCutchan, S. Osmanov, M. Peeters, D. Pieniazek, M. Salminen, P. M. Sharp, S. Wolinsky, and B. Korber. 2000. HIV-1 nomenclature proposal. Science 288:55-6.

Robey, E., and R. Axel. 1990. CD4: collaborator in immune recognition and HIV infection. Cell 60:697-700.

Rowe, D. C., A. F. McGettrick, E. Latz, B. G. Monks, N. J. Gay, M. Yamamoto, S. Akira, L. A. O'Neill, K. A. Fitzgerald, and D. T. Golenbock. 2006. The myristoylation of TRIF-related adaptor molecule is essential for Toll-like receptor 4 signal transduction. Proc Natl Acad Sci U S A 103:6299-304.

Ruiz, L. M., G. Armengol, E. Habeych, and S. Orduz. 2006. A theoretical analysis of codon adaptation index of the Boophilus microplus bm86 gene directed to the optimization of a DNA vaccine. J Theor Biol 239:445-9.

Saag, M. S. 1994. Natural History of HIV-1 Disease. Pages 21-43 *in Textbook of AIDS Medicine* (S. Broder, T. C. Merigan, and D. Bolognesi, eds.). Williams & Wilkins, Baltimore.

Sakai, H., R. Shibata, J. Sakuragi, S. Sakuragi, M. Kawamura, and A. Adachi. 1993. Cell-dependent requirement of human immunodeficiency virus type 1 Vif protein for maturation of virus particles. J Virol 67:1663-6.

Sakurai, N., and T. Utsumi. 2006. Posttranslational N-myristoylation is required for the anti-apoptotic activity of human tGelsolin, the C-terminal caspase cleavage product of human gelsolin. J Biol Chem 281:14288-95.

Schroder, A. R., P. Shinn, H. Chen, C. Berry, J. R. Ecker, and F. Bushman. 2002. HIV-1 integration in the human genome favors active genes and local hotspots. Cell 110:521-9.

Schwartz, S., B. K. Felber, D. M. Benko, E. M. Fenyo, and G. N. Pavlakis. 1990. Cloning and functional analysis of multiply spliced mRNA species of human immunodeficiency virus type 1. J Virol 64:2519-29.

Schwartz, S., B. K. Felber, and G. N. Pavlakis. 1992. Mechanism of translation of monocistronic and multicistronic human immunodeficiency virus type 1 mRNAs. Mol Cell Biol 12:207-19.

Selby, M. J., E. S. Bain, P. A. Luciw, and B. M. Peterlin. 1989. Structure, sequence, and position of the stem-loop in tar determine transcriptional elongation by tat through the HIV-1 long terminal repeat. Genes Dev 3:547-58.

Semon, M., J. R. Lobry, and L. Duret. 2006. No evidence for tissue-specific adaptation of synonymous codon usage in humans. Mol Biol Evol 23:523-9.

SenGupta, D. N., B. Berkhout, A. Gatignol, A. M. Zhou, and R. H. Silverman. 1990. Direct evidence for translational regulation by leader RNA and Tat protein of human immunodeficiency virus type 1. Proc Natl Acad Sci U S A 87:7492-6.

Serero, A., C. Giglione, A. Sardini, J. Martinez-Sanz, and T. Meinnel. 2003. An unusual peptide deformylase features in the human mitochondrial N-terminal methionine excision pathway. J Biol Chem 278:52953-63.

Sharp, P. M., and W. H. Li. 1987. The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res 15:1281-95.

Simon, F., P. Mauclere, P. Roques, I. Loussert-Ajaka, M. C. Muller-Trutwin, S. Saragosti, M. C. Georges-Courbot, F. Barre-Sinoussi, and F. Brun-Vezinet. 1998. Identification of a new human immunodeficiency virus type 1 distinct from group M and group O. Nat Med 4:1032-7.

Smith, D. M., J. K. Wong, G. K. Hightower, C. C. Ignacio, K. K. Koelsch, E. S. Daar, D. D. Richman, and S. J. Little. 2004. Incidence of HIV superinfection following primary infection. Jama 292:1177-8.

Smith, J., A. Azad, and N. Deacon. 1992. Identification of two novel human immunodeficiency virus type 1 splice acceptor sites in infected T cell lines. J Gen Virol 73 (Pt 7):1825-8.

Smith, R. A., L. A. Loeb, and B. D. Preston. 2005. Lethal mutagenesis of HIV. Virus Res 107:215-28.

Stein, B. S., S. D. Gowda, J. D. Lifson, R. C. Penhallow, K. G. Bensch, and E. G. Engleman. 1987. pH-independent HIV entry into CD4-positive T cells via virus envelope fusion to the plasma membrane. Cell 49:659-68.

Stein, D. S., J. A. Korvick, and S. H. Vermund. 1992. CD4+ lymphocyte cell enumeration for prediction of clinical course of human immunodeficiency virus disease: a review. J Infect Dis 165:352-63.

Steinman, R. M., A. Granelli-Piperno, M. Pope, C. Trumpfheller, R. Ignatius, G. Arrode, P. Racz, and K. Tenner-Racz. 2003. The interaction of immunodeficiency viruses with dendritic cells. Curr Top Microbiol Immunol 276:1-30.

Stormo, G. D., T. D. Schneider, L. Gold, and A. Ehrenfeucht. 1982. Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli. Nucleic Acids Res 10:2997-3011.

Svitkin, Y. V., A. Pause, and N. Sonenberg. 1994. La autoantigen alleviates translational repression by the 5' leader sequence of the human immunodeficiency virus type 1 mRNA. J Virol 68:7001-7.

Temin, H. M. 1991. Sex and recombination in retroviruses. Trends Genet 7:71-4.

Temin, H. M., and S. Mizutani. 1992. RNA-dependent DNA polymerase in virions of Rous sarcoma virus. 1970. Biotechnology 24:51-6.

Turner, B. G., and M. F. Summers. 1999. Structural biology of HIV. J Mol Biol 285:1-32.

UNAIDS. 2006. Joint United Nations Programme on HIV/AIDS (UNAIDS) and World Health Organization (WHO). AIDS epidemic update: December 2006.

van der Kuyl, A. C., K. Kozaczynska, R. van den Burg, F. Zorgdrager, N. Back, S. Jurriaans, B. Berkhout, P. Reiss, and M. Cornelissen. 2005. Triple HIV-1 infection. N Engl J Med 352:2557-9.

Vignuzzi, M., J. K. Stone, J. J. Arnold, C. E. Cameron, and R. Andino. 2006. Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. Nature 439:344-8.

Vilas, G. L., M. M. Corvi, G. J. Plummer, A. M. Seime, G. R. Lambkin, and L. G. Berthiaume. 2006. Posttranslational myristoylation of caspase-activated p21-activated protein kinase 2 (PAK2) potentiates late apoptotic events. Proc Natl Acad Sci U S A 103:6542-7.

Waterman, M. L., P. L. Sheridan, L. H. Milocco, C. T. Sheline, and K. A. Jones. 1991. Nuclear proteins implicated in HIV-1 transcriptional control. Pages 391-403 *in* Genetic Structure and Regulation of HIV (W. A. Haseltine, and F. Wong-Staal, eds.). Raven, New York.

Welker, R., H. Kottler, H. R. Kalbitzer, and H. G. Krausslich. 1996. Human immunodeficiency virus type 1 Nef protein is incorporated into virus particles and specifically cleaved by the viral proteinase. Virology 219:228-36.

WHO. 2004. *The World Health Report 2004: changing history.*

Wu, C. H., R. Apweiler, A. Bairoch, D. A. Natale, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, R. Mazumder, C. O'Donovan, N. Redaschi, and B. Suzek. 2006. The Universal Protein Resource (UniProt): an expanding universe of protein information. Nucleic Acids Res 34:D187-91.

Wu, G., D. E. Culley, and W. Zhang. 2005. Predicted highly expressed genes in the genomes of Streptomyces coelicolor and Streptomyces avermitilis and the implications for their metabolism. Microbiology 151:2175-87.

Xia, X. 1998. How optimized is the translational machinery in Escherichia coli, Salmonella typhimurium and Saccharomyces cerevisiae? Genetics 149:37-44.

Xia, X. 2001. Data analysis in molecular biology and evolution. Kluwer Academic Publishers, Boston.
Xia, X. 2007a. The +4G site in Kozak consensus is not related to the efficiency of translation initiation. PLoS ONE 2:e188.
Xia, X. 2007b. Bioinformatics and the cell: modern computational approaches in genomics, proteomics and transcriptomics. Springer.
Xia, X. 2007c. An Improved Implementation of Codon Adaptation Index. Evolutionary Bioinformatics 3:53-58.
Xia, X., and Z. Xie. 2001. DAMBE: software package for data analysis in molecular biology and evolution. J Hered 92:371-3.
Yamauchi, K. 1991. The sequence flanking translational initiation site in protozoa. Nucleic Acids Res 19:2715-20.
Yilmaz, A., C. Bolinger, and K. Boris-Lawrie. 2006. Retrovirus translation initiation: Issues and hypotheses derived from study of HIV-1. Curr HIV Res 4:131-9.
Yu, X., X. Yuan, Z. Matsuda, T. H. Lee, and M. Essex. 1992. The matrix protein of human immunodeficiency virus type 1 is required for incorporation of viral envelope protein into mature virions. J Virol 66:4966-71.
Zombek, D., N. R. Landau, and D. Trono. 1994. Vif is incorporated into HIV-1 particles: mapping of its packaging requirements. Pages p. 233 *in* Retroviruses '94 Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.

Copyright Permissions

Figures 1.1 and 1.2

From: "publicationpermissions" <publicationpermissions@unaids.org>
Subject: RE: Copyright permission (Global view of HIV infection, 2006)
Date: Tue, 28 August, 2007 2:53 am
To: "Sam Khalouei"

Dear Sam Khalouei

Thank you for your enquiry. We are pleased to let you know that you can use the image you specify in your thesis, please acknowledge 'Reproduced by kind permission of UNAIDS'. Please note that this permission is granted for one time use only - should you wish in future to publish your thesis for example it will be necessary to ask us again for permission.

Please accept our best wishes for success in your studies.

Alistair Craik

From: "publicationpermissions" <publicationpermissions@unaids.org>
Subject: RE: Copyright permission (PubMed ID: 17053344)
Date: Tue, 28 August, 2007 3:09 am
To: "Sam Khalouei"

Dear Sam Khalouei

The Journal AIDS is the publisher - the paper was written by colleagues here in UNAIDS and WHO. I am sure that my colleagues would be glad for you to quote their work in your thesis - it is not usually necessary to seek individual permissions for use of selected data from publications in the preparation of academic research which is not to be published. But, we appreciate your courtesy, and, as noted in another message if you wish to re-use the figure in any other context then please apply for permission from the publishers of the Journal.

With best wishes

Alistair Craik

Figure 1.4

From: "Chad Ishmael"
Subject: Re: Copyright permission
Date: Sat, 9 June, 2007 12:52 pm
To: "Sam Khalouei"

Dear Sam,

I apologise for the late reply, I have been out of the country and away from internet. You have my permission to include the image in your thesis, and I wish you luck with the project.

Sincerely,

Chad Ishmael