Analysis of protein-RNA and protein-peptide interactions in Equine Infectious Anemia Virus (EIAV) infection

by

Jae-Hyung Lee

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Bioinformatics and Computational Biology

Program of Study Committee:
Drena Dobbs, Co-Major Professor
Kai-Ming Ho, Co-Major Professor
Amy Andreotti
Robert Jernigan
Vasant Honavar

Iowa State University

Ames, Iowa

2007

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Macromolecular interactions are essential for virtually all cellular functions including signal transduction processes, metabolic processes, regulation of gene expression and immune responses. This dissertation focuses on the characterization of two important macromolecular interactions involved in the relationship between Equine Infectious Anemia Virus (EIAV) and its host cell in horse: i) the interaction between the EIAV Rev protein and its binding site, the Rev-responsive element (RRE) and ii) interactions between equine MHC class I molecules and epitope peptides derived from EIAV proteins.

EIAV, one of the most divergent members of the lentivirus family, has a single-stranded RNA genome and carries several regulatory and structural proteins within its viral particle. Rev is an essential EIAV regulatory encoded protein that interacts with the viral RRE, a specific binding site in the viral mRNA. Using a combination of experimental and computational methods, the interactions between EIAV Rev and RRE were characterized in detail. EIAV Rev was shown to have a bipartite RNA binding domain containing two arginine rich motifs (ARMs). The RRE secondary structure was determined and specific structural motifs that act as cis-regulatory elements for EIAV Rev-RRE interaction were identified. Interestingly, a structural motif located in the high affinity Rev binding site is well conserved in several diverse lentiviral genomes, including HIV-1.

Macromolecular interactions involved in the immune response of the horse to EIAV infection were investigated by analyzing complexes between MHC class I proteins and epitope peptides derived from EIAV Rev, Env and Gag proteins. Computational modeling results provided a mechanistic explanation for the experimental finding that a single amino acid change in the peptide binding domain of the equine MHC class I molecule differentially affects the recognition of specific epitopes by EIAV-specific CTL. Together, the findings in this dissertation provide novel insights into the strategy used by EIAV to replicate itself, and provide new details about how the host cell responds to and defends against EIAV upon the infection. Moreover, they have contributed to our understanding of the macromolecular recognition events that regulate these processes.

CHAPTER 1. GENERAL INTRODUCTION

This dissertation describes and characterizes the molecular details of interactions between regulatory proteins and their binding partners, including other proteins, RNA, and small peptides. In this study, a lentivirus, Equine Infectious Anemia Virus (EIAV) is used as a model system to analyze the interactions systematically. We focused on the interactions between the viral regulatory protein, Rev, and its binding target in viral mRNA, the Rev-Responsive Element (RRE) to characterize EIAV protein-RNA interactions essential for replication of the virus. We also observed and characterized protein-peptide interactions important in the immune response of the equine host by modeling of the equine MHC class I protein and its interactions with peptide epitopes derived from EIAV proteins, using docking approaches.

INTRODUCTION

Proteins are the most critical molecules in macromolecular interactions, playing essential roles in virtually all cellular functions. The regulation of DNA replication, transcription and translation, signal transduction pathways, metabolic processes, and immune responses are performed and controlled largely by protein-protein, protein-nucleic acid, and protein-small ligand interactions. Currently, macromolecular interactions are a central theme of structural and functional genomics, and genome-wide protein interaction networks are being generated and analyzed in model organisms by taking advantage of high-throughput experimental approaches (30). For example, in *Drosophila melanogaster*, the networks of interactions between proteins that control cell function are beginning to be revealed using two-hybrid-based protein-interaction maps (23). However, detailed mechanisms of macromolecular interactions, especially in protein-RNA and protein-small ligand complexes are still poorly understood, in part because of insufficient data.

Viruses are excellent model systems for studying macromolecular interactions. A virus contains only a small DNA or RNA genome, packaged together with a few enzymatic and/or structural proteins within a viral particle. To replicate themselves in the host, viruses use all kinds of macromolecular interactions, such as: protein-DNA (for replication of viral genomic DNA), protein-RNA (for reverse transcription and transport of mRNAs), and protein-small ligand (for escape from host immune system). Even though viruses have their own restricted resources, they are clever enough to direct these macromolecular interactions using their own proteins and various "borrowed" components of the host's machineries. Therefore, analyzing macromolecular interactions in viral model systems can give us understanding of some of the most effective ways for regulating macromolecular recognition processes. In this study, we have focused on one family of Retroviruses, the lentiviruses, to investigate two types of macromolecular interactions: i) protein-RNA interactions critical for replication of the virus in host cells, and ii) protein-peptide interactions involved in the host immune response to viral infection.

Equine Infectious Anemia Virus

Lentiviruses are a family of single-stranded, positive-strand RNA viruses that use an RNA-dependent DNA polymerase, called a reverse-transcriptase, encoded in viral genome (16). In a lentiviral infection, the virus attaches to the cell membrane of host cell and the core of virus is injected into the cytoplasm of infected cell. In the early stages of infection, the single-stranded RNA genome is reverse transcribed into DNA by the reverse transcriptase. The transcribed double-stranded DNA can be shuttled into the nucleus where it is integrated into the host genome and is maintained as a provirus. At some point, proviral genes are transcribed by the host's transcriptional machinery. Because lentiviruses have relatively small genomes, they utilize complex mechanisms to regulate viral replication and gene expression

at the transcription and post-transcriptional levels. One important mechanism is alternative splicing of viral transcripts, which generates several different spliced transcripts from the same region of the genome to express genes for later stages of viral replication (13). For example, in the case of HIV-1, there are more than 20 different spliced mRNAs for different viral gene products.

The Equine Infectious Anemia Virus (EIAV) is one of the most divergent members of the lentiviridae (17). EIAV has a complex genomic organization, and shares several common lentiviral genes essential for the viral life cycle. Unlike HIV-1, however, EIAV infection sometimes stimulates a rapid and variable disease course in the host, rather than a slow chronic and inevitably fatal disease. Figure 1.1 shows the organization of EIAV genome and spliced transcripts produced from it. The Gag gene encodes the inner structural proteins of the virion, which include the capsid, nucleocapsid and matrix proteins. The products of the Env gene are the envelope glycoproteins (surface and transmembrane glycoproteins) involved in virus-host cell fusion and entry of the virus. Also, there are genes involved in regulating viral replication, including the important and functionally conserved regulatory gene products, Tat and Rev. Tat is an RNA binding protein that interacts with a viral RNA structure called the TAR element. By binding to a stem-loop motif of the TAR element, Tat functions as a transcriptional activator, which increases the efficiency of transcription (10). The second critical regulatory protein, Rev, plays a role in regulating the expression of unspliced and incompletely spliced viral mRNAs (48). Rev protein in the cytoplasm enters the nucleus, binds to the specific viral mRNA element called the Rev-responsive Element (RRE), multimerizes, and directs the export of incompletely spliced mRNA to the cytoplasm, using CRM1 nuclear export pathway of the host cell. A detailed description of Rev and its functions will be provided below. In the late stage of viral infection, structural proteins and the progeny viral RNA genomes are targeted to the cell membrane, assembled as viral particles and finally released from the host cell.

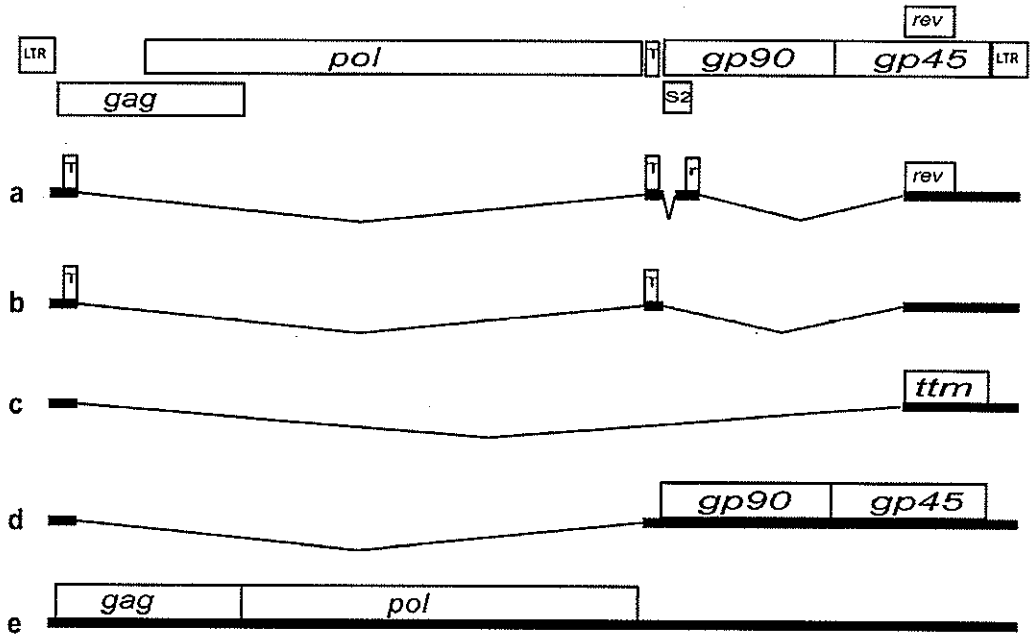


Figure 1.1. EIAV genome organization and transcript splicing patterns. Schematic of the EIAV genome and open reading frames (ORFs) within mRNA transcripts. **(a)** mRNA transcript "a" encodes both the Tat (T) and Rev (r, rev) proteins. **(b)** In the presence of Rev protein, EIAV exon 3 is skipped and the Tat (T) protein is produced from mRNA "b". **(c)** mRNA "c" encodes Ttm, a protein of unknown function. **(d, e)** Structural and enzymatic proteins are translated from mRNAs "d" and "e". Unspliced mRNA "e" corresponds to progeny RNA that is packaged to produce infectious virus.

The EIAV Rev protein and Rev-responsive element (RRE)

Rev is a regulatory protein essential for lentiviral replication. Many lentiviruses, including EIAV, BIV, FIV, visna virus, SIV, and CAEV, use a Rev-dependent RNA export pathway to achieve differential expression of incompletely spliced viral mRNAs, which encode structural and enzymatic proteins as well as the progeny viral RNA. Rev serves as the gear shift from the early expression of regulatory genes like Tat and Rev, to the late

expression of structural genes such as Gag and Env. Rev's function in the regulation of lentiviral replication have been best studied in HIV-1(28). Briefly, after entering the nucleus, the HIV-1 Rev protein binds to a specific cis-acting element, the Rev responsive element (RRE) within the viral pre-mRNA (12, 53), multimerizes (46, 54) and then facilitates nuclear export of the incompletely spliced viral mRNA via the CRM1 nuclear export pathway (18, 19) (Fig 1.2). Rev has discrete functional domains essential for the shuttling process, RNA binding, and multimerization. In the case of HIV-1 Rev, the N-terminal nuclear localization signal (NLS), which directly interacts with importin-β, is required for the nuclear transportation of Rev (25, 28, 44). The NLS includes an arginine-rich motif (ARM), which is important for binding to the HIV-1 RRE. An NMR structure of a Rev-RRE protein-RNA complex (a 23 aa Rev peptide bound to a 39 nt RRE fragment) was determined by Battiste and Williamson (2); in this complex, the Rev peptide formed an alpha-helix. Also, several studies showed that the N-terminal region of HIV-1 Rev, including the NLS/RNA binding domain (RBD), forms a helix-loop-helix structure (8, 52). Biochemical and biophysical experiments have suggested that Rev multimerization occurs on an extended RRE region as Rev concentration increases (14, 29, 37). After multimerization of Rev in Rev-RRE complexes, the nuclear export signal (NES) in the C-terminal of Rev interacts with host proteins, such as CRM1 or exportin 1. The NES motifs among different lentiviruses are conserved, in terms of their leucine-rich sequences, and it has been shown that the NES is functionally interchangeable between several lentiviral Rev proteins (19).

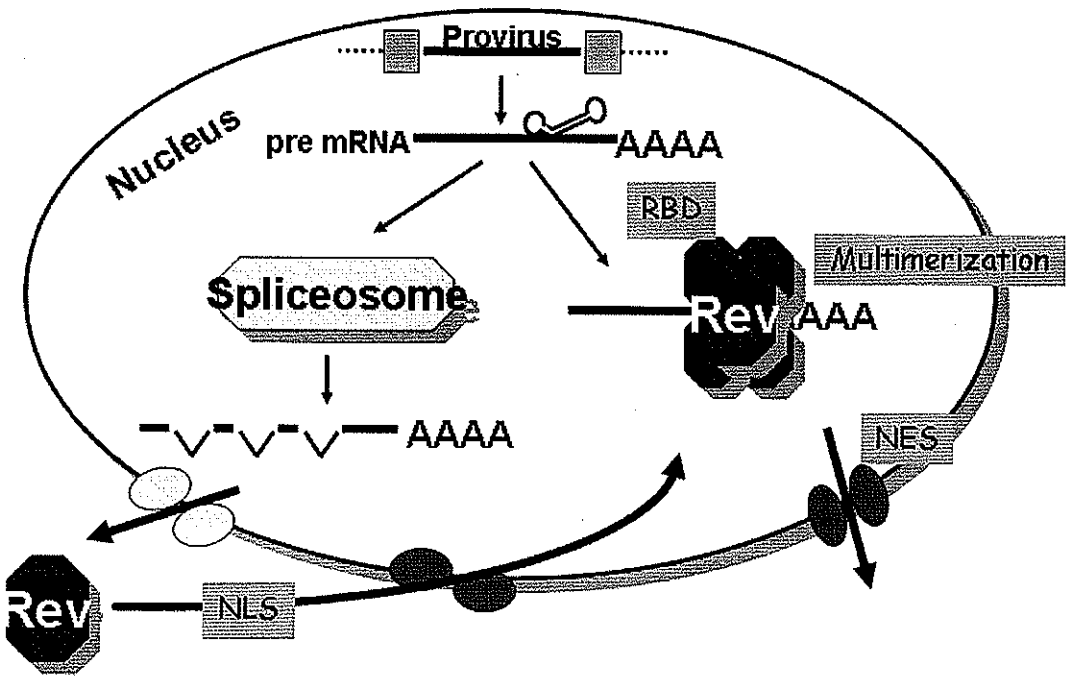


Figure 1.2. Schematic representation of Rev function. Rev enters the nucleus, facilitated by the interaction between its nuclear localization domain (NLS) and the host import machinery, including the importin-β protein. The RNA binding domain (RBD) of Rev binds to structured RRE sequences in unspliced or multiply spliced mRNAs. Rev then multimerizes on the mRNA and Rev-RRE (mRNA) complexes are exported from nucleus, through interactions between the nuclear export signal (NES) in Rev and components of the host nuclear export machinery, such as CRM1.

Rev proteins in all lentiviruses carry out similar functions, but the sequence similarities among them are quite low. In particular, Equine Infectious Anemia Virus (EIAV) Rev is genetically even more distant from HIV than other non-primate lentiviruses (17). As a result, at the primary sequence level, the organization of functional domains within the Rev proteins of HIV and EIAV Rev is very different (Fig 1.3). EIAV Rev has 165 amino acid residues compared to 116 amino acid residues in HIV-1 Rev. There are two exons in EIAV Rev and the first exon of Rev, which is the same ORF with *env* gene, has no known function

Rev in Rev export. The NES domain of EIAV Rev is located in the N-terminus, rather than C-terminus, as it is in HIV-1 Rev (26). A short arginine-rich stretch in the C-terminal region serves as a NLS in EIAV Rev, but HIV-1 Rev has a 17-residue arginine-rich motif (ARM) that functions as both an NLS and as a high affinity RNA binding domain. Although no typical ARM like that in HIV-1 Rev is found in EIAV Rev, previous studies have identified two specific motifs, RRDR (aa 76-89) and ERLE (aa 93-96), which affect functional activity of EIAV Rev in *in vivo* assays (26). Especially, the latter motif, ERLE has been also been shown, by alanine substitution mutations, to abrogate the RNA binding activity of Rev (11), and it has been assumed that the middle region of EIAV Rev was likely to be the RNA binding domain. But no extensive direct investigations of the RNA binding activity or specific RNA binding motifs in EIAV Rev had been reported when we initiated our study. In Figure 1.3A, the two putative RNA binding motifs within the potential RNA binding domain (RBD) discussed above, as well as the arginine-rich motif within the NLS of EIAV Rev (shown to be involved in RNA binding by our experiments, described in Chapter 2) as shown in red, for comparison with the well-characterized ARM within the RBD of HIV-1 Rev, which is also shown in red in Figure 1.3B.

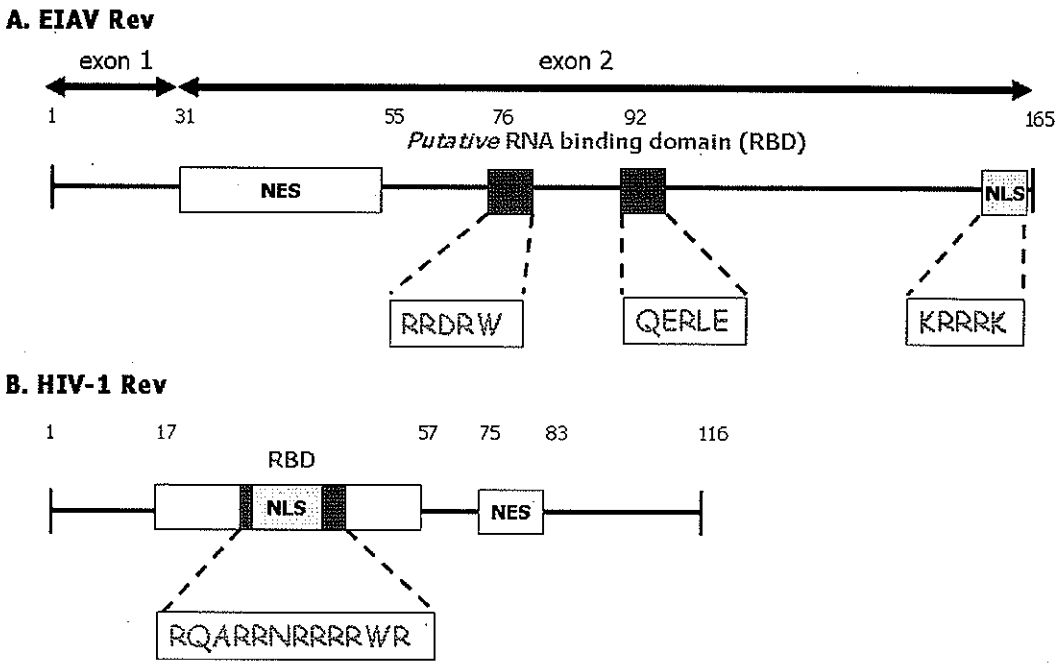


Figure 1.3. The domain organization of EIAV and HIV-1 Rev. **(A)** Domain organization of the EIAV Rev protein. NES = Nuclear export signal, located in the N-terminus of Rev exon 2. RBD = Putative RNA binding domain, including two putative RNA binding motifs (RRDRW and QERLE). NLS = Nuclear localization signal, located in the C-terminus of Rev, and including a KRRRK motif **(B)** Domain organization of the HIV-1 Rev protein. Note that the RBD and NLS are located in the N-terminal half of HIV-1 Rev. The NES is located in the C-terminal half of HIV-1 Rev. The ARM is shown in red.

The interaction between Rev protein and the RRE is a critical step in the export of incompletely spliced or unspliced lentiviral mRNA from the nucleus to the cytoplasm of infected cells. This step is key in regulation of lentiviral gene expression and genomic replication. In HIV-1, the RRE region within the viral mRNA is known to form specific RNA secondary structures required for recognition by the HIV-1 Rev protein. The HIV-1 RRE, and especially the secondary structure of the RRE, has been investigated extensively,

using biochemical and biophysical assays (15, 32, 37, 53). The HIV-1 RRE is highly structured, and contains stem-loop structures important for high affinity binding to the HIV-1 Rev protein and for interacting with cellular proteins. Mann et al. (37), showed that an extended region of the HIV-1 RRE is important for Rev multimerization. The RRE region in many lentiviruses, including HIV-1, HIV-2, SIV, VV, and CAEV, has been mapped to the junction between SU and TM genes, which corresponds to a protease cleavage site within the *env* gene (15, 33, 49, 51, 53). In case of FIV, the RRE is located in the 3' end of *env* gene (47). On the other hand, the essential RRE (ERRE) has been approximately mapped within the 3' end of the genome (38) in EIAV. Using functional assays, Belshan et al. (4), demonstrated that two separable elements of the RRE can efficiently mediate Rev-dependent pre-mRNA export in EIAV. Even though ERRE-1 (555 nt) is shorter than ERRE-2 (1698 nt), it supports a higher level of functional activity *in vivo* (~60% vs ~20%) (4). ERRE-1 includes an Exonic Splicing Enhancer (ESE) region, which contains several purine-rich tracts, two of which, Pu-A and Pu-B, have been shown to be important both for alternative splicing of EIAV mRNAs and for Rev binding (5, 11, 24, 36). ERRE-1 thus plays an important role in the complex interactions between viral pre-mRNAs, the viral Rev protein, and host cellular splicing factors, such as SF2/ASF, which also binds to ERRE-1 (3, 11, 24, 36). Prior to our work, however, detailed structural information regarding the EIAV RRE was not available, nor had interactions between the EIAV Rev protein and the EIAV RRE been investigated.

The Virus-specific CTL (Cytotoxic T Lymphocytes) Response

There are two kinds of immune responses to virus infection (16). One is the response of the innate immune system, which includes the activities of cytokines (such as interferon), complement (a complex collection of serum proteins), natural killer cells (cytolytic lymphocytes) and local sentinel cells (such as dendratic cells and macrophases). The innate

immune system has quick responses, is relatively non-specific and has no immunological "memory". The innate immune system is important not only because it is a first barrier against viral infection, but also because it plays a role in activating the adaptive immune responses.

The adaptive immune responses represent the second type of immune response to viral infection, and are complicated and sophisticated responses. There are two major aspects of the adaptive immune response: the humoral response and the cell-meditated response. Unlike the innate immune system, the adaptive immune system has highly antigen-specific recognition. In the humoral response, antibodies bind to virions in vessels and at mucosal surfaces to prevent an increase in viral infection. On the other hand, in the cell-mediated response, T cells recognize infected cells and synthesize cytokines, ultimately resulting in infected cells being killed by CTLs (cytotoxic T-lymphocytes). There are two kinds of effector T cells: T helper cells (Th cells) and cytotoxic T cells (cytotoxic T lymphocytes, or CTLs). Th cells presenting CD4 (cluster of differentiation marker-4), are able to interact with B cells and antigen-presenting cells, which display MHC class II proteins. The main roles of Th cells are the activation of B cell precursors, by direct interaction, and activation of CTL precursors, by secretion of cytokines such as interleukins and interferon-gamma. CTLs recognize the infected cell, which presents MHC class I proteins on its cell surface (16). MHC class I proteins are composed of two different subunits: a long alpha-chain (heavy chain) and a short β_2-microglobulin (β_2m, light chain). In the infected cell, the viral protein is proteolysed in the proteasome and short processed peptides bind to MHC class I proteins in the lumen of the endoplasmic reticulum (ER). Based on the known 3-dimensional (3D) structure of the MHC class I molecule, it has been shown that a peptide binds to the cleft between two helix domains (α-1 and α -2), and rests on a β strand domain chain (β-1) (7, 20). Because of polymorphic features in peptide binding domains (binding groove), the ability of MHC class I molecules to bind and display peptides on the cell surface varies from

individual to individual. The MHC I complex (MHC class I protein bound to viral peptide) on the surfaces of infected cells is recognized by the T cell receptor (TCR) of CTLs. The interaction between the MHC I complex and the TCR triggers the activation of CTLs and finally CTLs kill the infected cell. The detailed pathway of MHC class I antigen presentation is well reviewed elsewhere (27, 45). Figure 1.4 shows some of the processes involved in the virus-specific CTL response described above.

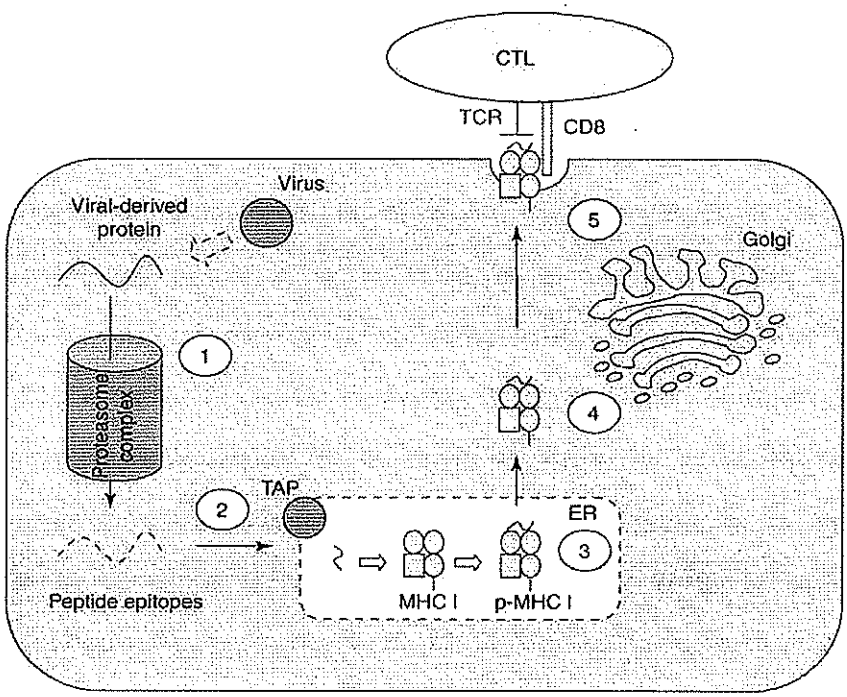


Figure 1.4. Schematic mechanism of virus-specific CTL response. **(1)** Viral-derived protein is proteolysed in the proteasome complex. **(2)** Processed peptide is transported into the endoplasmic reticulum (ER) by TAP protein. **(3)** In ER, a peptide binds to an MHC class I protein. **(4)** The MHC I-peptide complex is transported to the membrane surface through the Golgi apparatus. **(5)** The MHC I-peptide complex on the surface of membrane is recognized by a T-cell receptor on a Cytotoxic T lymphocyte (CTL). (Figure adapted from Nolan, D. et al., 2006 (45))

During lentiviral infection, the Env, Gag and Rev proteins are important targets for CTLs and the CTL response to epitope peptides derived from viral proteins is important for control of viral load and associated clinical disease. Several studies clearly show that HIV-1 Env, Gag and Rev-specific CTLs are tightly connected with the progression of AIDS in HIV-1 infected humans (1, 6, 9, 22). As they are in HIV-1 infection, Env, Gag and Rev proteins are critical targets for CTL responses in EIAV infected horses, and specific CTLs for Env and Gag peptides are observed in the progression of Equine Infectious Anemia (EIA) (41, 42). An important aspect of the virus-specific CTL response, is how the MHC class I molecule recognizes and interacts with viral epitope peptides, and subsequently activates CTL recognition processes. Although all MHC class I molecules belong to same structural supertype and have similar binding affinities for epitope peptides, polymorphism in the binding groove of MHC class I molecules affects the recognition of MHC class I-peptide complexes by CTLs (50). In human cases, many studies have investigated the correlation between specific MHC I alleles and the progress of disease (21, 31, 43). In horses, there are fewer known MHC I alleles, compared with humans. Recently, one of the classical MHC I molecules, designated 7-6, associated with the equine leukocyte Ag (ELA)-A1 haplotype has been identified (39).It has been shown that 7-6 presents both Env and Gag epitopes. Interestingly, Rev epitopes are also associated with the ELA-A1 haplotype, but the molecule presenting the Rev epitope has not been identified (39, 42). In addition, the lack of a 3D structure for the equine MHC class I molecule makes it difficult to understand the detailed mechanisms by which equine MHC class I molecules interact with epitopes from EIAV proteins, and how equine CTLs recognize equine MHC class I–peptide complexes. Therefore, we undertook a study of the structural features of equine MHC class I molecules, in complex with specific epitopes derived from EIAV proteins, using experimental and comparative computational approaches. Our long term goal in this work is not only to understand the detailed interactions that mediate recognition events among equine MHC class I molecules,

epitope peptides and CTLs in horses, but also the relationship between disease progression and the molecular mechanisms.

OVERALL GOALS

The overall goal of this research is to characterize regulatory interactions between proteins and RNA or small peptides. To achieve this goal, we proposed to accomplish the following:

1. Map and characterize the RNA binding domains and motifs in the Equine Infectious Anemia Virus Rev protein, focusing on those that interact with the Rev-responsive element (RRE) (Chapter 2)
2. Analyze the RNA secondary structure of the EIAV RRE and characterize the sequences and/or structures within it that are recognized and bound by the EIAV Rev protein (Chapter 3)
3. Generate computational models for equine MHC class I molecules in complexes with epitope peptides derived from EIAV proteins (Env, Gag and Rev), to investigate how MHC class I molecules differentially recognize the different viral epitope peptides (Chapter 4)

Also we proposed to explore other aspects of macromolecular structure and interactions in studies described in the Appendices:

4. Predict the structures of the N-terminal domain of the human and yeast telomerase reverse transcriptase enzymes, using a combination of homology

modeling and threading approaches; analyze and predict nucleic acid binding sites in the same domain, using machine learning approaches (Appendix A)

6. Analyze potentially regulatory conformational changes that occur in proteins upon proline *cis/trans* isomerization (Appendix B)

DISSERTATION ORGANIZATION

The dissertation has five chapters and two appendices. ***Chapter 1*** is a general introduction to the biology of Equine Infectious Anemia Virus (EIAV), and describes what was known about the interactions between EIAV Rev and the RRE when this study was initiated. A brief description of the equine immune response to the viral infection is included. ***Chapter 2*** is a paper published in the *Journal of Virology* in 2006 (35), in which mapping and characterization of functional domains in EIAV Rev are described. Identification of the domains essential for RNA binding activity demonstrated that the RNA binding domain of EIAV Rev is bipartite, with two critical arginine-rich RNA binding motifs (ARMs) separated by 76 amino acids in the primary sequence of the proteins. The contributions of other authors to this paper are as follows: *In vivo* analyses of the roles of specific residues in NLS motif were done by Sean C. Murphy and *in vivo* functional activity assays of Rev point mutation mutants (in Fig 2.5B) were performed by Wendy O. Sparks. Plasmid constructs I generated in this work were based on Rev-encoding plasmids from Michael Belshan's published work (4). I constructed all Rev deletion mutants, expressed and purified all proteins, performed all RNA binding assays, wrote the first draft of the paper and participated in revisions and editing. ***Chapter 3*** is a manuscript submitted recently to *Molecular and Cellular Biology*. It describes the detailed computational and experimental analysis of the secondary structure of the EIAV RRE and characterization of its interaction with the EIAV Rev protein. This work was done in collaboration with Gloria Culver. I

conceived of this study and designed and performed all of the chemical modification and footprint analyses, with assistance from members of Culver's lab, especially Laura Dutca. I performed all the computational analyses and predictions, wrote the first draft of the paper, and participated in revisions and editing. ***Chapter 4*** is a paper published in the *Journal of Immunology* in 2006 (40). This study was conceived by Robert H. Mealey, at Washington State University. All wet-lab experiments were performed in the Mealey lab, as were analyses of experiments shown in Figures. 4.1, 4.2 and 4.3. I performed all computational analyses and experiments, including the homology modeling and docking of equine MHC I molecules and epitope peptides. ***Chapter 5*** is a summary in which general conclusions of this dissertation study and future directions, based on my recommendations, are presented.

In the Appendices, several studies related to the main goal in this dissertation are introduced and described. ***Appendix A*** is a paper recently accepted for publication in the *Pacific Symposium on Biocomputing*, PSB08 (34). It describes structural modeling and functional residue prediction (DNA and RNA interfacial residues) for human and yeast telomerase reverse transcriptase (TERT) enzymes. The modeling N-terminal domains of human and yeast TERT proteins was performed Michael Hamilton and me. Prediction of DNA interacting residues was done by Cornelia Caragea in the lab of Vasant Honavar. Prediction of RNA interacting residues using *RNABindR* and comparisons between available experimental data and predictions on telomerases were carried out by Colin Gleeson, in the Dobbs lab. ***Appendix B*** is a study of the relationship between sequence, structure and function in folded proteins with prolines that undergo *cis/trans* isomerization. When complete, we plan to submit a manuscript decribing this look to *Journal of Molecular Biology*. All computational experiments, including the classification experiments using machine learning algorithms, and the large-scale identification and analysis of prolyl isomerization, were performed by me.

REFERENCES

1. **Addo, M. M., M. Altfeld, E. S. Rosenberg, R. L. Eldridge, M. N. Philips, K. Habeeb, A. Khatri, C. Brander, G. K. Robbins, G. P. Mazzara, P. J. Goulder, and B. D. Walker.** 2001. The HIV-1 regulatory proteins Tat and Rev are frequently targeted by cytotoxic T lymphocytes derived from HIV-1-infected individuals. Proc Natl Acad Sci U S A **98:**1781-6.

2. **Battiste, J. L., H. Mao, N. S. Rao, R. Tan, D. R. Muhandiram, L. E. Kay, A. D. Frankel, and J. R. Williamson.** 1996. Alpha helix-RNA major groove recognition in an HIV-1 rev peptide-RRE RNA complex. Science **273:**1547-51.

3. **Belshan, M., P. Baccam, J. L. Oaks, B. A. Sponseller, S. C. Murphy, J. Cornette, and S. Carpenter.** 2001. Genetic and biological variation in equine infectious anemia virus Rev correlates with variable stages of clinical disease in an experimentally infected pony. Virology **279:**185-200.

4. **Belshan, M., M. E. Harris, A. E. Shoemaker, T. J. Hope, and S. Carpenter.** 1998. Biological characterization of Rev variation in equine infectious anemia virus. J Virol **72:**4421-6.

5. **Belshan, M., G. S. Park, P. Bilodeau, C. M. Stoltzfus, and S. Carpenter.** 2000. Binding of equine infectious anemia virus rev to an exon splicing enhancer mediates alternative splicing and nuclear export of viral mRNAs. Mol Cell Biol **20:**3550-7.

6. **Betts, M. R., J. F. Krowka, T. B. Kepler, M. Davidian, C. Christopherson, S. Kwok, L. Louie, J. Eron, H. Sheppard, and J. A. Frelinger.** 1999. Human immunodeficiency virus type 1-specific cytotoxic T lymphocyte activity is inversely correlated with HIV type 1 viral load in HIV type 1-infected long-term survivors. AIDS Res Hum Retroviruses **15:**1219-28.

7. **Bjorkman, P. J., and P. Parham.** 1990. Structure, function, and diversity of class I major histocompatibility complex molecules. Annu Rev Biochem **59:**253-88.

8. **Blanco, F. J., S. Hess, L. K. Pannell, N. W. Rizzo, and R. Tycko.** 2001. Solid-state NMR data support a helix-loop-helix structural model for the N-terminal half of HIV-1 Rev in fibrillar form. J Mol Biol **313:**845-59.

9. **Borrow, P., H. Lewicki, B. H. Hahn, G. M. Shaw, and M. B. Oldstone.** 1994. Virus-specific CD8+ cytotoxic T-lymphocyte activity associated with control of viremia in primary human immunodeficiency virus type 1 infection. J Virol **68:**6103-10.

10. **Brady, J., and F. Kashanchi.** 2005. Tat gets the "green" light on transcription initiation. Retrovirology **2:**69.

11. **Chung, H., and D. Derse.** 2001. Binding sites for Rev and ASF/SF2 map to a 55-nucleotide purine-rich exonic element in equine infectious anemia virus RNA. J Biol Chem **276:**18960-7.

12. **Cook, K. S., G. J. Fisk, J. Hauber, N. Usman, T. J. Daly, and J. R. Rusche.** 1991. Characterization of HIV-1 REV protein: binding stoichiometry and minimal RNA substrate. Nucleic Acids Res **19:**1577-83.

13. **Cullen, B. R.** 1992. Mechanism of action of regulatory proteins encoded by complex retroviruses. Microbiol Rev **56:**375-94.

14. **Daly, T. J., R. C. Doten, P. Rennert, M. Auer, H. Jaksche, A. Donner, G. Fisk, and J. R. Rusche.** 1993. Biochemical characterization of binding of multiple HIV-1 Rev monomeric proteins to the Rev responsive element. Biochemistry **32:**10497-505.

15. **Dillon, P. J., P. Nelbock, A. Perkins, and C. A. Rosen.** 1990. Function of the human immunodeficiency virus types 1 and 2 Rev proteins is dependent on their ability to interact with a structured region present in env gene mRNA. J Virol **64:**4428-37.

16. **Flint, S. J., L. W. Enquist, V. R. Racaniello, and A. M. Skalka.** 2004. Principles of virology, Second edition ed. ASM Press.

17. **Foley, B. T.** 2000. An overview of the molecular phylogeny of lentiviruses. HIV Sequence Compendium 2000:35-43.

18. **Fridell, R. A., H. P. Bogerd, and B. R. Cullen.** 1996. Nuclear export of late HIV-1 mRNAs occurs via a cellular protein export pathway. Proc Natl Acad Sci U S A **93:**4421-4.

19. **Fridell, R. A., K. M. Partin, S. Carpenter, and B. R. Cullen.** 1993. Identification of the activation domain of equine infectious anemia virus rev. J Virol **67:**7317-23.

20. **Gao, G. F., B. E. Willcox, J. R. Wyer, J. M. Boulter, C. A. O'Callaghan, K. Maenaka, D. I. Stuart, E. Y. Jones, P. A. Van Der Merwe, J. I. Bell, and B. K. Jakobsen.** 2000. Classical and nonclassical class I major histocompatibility complex molecules exhibit subtle conformational differences that affect binding to CD8alphaalpha. J Biol Chem **275:**15232-8.

21. **Gao, X., G. W. Nelson, P. Karacki, M. P. Martin, J. Phair, R. Kaslow, J. J. Goedert, S. Buchbinder, K. Hoots, D. Vlahov, S. J. O'Brien, and M. Carrington.** 2001. Effect of a single amino acid change in MHC class I molecules on the rate of progression to AIDS. N Engl J Med **344:**1668-75.

22. **Gea-Banacloche, J. C., S. A. Migueles, L. Martino, W. L. Shupert, A. C. McNeil, M. S. Sabbaghian, L. Ehler, C. Prussin, R. Stevens, L. Lambert, J. Altman, C. W. Hallahan, J. C. de Quiros, and M. Connors.** 2000. Maintenance of large numbers of virus-specific CD8+ T cells in HIV-infected progressors and long-term nonprogressors. J Immunol **165:**1082-92.

23. **Giot, L., J. S. Bader, C. Brouwer, A. Chaudhuri, B. Kuang, Y. Li, Y. L. Hao, C. E. Ooi, B. Godwin, E. Vitols, G. Vijayadamodar, P. Pochart, H. Machineni, M. Welsh, Y. Kong, B. Zerhusen, R. Malcolm, Z. Varrone, A. Collis, M. Minto, S. Burgess, L. McDaniel, E. Stimpson, F. Spriggs, J. Williams, K. Neurath, N. Ioime, M. Agee, E.**

Voss, K. Furtak, R. Renzulli, N. Aanensen, S. Carrolla, E. Bickelhaupt, Y. Lazovatsky, A. DaSilva, J. Zhong, C. A. Stanyon, R. L. Finley, Jr., K. P. White, M. Braverman, T. Jarvie, S. Gold, M. Leach, J. Knight, R. A. Shimkets, M. P. McKenna, J. Chant, and J. M. Rothberg. 2003. A protein interaction map of Drosophila melanogaster. Science **302:**1727-36.

24. **Gontarek, R. R., and D. Derse.** 1996. Interactions among SR proteins, an exonic splicing enhancer, and a lentivirus Rev protein regulate alternative splicing. Mol Cell Biol **16:**2325-31.

25. **Gorlich, D., and I. W. Mattaj.** 1996. Nucleocytoplasmic transport. Science **271:**1513-8.

26. **Harris, M. E., R. R. Gontarek, D. Derse, and T. J. Hope.** 1998. Differential requirements for alternative splicing and nuclear export functions of equine infectious anemia virus Rev protein. Mol Cell Biol **18:**3889-99.

27. **Hewitt, E. W.** 2003. The MHC class I antigen presentation pathway: strategies for viral immune evasion. Immunology **110:**163-9.

28. **Hope, T. J.** 1999. The ins and outs of HIV Rev. Arch Biochem Biophys **365:**186-91.

29. **Jain, C., and J. G. Belasco.** 2001. Structural model for the cooperative assembly of HIV-1 Rev multimers on the RRE as deduced from analysis of assembly-defective mutants. Mol Cell 7:603-14.

30. **Janin, J.** 2005. Assessing predictions of protein-protein interaction: the CAPRI experiment. Protein Sci **14:**278-83.

31. **Jin, X., X. Gao, M. Ramanathan, Jr., G. R. Deschenes, G. W. Nelson, S. J. O'Brien, J. J. Goedert, D. D. Ho, T. R. O'Brien, and M. Carrington.** 2002. Human immunodeficiency virus type 1 (HIV-1)-specific CD8+-T-cell responses for groups of HIV-1-infected individuals with different HLA-B*35 genotypes. J Virol **76:**12603-10.

32. **Kjems, J., M. Brown, D. D. Chang, and P. A. Sharp.** 1991. Structural analysis of the interaction between the human immunodeficiency virus Rev protein and the Rev response element. Proc Natl Acad Sci U S A **88:**683-7.

33. **Le, S. Y., M. H. Malim, B. R. Cullen, and J. V. Maizel.** 1990. A highly conserved RNA folding region coincident with the Rev response element of primate immunodeficiency viruses. Nucleic Acids Res **18:**1613-23.

34. **Lee, J. H., M. Hamilton, C. Gleeson, C. Caragea, P. Zaback, J. D. Sander, X. Li, F. Wu, M. Terribilini, V. Honavar, and D. Dobbs.** 2008. Presented at the Pacific Symposium on Biocomputing (PSB) 2008.

35. **Lee, J. H., S. C. Murphy, M. Belshan, W. O. Sparks, Y. Wannemuehler, S. Liu, T. J. Hope, D. Dobbs, and S. Carpenter.** 2006. Characterization of functional domains of equine infectious anemia virus Rev suggests a bipartite RNA-binding domain. J Virol **80:**3844-52.

36. **Liao, H. J., C. C. Baker, G. L. Princler, and D. Derse.** 2004. cis-Acting and trans-acting modulation of equine infectious anemia virus alternative RNA splicing. Virology **323:**131-40.

37. **Mann, D. A., I. Mikaelian, R. W. Zemmel, S. M. Green, A. D. Lowe, T. Kimura, M. Singh, P. J. Butler, M. J. Gait, and J. Karn.** 1994. A molecular rheostat. Co-operative rev binding to stem I of the rev-response element modulates human immunodeficiency virus type-1 late gene expression. J Mol Biol **241:**193-207.

38. **Martarano, L., R. Stephens, N. Rice, and D. Derse.** 1994. Equine infectious anemia virus trans-regulatory protein Rev controls viral mRNA stability, accumulation, and alternative splicing. J Virol **68:**3102-11.

39. **McGuire, T. C., S. R. Leib, R. H. Mealey, D. G. Fraser, and D. J. Prieur.** 2003. Presentation and binding affinity of equine infectious anemia virus CTL envelope and

matrix protein epitopes by an expressed equine classical MHC class I molecule. J Immunol **171:**1984-93.

40. **Mealey, R. H., J. H. Lee, S. R. Leib, M. H. Littke, and T. C. McGuire.** 2006. A single amino acid difference within the alpha-2 domain of two naturally occurring equine MHC class I molecules alters the recognition of Gag and Rev epitopes by equine infectious anemia virus-specific CTL. J Immunol **177:**7377-90.

41. **Mealey, R. H., A. Sharif, S. A. Ellis, M. H. Littke, S. R. Leib, and T. C. McGuire.** 2005. Early detection of dominant Env-specific and subdominant Gag-specific CD8+ lymphocytes in equine infectious anemia virus-infected horses using major histocompatibility complex class I/peptide tetrameric complexes. Virology **339:**110-26.

42. **Mealey, R. H., B. Zhang, S. R. Leib, M. H. Littke, and T. C. McGuire.** 2003. Epitope specificity is critical for high and moderate avidity cytotoxic T lymphocytes associated with control of viral load and clinical disease in horses with equine infectious anemia virus. Virology **313:**537-52.

43. **Muller, D., K. Pederson, R. Murray, and J. A. Frelinger.** 1991. A single amino acid substitution in an MHC class I molecule allows heteroclitic recognition by lymphocytic choriomeningitis virus-specific cytotoxic T lymphocytes. J Immunol **147:**1392-7.

44. **Nigg, E. A.** 1997. Nucleocytoplasmic transport: signals, mechanisms and regulation. Nature **386:**779-87.

45. **Nolan, D., S. Gaudieri, and S. Mallal.** 2006. Host genetics and viral infections: immunology taught by viruses, virology taught by the immune system. Curr Opin Immunol **18:**413-21.

46. **Olsen, H. S., A. W. Cochrane, P. J. Dillon, C. M. Nalin, and C. A. Rosen.** 1990. Interaction of the human immunodeficiency virus type 1 Rev protein with a structured

region in env mRNA is dependent on multimer formation mediated through a basic stretch of amino acids. Genes Dev **4:**1357-64.

47. **Phillips, T. R., C. Lamont, D. A. Konings, B. L. Shacklett, C. A. Hamson, P. A. Luciw, and J. H. Elder.** 1992. Identification of the Rev transactivation and Rev-responsive elements of feline immunodeficiency virus. J Virol **66:**5464-71.

48. **Pollard, V. W., and M. H. Malim.** 1998. The HIV-1 Rev protein. Annu Rev Microbiol **52:**491-532.

49. **Saltarelli, M. J., R. Schoborg, G. N. Pavlakis, and J. E. Clements.** 1994. Identification of the caprine arthritis encephalitis virus Rev protein and its cis-acting Rev-responsive element. Virology **199:**47-55.

50. **Sette, A., and J. Sidney.** 1998. HLA supertypes and supermotifs: a functional perspective on HLA polymorphism. Curr Opin Immunol **10:**478-82.

51. **Tiley, L. S., and B. R. Cullen.** 1992. Structural and functional analysis of the visna virus Rev-response element. J Virol **66:**3609-15.

52. **Watts, N. R., M. Misra, P. T. Wingfield, S. J. Stahl, N. Cheng, B. L. Trus, A. C. Steven, and R. W. Williams.** 1998. Three-dimensional structure of HIV-1 Rev protein filaments. J Struct Biol **121:**41-52.

53. **Zapp, M. L., and M. R. Green.** 1989. Sequence-specific RNA binding by the HIV-1 Rev protein. Nature **342:**714-6.

54. **Zapp, M. L., T. J. Hope, T. G. Parslow, and M. R. Green.** 1991. Oligomerization and RNA binding domains of the type 1 human immunodeficiency virus Rev protein: a dual function for an arginine-rich binding motif. Proc Natl Acad Sci U S A **88:**7734-8.

CHAPTER 2. CHARACTERIZATION OF FUNCTIONAL DOMAINS OF EQUINE INFECTIOUS ANEMIA VIRUS REV SUGGESTS A BIPARTITE RNA-BINDING DOMAIN

A paper published in *Journal of Virology*

Jae-Hyung Lee, Sean C. Murphy, Michael Belshan, Wendy O. Sparks, Yvonne Wannemuehler, Sijun Liu, Thomas J. Hope, Drena Dobbs, and Susan Carpenter

ABSTRACT

Equine infectious anemia virus (EIAV) Rev is an essential regulatory protein that facilitates expression of viral mRNAs encoding structural proteins and genomic RNA and regulates alternative splicing of the bicistronic tat/rev mRNA. EIAV Rev is characterized by a high rate of genetic variation in vivo, and changes in Rev genotype and phenotype have been shown to coincide with changes in clinical disease. To better understand how genetic variation alters Rev phenotype, we undertook deletion and mutational analyses to map functional domains and to identify specific motifs that are essential for EIAV Rev activity. All functional domains are contained within the second exon of EIAV Rev. The overall organization of domains within Rev exon 2 includes a nuclear export signal, a large central region required for RNA binding, a nonessential region, and a C-terminal region required for both nuclear localization and RNA binding. Subcellular localization of green fluorescent protein-Rev mutants indicated that basic residues within the KRRRK motif in the C-terminal region of Rev are necessary for targeting of Rev to the nucleus. Two separate regions of Rev were necessary for RNA binding: a central region encompassing residues 57 to 130 and a C-terminal region spanning residues 144 to 165. Within these regions were two distinct, short

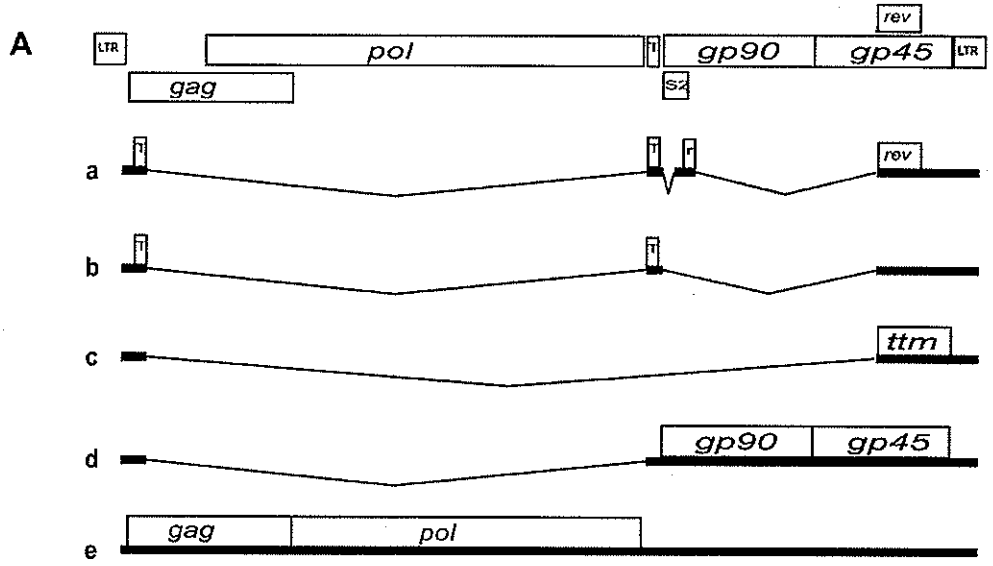
arginine-rich motifs essential for RNA binding, including an RRDRW motif in the central region and the KRRRK motif near the C terminus. These findings suggest that EIAV Rev utilizes a bipartite RNA-binding domain.

INTRODUCTION

Equine infectious anemia virus (EIAV) infection of horses can result in a rapid, variable, and dynamic disease course. Moreover, horses that survive the early clinical episodes of disease are generally able to control virus replication and remain clinically normal, inapparent carriers of EIAV. The unique features of clinical disease, and the ability of some infected horses to eventually control virus replication, provide an excellent system for longitudinal analyses of virus and host factors important in lentivirus persistence and pathogenesis. Genetic diversity is a hallmark of lentiviruses and is considered an important mechanism of virus persistence and pathogenesis. Previous studies have identified a high rate of genetic variation in EIAV in the region overlapped by the transmembrane protein gp45 (TM) and the major exon of Rev (2, 30). Genetic variation in rev/tm can significantly alter Rev activity (7), and in vivo studies suggest that changes in Rev phenotype correlate with changes in the clinical stage of disease (4, 6). In particular, Rev is significantly less active during the inapparent compared to the chronic stage of disease, suggesting that the Rev phenotype contributes to selection of virus variants in vivo. Insight into the genetic changes and factors that contribute to Rev selection in vivo requires identification of the functional domains and motifs that mediate EIAV Rev activity.

The Rev/Rex proteins of complex retroviruses differentially regulate expression of incompletely spliced mRNAs encoding virion structural and enzymatic proteins and progeny RNA molecules (reviewed in reference 17). The prototypical member of this family, human immunodeficiency type 1 (HIV-1) Rev, binds to the viral pre-mRNA at a specific sequence

called the Rev-responsive element (RRE) (15, 48), multimerizes (37, 47), and facilitates export of incompletely spliced RNAs from the nucleus via a nucleoporin pathway distinct from that used by most cellular mRNAs (18, 19). Mutational analyses indicate that the activities of HIV-1 Rev are mediated by discrete functional domains, including an N-terminal arginine-rich RNA-binding motif (ARM), which also functions as a nuclear localization signal (NLS) (24, 32), and a C-terminal leucine-rich nuclear export signal (NES) (18, 19). EIAV Rev is a 165-amino-acid protein translated from exons 3 and 4 of a multiply spliced, 4-exon, bicistronic mRNA that also encodes the trans-activating protein Tat (Fig. 2.1A) (12). EIAV Rev is functionally homologous to HIV-1 Rev (20, 33) but is less well characterized. The N-terminal leucine-rich NES, which maps to amino acids 31 to 55 (Fig. 2.1B) (20, 23, 33), is similar to other leucine-rich viral and cellular export proteins that interact with the nuclear exporter CRM1; however, EIAV Rev is atypical in the spacing of the leucine residues within the NES (23, 38). The C-terminal basic region has been found to be important for nuclear localization, while the central region of the protein has been implicated in RNA binding (14, 23).



B

1 MAESKEARDQEMNLKEESKEEKRRNDWWKIDPQGPLESDQWCRVLRQSLPEEKIISQTCI 60

61 ARRHLGPGPTQHTPSRRDRWIRGQILQAEVLQERLEWRIRGVQQVAKELGEVNRGIWREL 120

121 YFREDQRGDFSAWGGYQRAQERLWGEQSSPRVLRPGDSKRRRKHL 165

Figure 2.1. Organization and splicing patterns of EIAV and the EIAV Rev amino acid sequence. **(A)** Schematic of the EIAV genome showing open reading frames (ORF) and predominant mRNAs **(a to e)** isolated from virus-infected tissue culture cells (27). Regulatory proteins Tat (T) and Rev (r, rev) are translated from the four-exon mRNA **(a).** In the presence of Rev, EIAV exon 3 is skipped, resulting in a three-exon multiply spliced mRNA that encodes only Tat **(b).** Structural proteins and progeny RNA molecules are translated from singly spliced **(d)** and unspliced mRNAs **(e).** Ttm, a protein of unknown function (5), is encoded by a two-exon mRNA **(c). (B)** The amino acid sequence of the wild-type EIAV Rev H21 (7) is shown and numbered 1 to 165. The nuclear export signal (aa 31 to 55) is boxed.

In addition to promoting nuclear export of incompletely spliced RNA, EIAV Rev regulates alternative splicing of the viral RNA. In the presence of Rev, the multiply spliced

mRNA, which lacks exon 3 (Fig. 2.1A), is produced (34). Exon 3 contains the translational start site for Rev, and alternative splicing was originally proposed as a novel mechanism for autoregulation of Rev expression (21, 23). Exon 3 is flanked by a suboptimal splice acceptor and contains a purine-rich exonic splicing enhancer (ESE) that interacts with the SR protein SF2/ASF (21). The ESE also functions as an EIAV RRE (6), and we have suggested that EIAV Rev mediates alternative splicing of exon 3 through protein-RNA interactions required for efficient export of incompletely spliced viral RNAs. While the exact mechanism of alternative splicing is not known, current models (6, 14) agree that alternative splicing requires both nuclear localization and RNA binding and perhaps an as yet unidentified domain(s) of Rev that interacts directly with SF2/ASF and/or other cellular splicing factors. To date, the specific amino acids that mediate EIAV Rev nuclear import, RNA binding, and alternative splicing have not been identified. Using a series of deletions and mutations in EIAV Rev cDNA, we mapped functional domains of EIAV Rev and identified specific motifs required for nuclear localization and RNA binding. Two noncontiguous, short ARMs were required for RNA binding, suggesting that EIAV Rev contains a bipartite RNA-binding domain.

MATERIALS AND METHODS

Construction of Rev mutants

EIAV Rev cDNA deletion mutants were generated in the previously described plasmid pRevWT (7). Rev mutants containing internal deletions were constructed by PCR-ligation-PCR mutagenesis as described by Ali and Steinkasserer (3). Briefly, upstream and downstream blunt-ended cDNA fragments of each mutant were amplified from pRevWT using the primers designed from the EIAV Wyoming cell culture-adapted isolate (GenBank

accession no. M16575) (27). Specific primer sequences used for cloning are available upon request. Fragments were gel purified, and downstream fragments were phosphorylated and ligated to the corresponding upstream fragment. Ligation products were amplified by PCR using wild-type 5′ (CAGCATGGCAGAATCGAAGG) and wild-type 3′ (CGAGAGTTCCTTCTTGGGA) primers. Following the second PCR, the cDNAs were TA cloned into pCR3.1 (Invitrogen, Carlsbad, CA) and transformed into Escherichia coli DH5-{alpha}, and transformants were screened by colony blot hybridization. The C-terminal deletion mutants were PCR amplified using the 5′ wild-type flanking primer and unique 3′ primers that generated premature stop codons. PCR was performed using standard methods, and cDNAs were TA cloned into pCR3.1 as described above. Site-specific mutations were introduced by PCR-based mutagenesis. All constructs were confirmed by sequencing, and protein expression was verified by Western blotting using EIAV convalescent horse sera (10).

Rev fusion proteins

To construct green fluorescent protein (GFP) fusion proteins, Rev sequences were PCR amplified from wild-type and deletion constructs with primers that introduced 5′ EcoRI and 3′ BamHI restriction sites. PCR products were digested with EcoRI and BamHI and cloned into the GFP expression vector pEGFP-C2 (Clontech, Palo Alto, CA). The C-terminal region of Rev (specifying amino acids 145 to 165) was synthesized as complementary oligonucleotides that created 5′ EcoRI and 3′ BamHI overhangs after annealing. Wild-type and mutant oligonucleotides were synthesized, annealed, digested with EcoRI and BamHI, and cloned into similarly restricted pGFP-RDM12. All plasmids were sequenced to verify that each mutant Rev was translated from a single open reading frame.

A series of maltose-binding protein (MBP)-Rev fusion proteins were used for RNA-binding studies. ERev fragments amplified from EIAV R1A (6) or Rev cDNA plasmids were

cloned in pHMTc (43), which is based on the pMal-c2x expression vector (New England Biolabs, Beverly, MA). For most MBP-ERev constructs, Rev cDNA templates were amplified from EIAV variant R1A; MRD8 contains the Rev deletion mutant RDM11, which is based on pRevWT (see above). MBP-ERev constructs containing point mutations were cloned using amplified mutated PCR fragments based on R1A Rev cDNA template. All plasmid constructs were confirmed by DNA sequence analysis.

CAT assays

Rev nuclear export activity was quantified in transient transfection assays using a pDM138-based EIAV Rev reporter construct, pERRE-All, as previously described (7). 293T cells were seeded in triplicate at 1 x 10^5 to 5 x 10^5 cells/well in 6-well tissue culture dishes and maintained in Dulbecco's modified Eagle medium (DMEM) supplemented with 10% fetal calf serum (FCS) and penicillin-streptomycin (P-S) at 37°C and 8% CO_2. One day after seeding, cells were transfected with 1.0 µg Rev cDNA or empty vector, 0.2 µg pERRE-All reporter plasmid DNA, 0.2 µg beta-galactosidase reporter plasmid DNA, and sufficient pUC19 to bring each well to 2.0 µg total transfected DNA. Medium was changed the next day, and cells were harvested at 48 h posttransfection, lysed by freezing-thawing, and normalized for transfection efficiency by measuring beta-galactosidase activity as described previously (7). Cell lysates were assayed for ß-galactosidase activity and normalized reaction volumes were assayed for chloramphenicol acetyltransferase (CAT) enzyme activity or for CAT protein using thin-layer chromatography or a commercial CAT enzyme-linked immunosorbent assay kit (Roche Molecular Biochemicals, Indianapolis, IN), respectively (7, 8). All assays were performed in triplicate, and results represent at least six independent transfections. Statistical analysis was performed using analysis of variance and Student's t test assuming unequal variance among groups.

Trans-complementation assays

Cf2Th (ATCC no. CRL-1430) clonal cell lines containing Rev (+) or Rev (–) EIAV proviral DNA were used to characterize nuclear export activity (7). Cf2Th/112 cells contain a Rev-competent proviral clone and produce supernatant reverse transcriptase (RT) activity, viral structural proteins, and all classes of viral mRNAs. Cf2Th/51 cells contain Rev-defective proviral DNA, express only fully spliced mRNA, and lack detectable levels of viral structural proteins or RT activity. *Trans*-complementation of Cf2Th/51 cells with Rev cDNA results in viral protein expression and RT activity (7). Cells were seeded at 5×10^5 cells/well in 6-well tissue culture plates and maintained in DMEM supplemented with 10% FCS and P-S at 37°C and 8% CO_2. The next day, cells were transfected with 2 μg of wild-type Rev cDNA, Rev mutant cDNA, or empty vector DNA using the liposome-mediated transfection reagent Lipofectamine (Invitrogen). At 3 days posttransfection, supernatant was collected and assayed for RT activity as previously described (11).

Nuclear localization assays and microscopy

Cf2Th cells were plated at 2×10^4 cells/cm^2 in 6- or 24-well plates or 8-well glass chamber slides (Nunc, Rochester, NY). Twenty-four hours after plating, cells were transiently transfected with 0.25 μg GFP or GFP-Rev plasmid DNA per 2×10^4 cells using the liposome-mediated transfection reagent TransIT-LT1 according to manufacturer instructions (Mirus, Madison, WI). In certain experiments, replicate cultures were treated at 20 h posttransfection with 5 nM leptomycin B (LMB) in dimethyl sulfoxide. At 24 h after transfection, cells were fixed in 3.7% formaldehyde in 10 mM phosphate-buffered saline (PBS) for 30 min at 25°C and washed twice with complete DMEM supplemented with 10% FCS and P-S. Nuclei of fixed cells were stained with 0.5 μg/ml Hoechst 33258 dye (Sigma,

St. Louis, MO) in 0.5% NP-40 and 10 mM PBS for 15 min at 25°C. Cells were subsequently washed twice in complete DMEM and once in 10 mM PBS. All transfections were performed in triplicate.

Fixed cells in 6- and 24-well plates were examined with an inverted Nikon Diaphot fluorescence microscope with a 40x objective and a 100-W high-pressure mercury lamp; epifluorescence filters were used to visualize Hoechst-stained nuclei and GFP. For confocal microscopy, chambers were removed from slides and a coverslip was sealed over fixed cells; slides were examined with a Leica TCS NT laser confocal microscope using a 63x oil-immersion objective; digital filters with 400- to 480-nm and 500- to 560-nm excitation wavelengths were necessary to visualize Hoechst-stained DNA and GFP, respectively. Brightness and contrast of images obtained by confocal microscopy were adjusted with Adobe Photoshop 4.0.

RNA-binding assays

MBP-Rev fusion proteins were expressed in Rosetta-Gami DE3 (pLacI) (Novagen, Madison, WI). Harvested cells were lysed by freezing-thawing, and His-tagged fusion proteins were purified under native conditions using Ni^{2+}-charged resin (Invitrogen). The purity of fusion proteins was confirmed by sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) followed by Coomassie staining or Western blotting using EIAV convalescent-phase sera or polyclonal antibodies directed against the His Tag (MBL International, Woburn, MA). Purified MBP-Rev proteins were dialyzed and stored in 50 mM Tris-HCl (pH 8.0), 50 mM NaCl at 4°C. The EIAV Rev-responsive element (RRE) (nucleotides 5443 to 5565) was amplified by PCR from pERRE-All (8) using a 5′ primer containing a T7 promoter site. The PCR product was purified using QIAquick PCR purification columns (QIAGEN, Valencia, CA), and RNA was generated by in vitro

transcription (T7-MEGAshortscript; Ambion, Austin, TX) in the presence of [α-^{32}P]UTP. Transcribed radiolabeled RNA was purified on a G50 column (Roche, Indianapolis, IN), denatured at 80°C for 5 min in 20 mM Tris-HCl, pH 7.5, 100 mM NaCl, annealed by slow cooling, and stored at –80°C.

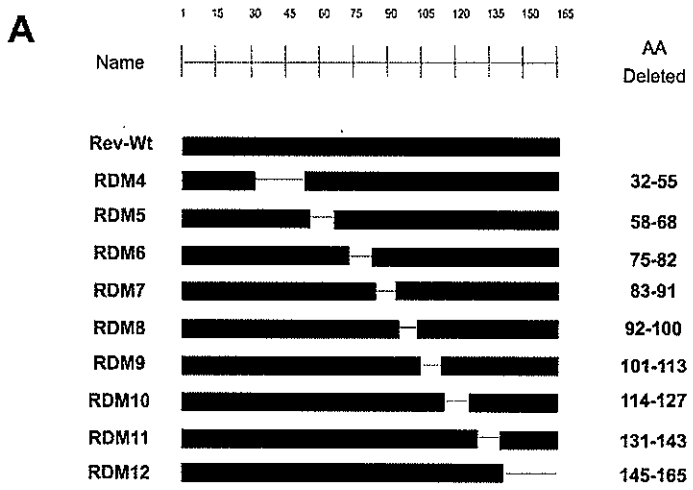
For RNA-binding reactions, 2 to 4 μg purified MBP-Rev fusion protein was incubated with 10^4 cpm of ^{32}P-labeled EIAV RRE RNA in binding buffer (10 mM HEPES-KOH, pH 7.5, 100 mM KCl, 1 mM $MgCl_2$, 0.5 mM EDTA, 1 mM dithiothreitol, 50 μg/ml E. coli tRNA, and 10% glycerol) for 20 min at room temperature. After incubation, reaction samples were irradiated with 3 x 10^5 μJ at 254 nm for 7 min. Samples were treated with 0.1 mg/ml RNase A at 37°C for 20 min; the reaction was terminated by boiling 5 min in an equal volume of SDS and separated in SDS-12% PAGE in Tris-glycine buffer. Gels were fixed in 50% methanol-10% acetic acid, dried, and exposed to PhosphorImager screens overnight. UV cross-linked complexes were quantified using a PersonalFX scanner and Quantity One software (Bio-Rad, Hercules, CA).

RESULTS

Functional analyses of Rev deletion mutants.

Lentiviral Rev proteins utilize discrete functional domains to control expression of viral mRNAs and structural proteins. To aid in identification of EIAV Rev functional domains, we constructed a nested set of Rev deletion mutants (RDM) (Fig. 2.2A) and tested them for nuclear export activity in transient transfection assays using the CAT reporter construct, pERRE-All (7). Nuclear export activity was highly sensitive to deletions in the Rev protein (Fig. 2.2B, panel 1). Mutants RDM4, RDM6-RDM10, and RDM12 showed no CAT activity, whereas activity was significantly reduced in RDM5. Only one mutant,

RDM11, showed levels of CAT activity comparable to wild-type Rev. To ensure that these results were not an artifact of the CAT reporter construct, all mutants were also tested by *trans*-complementation of a Rev-defective clonal cell line, Cf2Th/51 (Fig. 2.2B, panel 2). Overall, there was good agreement between the two assays: transfection with mutants RDM4-RDM10 and RDM12 showed no RT activity, while cells transfected with mutant RDM11 had activity similar to that of wild-type Rev. Western blot analysis using anti-EIAV polyclonal horse sera was done to confirm that all mutants expressed Rev. Some variation in protein expression was observed; however, the levels of expression did not correlate with levels of Rev activity. Mutants RDM4, RDM5, RDM6, RDM11, and RDM12 expressed at levels equal to or higher than Rev-WT, while RDM7, RDM8, and RDM10 expression levels were somewhat lower than Rev-WT. Only RDM9 showed markedly reduced levels of protein expression compared to other mutants. With the exception of RDM9, therefore, the loss of nuclear export activity was likely due to deletion of discrete functional domains and/or the loss of tertiary structure(s) required for activity.


A
1 15 30 45 60 75 90 105 120 135 150 165
Name
AA Deleted
Rev-Wt
RDM4 32-55
RDM5 58-68
RDM6 75-82
RDM7 83-91
RDM8 92-100
RDM9 101-113
RDM10 114-127
RDM11 131-143
RDM12 145-165

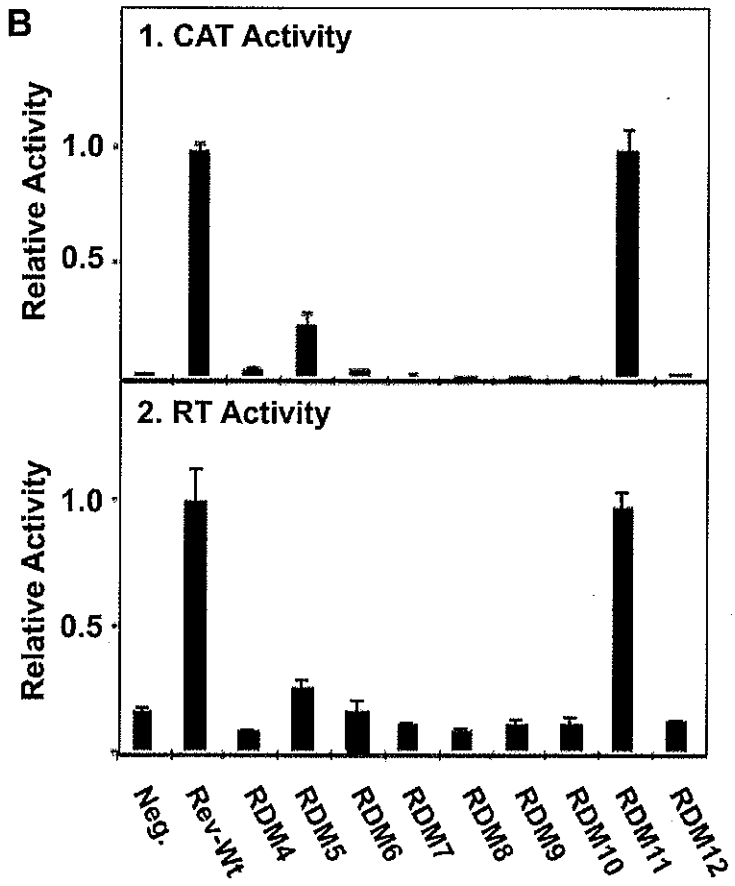

B
1. CAT Activity
Relative Activity
1.0
0.5
2. RT Activity
Relative Activity
1.0
0.5
Neg.
Rev-Wt
RDM4
RDM5
RDM6
RDM7
RDM8
RDM9
RDM10
RDM11
RDM12

Figure 2.2. Functional analyses of Rev deletion mutants. **(A)** Terminal and internal deletions in Rev cDNA were generated from pRevWT by PCR-ligation-PCR as described in the text. The region of amino acids deleted in each cDNA is indicated. **(B)** Activity of Rev deletion mutants was assayed in transient expression assays. Results are reported as activity relative to pRevWT. Cells transfected with the pCR3.1 vector DNA were used as the negative control (Neg.) in all assays. Error bars denote the standard error of the mean. Graph 1 shows nuclear export activity measured using a CAT reporter assay (7). Lysates were normalized by ß-galactosidase activity, and CAT activity was measured as the percentage of acetylation. Individual experiments included triplicate wells, and the data shown represent the means of at least three separate experiments. Graph 2 shows supernatant reverse transcriptase (RT) activity following *trans*-complementation of a Rev-defective clonal cell line, Cf2Th/51, with Rev deletion mutants. Wt, wild type.

The KRRRK motif in the C terminus of Rev is required for nuclear localization.

The highly basic C terminus of EIAV Rev contains residues essential for nuclear entry (23); however, the precise NLS has not been identified. To more specifically identify the amino acids necessary for nuclear localization, each of the Rev deletion mutants shown in Fig. 2.2A were used to construct GFP-Rev fusion proteins. Plasmid DNA was transiently transfected into Cf2Th cells, and subcellular localization of fusion proteins was assessed by fluorescence and confocal microscopy. Only two of the nine GFP-Rev deletion mutants exhibited subcellular localization patterns different from that of wild-type Rev (data not shown). GFP-RDM4, which contains a deletion in the NES, was found only in the nucleus, and GFP-RDM12, which lacks the 21 C-terminal residues of Rev, was found exclusively in the cytoplasm. The C terminus contains a strongly basic KRRRK motif that is similar to other basic NLSs (22, 25, 46) and was previously suggested to be a component of the EIAV Rev NLS (23). To identify critical residues within the C-terminal region, we introduced additional deletions or alanine substitutions across the last seven residues of GFP-Rev (Fig.

2.3A). Mutants were transfected into Cf2Th cells and assayed for subcellular localization in the presence and absence of leptomycin B (LMB) (Fig. 2.3B). Rev is a nucleocytoplasmic shuttling protein, and LMB was used to block export of fusion proteins once they had translocated to the nucleus. Deletion of six (GFP-RDM13) or two (GFP-RDM14) amino acids from the C terminus of EIAV Rev abrogated nuclear localization. Alanine substitution within the KRRRK motif indicated that mutation of the middle arginine (C3 = KRARK) or the terminal lysine (C6 = KRRRA) did not alter nuclear import; however, Rev-GFP proteins with alanine substitutions of any two adjacent basic residues within amino acids 159 to 163 (C1 = AARRK, C2 = KAARK, C4 = KRAAK, C5 = KRRAA) remained in the cytoplasm. Therefore, nuclear localization was dependent on the presence of a cluster of basic amino acids within the KRRRK motif in the C-terminal region.

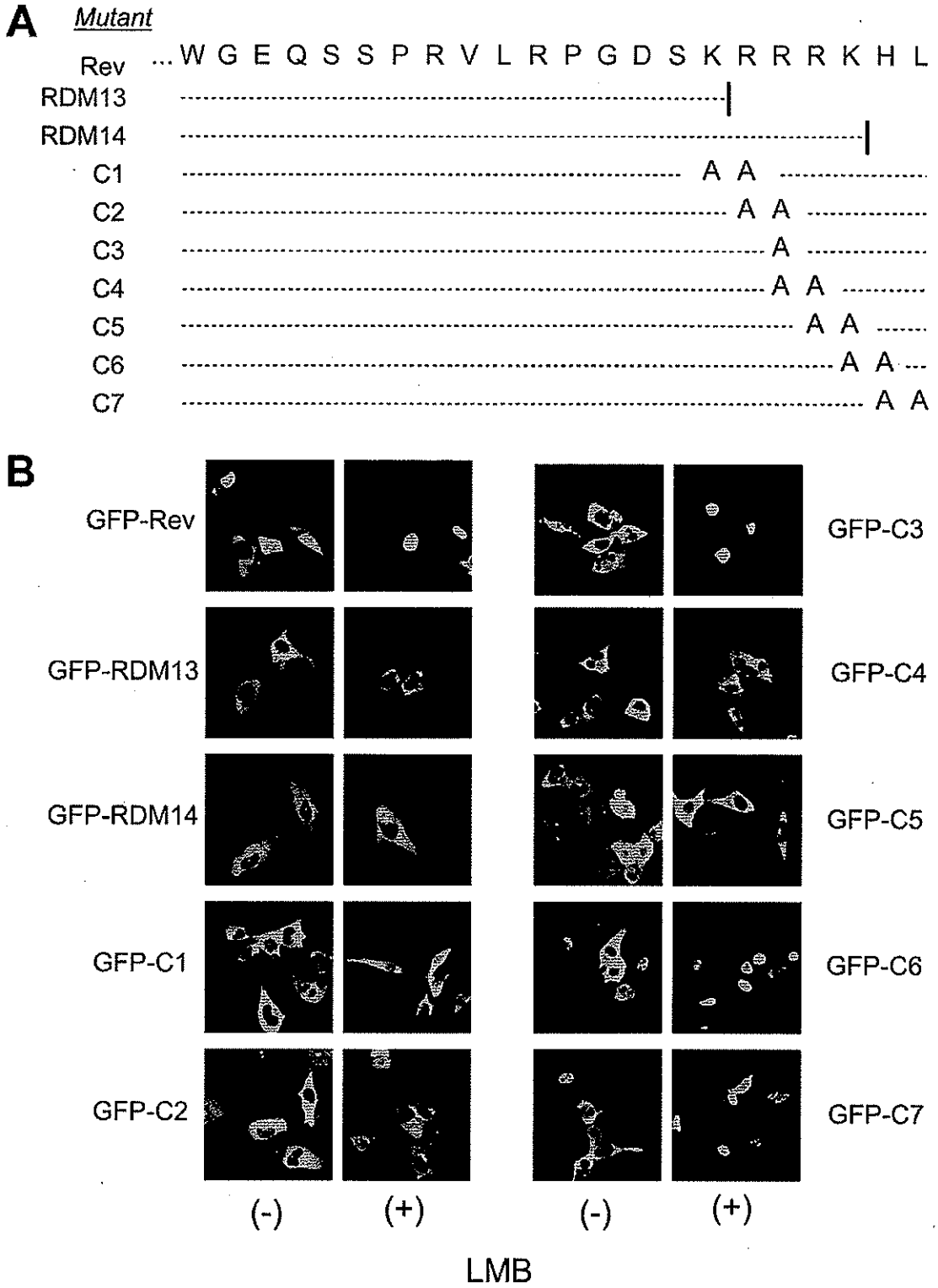

A
Mutant
Rev ... W G E Q S S P R V L R P G D S K R R R K H L
RDM13
RDM14
C1 A A
C2 A A
C3 A
C4 A A
C5 A A
C6 A A
C7 A A
B
GFP-Rev
GFP-RDM13
GFP-RDM14
GFP-C1
GFP-C2
GFP-C3
GFP-C4
GFP-C5
GFP-C6
GFP-C7
(-)
(+)
(-)
(+)
LMB

Figure 2.3. The KRRRK motif in the C terminus of Rev is required for nuclear localization. **(A)** Amino acid sequence of the C terminus of wild-type Rev (aa 144 to 165) and the location of deletions or alanine substitutions introduced into wild-type GFP-Rev mutants. **(B)** Subcellular location of GFP-Rev deletion mutants (GFP-RDM13 and GFP-RDM14) and GFP-Rev containing alanine substitutions in the C terminus. Cf2Th cells were transfected with the specified Rev-GFP cDNAs in the presence or absence of 5 nM leptomycin B (LMB). Images of fixed and stained Cf2Th cells were obtained by confocal laser microscopy. Brightness and contrast were adjusted with Adobe Photoshop 4.0.

Mapping the RNA-binding domain of EIAV Rev.

In HIV-1 Rev, the RNA-binding domain overlaps the NLS (24). In EIAV Rev, however, the NLS is located in the C-terminal basic region, while the RNA-binding domain is thought to reside within the central region of EIAV Rev (23). To better define the RNA-binding domain of EIAV Rev, we assessed RNA binding activity of a series of truncated MBP-Rev deletion (MRD) mutants in UV cross-linking assays (Fig. 2.4A). The purity and integrity of the MBP-Rev fusion proteins were confirmed by SDS-PAGE (Fig. 2.4B). RNA-protein complexes were observed in reactions containing either MBP-Rev or MRD1 and the sense strand of the RRE but not in reactions containing the antisense RRE RNA (Fig. 2.4C). RNA-protein complexes included a clearly defined, faster migrating band and diffuse slower migrating products. No RNA-protein complexes were observed with either bovine serum albumin or MBP, demonstrating the specificity of Rev-RNA binding (Fig. 2.4C).

Our analysis of the Rev-MBP deletion mutants showed that MRD1 and MRD3 each formed an RNA-protein complex, indicating that neither exon 1 (amino acids [aa] 1 to 30) nor the N-terminal leucine-rich NES (aa 31 to 56) was required for RNA binding. Constructs lacking either the last 20 or 40 C-terminal residues (MRD2, MRD4, and MRD5) did not form a complex with the EIAV RRE (Fig. 2.4D). However, no RRE binding was detected using

fusion proteins containing only C-terminal residues (MRD6 and MRD7). Therefore, residues within the C-terminal domain were necessary, but not sufficient, for RNA binding. To aid in identifying the boundaries of the Rev RNA binding domain, we constructed MRD8, a mutant with an internal deletion of aa 131 to 143. This is the same region deleted in RDM11 (Fig. 2.2), a deletion mutant with wild-type levels of nuclear export activity. MRD8 bound the RRE RNA at levels comparable to wild-type Rev and MRD1 (Fig. 2.4E). There appeared to be a relative increase in the slower migrating RNA-protein complexes (compare Fig. 2.4E with C and D); however, this may be due to variability between experiments rather than differences in Rev binding. The RNA-binding activity of MRD8, together with the lack of RNA binding in MRD5, MRD6, and MRD7, indicate that two, noncontiguous regions of EIAV are required for RNA-binding activity. One region encompasses aa 57 to 130, and the second region is located in the C terminus of Rev, encompassing aa 144 to 165.

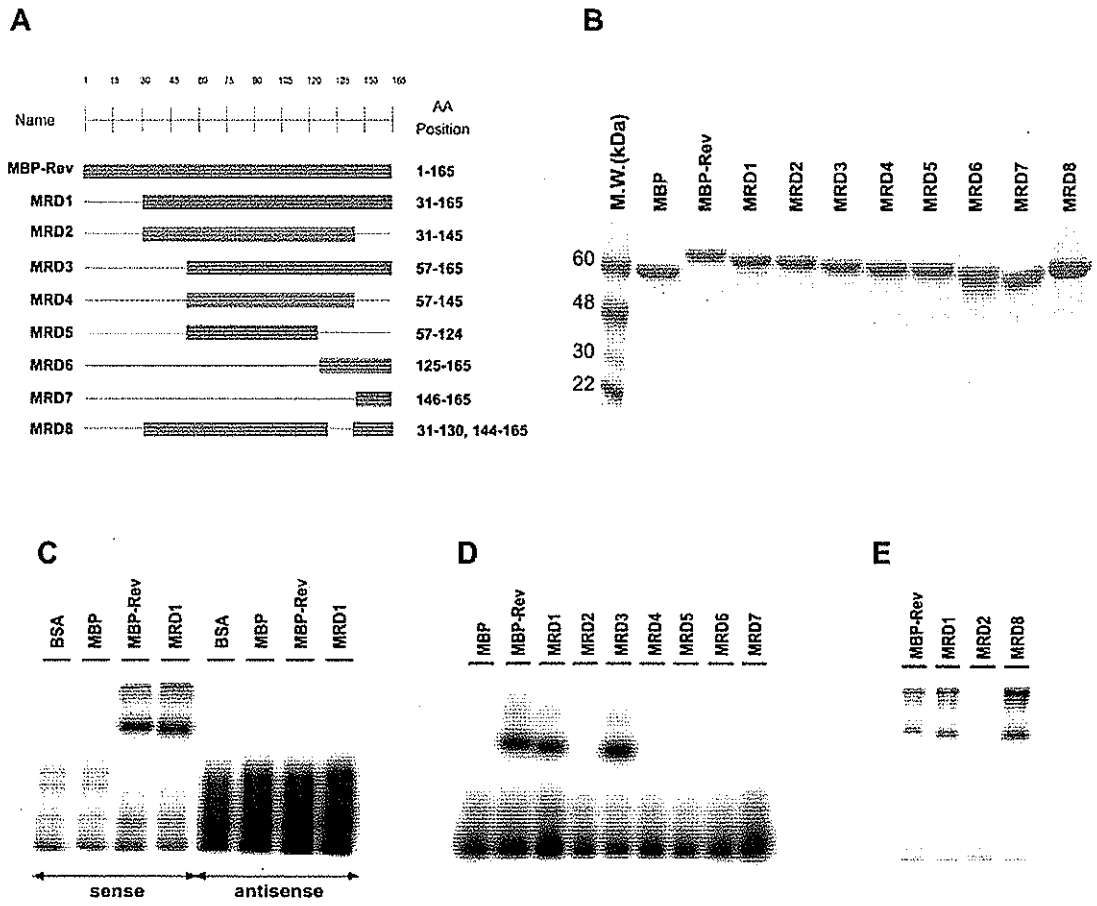


Figure 2.4. Identification of sequences critical for RNA-binding activity of EIAV Rev. **(A)** MBP-Rev deletion mutants (MRD) containing 5′ or 3′ deletions used to map regions of Rev required for RNA binding. The Rev amino acids retained in each construct are indicated. **(B)** Expression and purity of MRD mutants were assessed by Coomassie staining of SDS-PAGE gels. The molecular size of each deletion mutant was confirmed using molecular size (MW) markers. **(C)** RNA-binding activity of MBP-Rev fusion proteins was determined by UV cross-linking and SDS-PAGE. Rev fusion proteins were incubated with radiolabeled EIAV RRE (nucleotides 5443 to 5565), cross-linked with UV irradiation, and treated with RNase. RNA-protein complexes were separated by SDS-PAGE and quantified using a PersonalFX scanner and Quantity One software (Bio-Rad). Negative controls included bovine serum albumin (BSA), MBP, and antisense RRE. (D and E) RRE-binding activity of Rev deletion mutants as described for panel C.

Two noncontiguous ARMs are required for RNA binding and nuclear export activity.

Arginine-rich motifs are among the most common and well-characterized RNA-binding motifs (9) and are essential for RNA-binding activities of HIV Rev and Tat (24, 29, 41, 45). In HIV-1 Rev, a single 12-aa-long ARM located within the N-terminal domain functions in both RNA binding and nuclear localization (24). EIAV Rev contains two short Arg-rich motifs: RRDRW (aa 76 to 80) within the central region and KRRRK (aa 159 to 163) in the C-terminal region. Another motif within the central domain, ERLE (aa 93 to 96), was proposed as an RNA-binding motif based on studies showing that replacement of the motif with alanines abolished nuclear export and alternative splicing activity (23) and abrogated RNA binding in gel mobility shift assays (14). All three motifs are located within the regions we identified as essential for RNA-binding activity of EIAV Rev (Fig. 2.4). To directly test their contribution to RRE binding, we introduced alanine substitutions into each motif in the background of MBP Rev construct MRD1. The amount of RNA-protein complexes was greatly reduced with either MRD1-AADAA, where all of the residues except the Asp in the RRDRW motif were replaced by Ala, or MRD1-KAAAK, in which all Arg residues in the KRRRK motif were replaced by Ala (Fig. 2.5A). However, near wild-type levels of RNA-protein complexes were observed with MRD1-AALA, where the charged residues (Arg and Glu) in the ERLE motif were replaced by Ala. Thus, Arg residues in both the RRDRW and the KRRRK motif are essential for RNA binding. The finding that two separate, noncontiguous Arg-rich motifs are required for RNA binding indicates that EIAV Rev contains a bipartite RNA-binding domain, one portion of which overlaps the nuclear localization signal. This RNA-binding arrangement of EIAV Rev differs markedly from that of HIV-1 Rev.

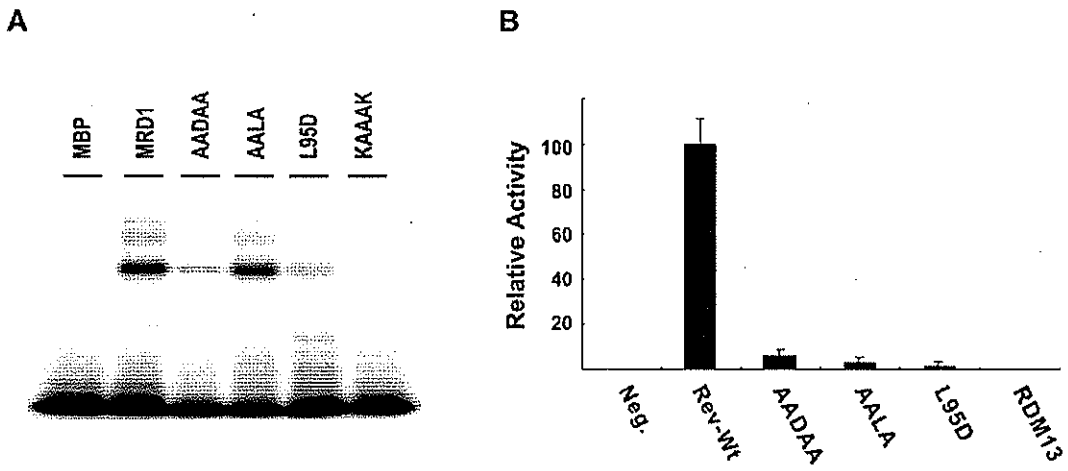


Figure 2.5. RRE-binding activity and functional analyses of Rev mutants. **(A)** Identification of specific motifs essential for RRE-binding activity of EIAV Rev. Point mutations introduced into the RRDRW, ERLE, and KRRRK motifs include alanine substitution of charged and/or Arg residues (AADAA, AALA, and KAAAK) and substitution of Asp for Leu in the ERLE motif (L95D). Binding activity was assessed as described in the legend to Fig. 2.4C. **(B)** Activity of Rev point mutation mutants (AADAA, AALA, KAAAK, and L95D) and deletion mutant (RDM13) were assayed as described in the legend to Fig. 2.2B. Neg., negative control. Wt, wild type.

Although others have reported a loss of RNA-binding activity due to alanine substitutions for the ERLE motif in the central domain of EIAV Rev (14), we observed no decrease in RNA binding when Ala was substituted for only the charged residues in the ERLE motif (MRD1-AALA). This suggested that the Leu residue at position 95 was critical for the RNA-binding activity. The predicted secondary structure of EIAV Rev, based on consensus results using five different prediction programs (16, 28, 29, 39, 42), places the ERLE motif in the middle of an {alpha}-helix, and it was previously suggested that the Ala substitutions may disrupt the {alpha}-helical structure (14). To explore this, we substituted an Asp for the Leu residue at position 95 (L95D). The Asp can act not only as a helix breaker

but also can disrupt hydrophobic interactions between helices. The L95D mutation resulted in an 80% reduction of the maximum binding of Rev to RRE (Fig. 2.5A). Therefore, the ERLE may play a role in stabilizing the protein structure required for RNA-binding activity.

To extend our analyses of the RRDRW and ERLE motifs, we introduced the Ala or Asp substitutions in R1A cDNA and analyzed the effect of mutations on Rev nuclear export activity in transient transfection assays (Fig. 2.5B). All of the RRDRW and ERLE mutants had significantly reduced Rev nuclear export activity. In addition, deletion of the KRRRK motif (RDM13), which is required for both RNA binding and nuclear export, eliminated Rev nuclear export activity in cell-based transfection assays. The results of these functional studies support our RNA-binding studies and establish a critical role for these motifs in Rev nuclear export activity. Interestingly, the AALA mutant was able to bind to viral RNA but was defective for nuclear export activity. Although its exact role is not known, this confirms previous findings that ERLE is critical for Rev activity (14, 23). More detailed analyses of Rev-RRE interactions will further enhance the results of our in vitro RNA-binding assays.

DISCUSSION

The Rev/Rex proteins of complex retroviruses are essential regulatory proteins that mediate nuclear export of incompletely spliced viral mRNAs via discrete functional domains that interact with cellular proteins and viral RNA. This family of proteins contains an interchangeable nuclear export signal that interacts with the cellular protein Crm1 and an ARM that binds specifically to homologous viral mRNA. Compared to most lentiviral Rev proteins, EIAV Rev is atypical with respect to the organization of functional domains (20, 23), the spacing of critical residues in the NES (23, 33), the use of a purine-rich, exonic splicing element as an RRE (8, 34), and the regulation of alternative splicing of the bicistronic *tat/rev* mRNA (34). In addition, EIAV Rev is highly variable in vivo, and we have

shown that changes in Rev phenotype correlate with changes in clinical stages of EIAV infection (4, 6, 7). To better understand the significance of Rev variation in vivo, we undertook more detailed analyses of the functional domains of EIAV Rev. Previous studies identified point mutations in EIAV Rev that reduced nuclear export (7, 20, 23, 33, 35), alternative splicing (7, 23), RNA binding (14), and nuclear localization activities (23). Here, we extend those studies to map functional domains and identify specific motifs that are essential for EIAV Rev activity. All domains essential for Rev activity are contained within the second exon of EIAV Rev (20, 23). The overall functional organization of EIAV Rev exon 2 includes the N-terminal nuclear export signal, a large central region that contains amino acids required for RNA binding, a nonessential region, and a C-terminal region required for both nuclear localization and RNA binding. Two short, noncontiguous ARMs are necessary for RNA binding: a central RRDRW motif and a C-terminal KRRRK motif that is also essential for nuclear localization. The bipartite ARM is unique among lentivirus Rev proteins and provides another example of how EIAV uses an unusual protein arrangement to carry out a common lentiviral function.

In HIV-1 Rev, the prototypical lentiviral Rev, a 17-aa ARM is located in the N-terminal half of the protein and serves as both a sequence-specific RNA-binding domain and the NLS (32, 37, 40). No typical ARM is found within EIAV Rev, and prior studies have not directly examined the RNA-binding domain of EIAV Rev. However, previous studies have identified several mutants defective in either nuclear localization or RNA binding (14, 23). Mutants containing alanine substitutions of 157DSKR160 (M15) or 161RR162 (M1) were defective in nuclear entry and showed reduced alternative splicing activity (23). Alanine substitutions of 76RRDR79 (M11) and 93ERLE96 (M27) resulted in loss of nuclear export activity (23), and the latter was defective in RNA binding and alternative splicing (14). Based on these results, it appeared that there was no overlap between the NLS and RNA-binding

domain and that the EIAV Rev RNA-binding domain was distinct from the ARMs characteristic of other complex retroviruses.

The present study is the first to directly characterize the RNA-binding activity of EIAV Rev and provides further insight into the specific motifs essential for EIAV Rev activity. The EIAV Rev NLS is relatively compact and requires the KRRRK motif contained within the C terminus of Rev. Mutants that retained four or five basic residues at amino acids 159 to 163 were able to translocate to the cell nucleus (C3 and C6), while mutants with less than four basic residues in this motif remained in the cytoplasm (C1, C2, C4, and C5). The KRRRK motif is similar to the arginine-rich NLSs of HIV-1 Rev and HTLV-1 Rex (32, 36, 40) and is nearly identical to the KRRR nuclear localization motifs found in the Drosophila melanogaster gcm gene product (1) and the RNA-binding human DEDD caspase protein (44). The ERLE and RRDR motifs identified in previous mutants M11 and M27 (23) were both found to be important in the RNA-binding activity of EIAV Rev, but they act through different mechanisms. Mutations of the charged residues in the ERLE motif had no effect on RNA binding, whereas the L95D mutation reduced binding by more than 80%. Therefore, the ERLE motif may play a role in stabilizing the protein structure required for binding of Rev to the viral RNA. In contrast, the arginine residues in both the RRDRW and KRRRK motifs were required for RNA binding, suggesting these short ARMs directly contact the RRE. Based on the results of our deletion analyses, other residues in the central domain likely contribute to RNA binding through direct contact with RNA and/or stabilization of protein structure.

Arginine rich motifs are short, arginine-containing regions of 10 to 20 amino acids that mediate RNA binding of a number of viral and ribosomal proteins (9). The arginine residues are thought to play two general roles in RNA binding: first, as a probe to search for a high-affinity binding site, and second, to form a network of specific hydrogen bonds with the RNA backbone and specific bases (9). In the present study, we found that RNA binding

of EIAV Rev required two short ARMs separated by 79 amino acids in the primary protein sequence. One motif is located in the central region, and the second overlaps with the NLS in the C terminus. The results of UV cross-linking experiments showed that the C-terminal 20 amino acids of EIAV Rev are required, but not sufficient, for binding to the RRE. Importantly, site-specific mutation of arginine residues in either the central ARM or the C-terminal ARM abolished RRE binding. These results strongly suggest that the two ARMs interact with the EIAV RRE in concert and thus comprise a bipartite RNA-binding domain. Coordinated action of two, noncontiguous ARMs was also found to be necessary for binding of hepatitis delta antigen to viral RNA (13). It is not clear how two ARMs interact to bind viral RNA. The two domains may be in close proximity within the three-dimensional structure of the folded protein, where they could form a single RNA-binding domain containing at least seven arginine residues. Alternately, the two ARMs could interact with different regions of the viral RNA. Further analyses of EIAV Rev-RRE interactions will provide insight into this potentially novel class of RNA-binding motifs.

In addition to promoting export of incompletely spliced viral RNA, EIAV Rev regulates alternative splicing of exon 3 in the bicistronic *tat/rev* mRNA (34). Exon 3 contains a purine-rich region that binds both Rev and the SR protein SF2/ASF and thus functions as both the EIAV RRE and an exonic splicing enhancer (7, 21, 31). Alternative splicing, or exon 3 skipping, likely results from competition between these two functions at either the RNA binding step or at a downstream step involved in spliceosomal assembly or activation (7, 14, 21, 31). Previous studies reported a loss in alternative splicing activity in Rev mutants containing alanine substitution of the ERLE motif or in both the RRDR and KRRRK motifs (14). Mutational analyses of cis-acting sequences required for exon 3 skipping show a close correlation between alternative splicing activity and RNA-binding activity (8, 14). Together, these data suggest that RNA binding plays an essential role in alternative splicing. Studies to date, however, have been unable to identify an alternative splicing domain distinct from other

Rev functional domains (23). Therefore, it has been difficult to assess the biological significance of alternative splicing in vivo. Exon 3 skipping was originally proposed as a novel mechanism for autoregulation of Rev (34); however, mRNAs lacking exon 3 comprise only a small percentage of viral mRNA in Rev-expressing cells (31). At present, the primary function of EIAV Rev appears similar to other lentiviral Revs in regulating the shift from production of multiply spliced to incompletely spliced viral mRNAs.

There is a relatively high rate of genetic variation in the region of the EIAV genome where the second exon of Rev overlaps the cytoplasmic portion of the transmembrane protein gp45 (2, 6, 7, 30). The NLS and RNA-binding motifs are highly conserved, while the region identified here as nonessential for Rev function is the site of a number of amino acid changes that significantly altered Rev phenotypes (6). The nonessential region spans a predicted {alpha}-helix, and our results indicate that deletion of this region does not alter the structure of the Rev in a way that compromises Rev function. Similarly, the overlapping region of TM is found within the cytoplasmic domain, which may also be able to accommodate some degree of genetic variation without loss of function. As such, this region of the viral genome may be able to tolerate genetic mutations in Rev and/or TM that confer a selective advantage in vivo. Future studies may reveal how genetic variation in Rev contributes to viral persistence while preserving the function of this critical lentiviral protein.

ACKNOWLEDGEMENT

We thank Kai-Ming Ho and Yungok Ihm for helpful discussions and Pamela Bruellman and Sue Pritchard for excellent technical assistance. The MBP vector was kindly provided by Jamie Williamson. This work was supported in part by funding from the National Institutes of Health grant CA97936 and the National Research Initiative of the

USDA Cooperative State Research, Education, and Extension Service grant number 2002-35204-12699.

REFERENCES

1. **Akiyama, Y., T. Hosoya, A. M. Poole, and Y. Hotta.** 1996. The gcm-motif: a novel DNA-binding motif conserved in Drosophila and mammals. Proc. Natl. Acad. Sci. USA **93:**14912-14916.

2. **Alexandersen, S., and S. Carpenter.** 1991. Characterization of variable regions in the envelope and S3 open reading frame of equine infectious anemia virus. J. Virol. **65:**4255-4262.

3. **Ali, S. A., and A. Steinkasserer.** 1995. PCR-ligation-PCR mutagenesis: a protocol for creating gene fusions and mutations. BioTechniques **18:**746-750.

4. **Baccam, P., R. J. Thompson, Y. Li, W. O. Sparks, M. Belshan, K. S. Dorman, Y. Wannemuehler, J. L. Oaks, J. L. Cornette, and S. Carpenter.** 2003. Subpopulations of equine infectious anemia virus Rev coexist in vivo and differ in phenotype. J. Virol. **77:**12122-12131.

5. **Beisel, C. E., J. F. Edwards, L. L. Dunn, and N. R. Rice.** 1993. Analysis of multiple mRNAs from pathogenic equine infectious anemia virus (EIAV) in an acutely infected horse reveals a novel protein, ttm, derived from the carboxy terminus of the EIAV transmembrane protein. J. Virol. **67:**832-842.

6. **Belshan, M., P. Baccam, J. L. Oaks, B. A. Sponseller, S. C. Murphy, J. Cornette, and S. Carpenter.** 2001. Genetic and biological variation in equine infectious anemia virus Rev correlates with variable stages of clinical disease in an experimentally infected pony. Virology **279:**185-200.

7. **Belshan, M., M. E. Harris, A. E. Shoemaker, T. J. Hope, and S. Carpenter.** 1998. Biological characterization of Rev variation in equine infectious anemia virus. J. Virol. **72:**4421-4426.

8. **Belshan, M., G. S. Park, P. Bilodeau, C. M. Stoltzfus, and S. Carpenter.** 2000. Binding of equine infectious anemia virus Rev to an exon splicing enhancer mediates alternative splicing and nuclear export of viral mRNAs. Mol. Cell. Biol. **20:**3550-3557.

9. **Burd, C. G., and G. Dreyfuss.** 1994. Conserved structures and diversity of functions of RNA-binding proteins. Science **265:**615-621.

10. **Carpenter, S., S. Alexandersen, M. J. Long, S. Perryman, and B. Chesebro.** 1991. Identification of a hypervariable region in the long terminal repeat of equine infectious anemia virus. J. Virol. **65:**1605-1610.

11. **Carpenter, S., and B. Chesebro.** 1989. Change in host cell tropism associated with in vitro replication of equine infectious anemia virus. J. Virol. **63:**2492-2496.

12. **Carroll, R., and D. Derse.** 1993. Translation of equine infectious anemia virus bicistronic tat-rev mRNA requires leaky ribosome scanning of the tat CTG initiation codon. J. Virol. **67:**1433-1440.

13. **Chou, H.-C., T.-Y. Hsieh, G.-T. Sheu, and M. M. C. Lai.** 1998. Hepatitis delta antigen mediates the nuclear import of hepatitis delta virus RNA. J. Virol. **72:**3684-3690.

14. **Chung, H.-K., and D. Derse.** 2001. Binding sites for Rev and ASF/SF2 map to a 55-nucleotide purine-rich exonic element in equine infectious anemia virus RNA. J. Biol. Chem. **276:**18960-18967.

15. **Cook, K., G. Sue, J. Fisk, J. Hauber, N. Usman, T. J. Daly, and J. R. Rusche.** 1991. Characterization of HIV-1 REV protein: binding stoichiometry and minimal RNA substrate. Nucleic Acids Res. **19:**1577-1583.

16. **Cuff, J. A., M. E. Clamp, A. S. Siddiqui, M. Finlay, and G. J. Barton.** 1998. JPred: a consensus secondary structure prediction server. Bioinformatics **14:**892-893.

17. **Cullen, B. R.** 1992. Mechanism of action of regulatory proteins encoded by complex retroviruses. Microbiol. Rev. **56:**375-394.

18. **Fischer, U., J. Huber, W. C. Boelens, I. W. Mattal, and R. Luhrmann.** 1995. The HIV-1 Rev activation domain is a nuclear export signal that accesses an export pathway used by specific cellular RNAs. Cell **82:**475-483.

19. **Fridell, R. A., H. P. Bogerd, and B. R. Cullen.** 1996. Nuclear export of late HIV-1 mRNAs occurs via a cellular protein export pathway. Proc. Natl. Acad. Sci. USA **93:**4421-4424.

20. **Fridell, R. A., K. M. Partin, S. Carpenter, and B. R. Cullen.** 1993. Identification of the activation domain of equine infectious anemia virus Rev. J. Virol. **67:**7317-7323.

21. **Gontarek, R. R., and D. Derse.** 1996. Interactions among SR proteins, an exonic splicing enhancer, and a lentivirus Rev protein regulate alternative splicing. Mol. Cell. Biol. **16:**2325-2331.

22. **Gorlich, D., and I. W. Mattaj.** 1996. Nucleocytoplasmic transport. Science **271:**1513-1518.

23. **Harris, M. E., R. R. Gontarek, D. Derse, and T. J. Hope.** 1998. Differential requirements for alternative splicing and nuclear export functions of equine infectious anemia virus Rev protein. Mol. Cell. Biol. **18:**3889-3899.

24. **Hope, T. J., D. McDonald, X. Huang, J. Low, and T. G. Parslow.** 1990. Mutational analysis of the human immunodeficiency virus type 1 Rev transactivator: essential residues near the amino terminus. J. Virol. **64:**5360-5366.

25. **Izaurralde, E., and S. Adam.** 1998. Transport of macromolecules between the nucleus and the cytoplasm. RNA **4:**351-364.

26. **Jones, D. T.** 1999. Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. **292:**195-202.

27. **Kawakami, T., L. Sherman, J. Dahlbert, A. Gazit, A. Yaniv, S. R. Tronick, and S. A. Aaronson.** 1987. Nucleotide sequence analysis of equine infectious anemia proviral DNA. Virology **158:**300-312.

28. **Kloczkowski, A., K. L. Ting, R. L. Jernigan, and J. Garnier.** 2002. Combining the GOR V algorithm with evolutionary information for protein secondary structure prediction from amino acid sequence. Proteins **49:**154-166.

29. **Laspia, M. F., A. P. Rice, and M. B. Mathews.** 1989. HIV-1 Tat protein increases transcriptional initiation and stabilizes elongation. Cell **59:**283-292.

30. **Leroux, C., C. J. Issel, and R. C. Montelaro.** 1997. Novel and dynamic evolution of equine infectious anemia virus genomic quasispecies associated with sequential disease cycles in an experimentally infected pony. J. Virol. **71:**9627-9639.

31. **Liao, H.-J., C. C. Baker, G. L. Princler, and D. Derse.** 2004. Cis-acting and trans-acting modulation of equine infectious anemia virus alternative RNA splicing. Virology **323:**131-140.

32. **Malim, M. H., S. Böhnlein, J. Hauber, and B. R. Cullen.** 1989. Functional dissection of the HIV-1 Rev trans-activator - derivation of a trans-dominant repressor of Rev function. Cell **58:**205-214.

33. **Mancuso, V. A., T. J. Hope, L. Zhu, D. Derse, T. Phillips, and T. G. Parslow.** 1994. Posttranscriptional effector domains in the Rev proteins of feline immunodeficiency virus and equine infectious anemia virus. J. Virol. **68:**1998-2001.

34. **Martarano, L., R. Stephens, N. Rice, and D. Derse.** 1994. Equine infectious anemia virus trans-regulatory protein Rev controls viral mRNA stability, accumulation, and alternative splicing. J. Virol. **68:**3102-3111.

35. **Meyer, B. E., J. L. Meinkoth, and M. H. Malim.** 1996. Nuclear transport of human immunodeficiency virus type 1, visna virus, and equine infectious anemia virus Rev proteins: identification of a family of transferable nuclear export signals. J. Virol. **70:**2350-2359.

36. **Nakielny, S., and G. Dreyfuss.** 1999. Transport of proteins and RNAs in and out of the nucleus. Cell **99:**677-690.

37. **Olsen, H., A. Cochrane, P. Dillon, C. Nalin, and C. Rosen.** 1990. Interaction of the human immunodeficiency virus type 1 rev protein with a structured region in env mRNA is dependent on multimer formation mediated through a basic stretch of amino acids. Genes Dev. **4:**1357-1364.

38. **Otero, G. C., M. E. Harris, J. E. Donello, and T. J. Hope.** 1998. Leptomycin B inhibits equine infectious anemia virus Rev and feline immunodeficiency virus Rev function but not the function of the hepatitis B virus postranscriptional regulatory element. J. Virol. **72:**7593-7597.

39. **Ouali, M., and R. D. King.** 2000. Cascaded multiple classifiers for secondary structure prediction. Protein Sci. **9:**1162-1176.

40. **Perkins, A., A. W. Cochrane, S. M. Ruben, and C. A. Rosen.** 1989. Structural and functional characterization of the human immunodeficiency virus rev protein. J. Acquir. Immune Defic. Syndr. **2:**256-263.

41. **Pollard, V. W. 1998. The HIV-1 Rev protein. Annu. Rev. Microbiol.** 52:491-532.

42. **Rost, B., and C. Sander.** 1993. Secondary structure prediction of all-helical proteins in two states. Protein Eng. **6:**831-836.

43. **Ryder, S. P., L. A. Frater, D. L. Abramovitz, E. B. Goodwin, and J. R. Williamson.** 2004. RNA target specificity of the STAR/GSG domain post-transcriptional regulatory protein GLD-1. Nat. Struct. Mol. Biol. **11:**20-28.

44. **Stegh, A. H., O. Schickling, A. Ehret, C. Scaffidi, C. Peterhansel, T. G. Hofmann, I. Grummt, P. H. Krammer, and M. E. Peter.** 1998. DEDD, a novel death effector domain-containing protein, targeted to the nucleolus. EMBO J. **17:**5974-5986.

45. **Weiss, M. A., and N. Narayana.** 1998. RNA recognition by arginine-rich peptide motifs. Biopolymers **48:**167-180.

46. **Whittaker, G., and A. Helenius.** 1998. Nuclear import and export of viruses and virus genomes. Virology **246:**1-23.

47. **Zapp, M., T. Hope, T. Parslow, and M. Green.** 1988. Oligomerization and RNA binding domains of the type 1 human immunodeficiency virus rev protein: a dual function for an arginine-rich motif. Proc. Natl. Acad. Sci. USA **88:**7734-7738.

48. **Zapp, M. L., and M. R. Green.** 1989. Sequence-specific RNA binding by the HIV-1 Rev protein. Nature **342:**714-716.

CHAPTER 3. A CONSERVED RNA STRUCTURAL MOTIF IS REQUIRED FOR HIGH AFFINITY REV BINDING IN BOTH EIAV AND HIV-1

A paper to be submitted in *Molecular and Cellular Biology*

Jae-Hyung Lee, Gloria Culver, Susan Carpenter, and Drena Dobbs

ABSTRACT

A *cis*-acting RNA regulatory element, the Rev-responsive element (RRE), is essential for regulation of gene expression and genomic replication in lentiviruses, including human immunodeficiency virus (HIV-1) and equine infection anemia virus (EIAV). Despite its potential as a clinical target, little is known about the detailed molecular structure and mechanisms of RRE function. In this study, we investigate the secondary structure of the EIAV RRE and its interaction with the EIAV Rev protein, a critical *trans*-acting factor that effects several key RRE-mediated functions. A combination of computational prediction and detailed chemical probing and footprinting experiments were used to determine the RNA secondary structure of EIAV RRE-1, a 555 nt region previously shown to function as the EIAV RRE in vivo. Chemical probing experiments confirmed the presence of several predicted loop and stem-loop structures, which are conserved among 140 EIAV sequence variants. Footprinting experiments revealed that Rev binding induces significant structural rearrangement in two conserved domains characterized by stable stem-loop structures. Rev binding region-1 (RBR-1) corresponds to a genetically-defined Rev binding region that overlaps exon 1 of the EIAV Rev gene and contains an Exonic Slicing Enhancer (ESE). RBR-2, characterized for the first time in this study, is required for high affinity binding of

EIAV Rev to ERRE-1. RBR-2 contains a conserved RNA structural motif that is also found within the high affinity Rev binding site in HIV-1 RRE (stem-loop IIB), and within or near RRE regions of four additional lentiviruses. Taken together, the experimentally-validated RNA secondary structure of the EIAV RRE, and the discovery that high affinity Rev binding sites within the RREs of HIV-1 and EIAV share an RNA structural motif also found in at least six phylogenetically-diverse lentiviruses, provide the basis for detailed comparative analyses that should provide additional insight into the molecular mechanisms of Rev-RRE interactions in HIV-1 and other lentiviruses.

INTRODUCTION

Retroviruses employ a variety of mechanisms to express differentially spliced viral mRNAs transcribed from a single promoter. In all retroviruses, the presence of suboptimal splice sites allows for expression of several mRNAs from a single pre-RNA. Regulation of splice-site selection can be further regulated by *cis*-acting RNA sequences that either enhance or repress recognition of a splice site by the cellular splicing factors. In some retroviruses, viral pre-mRNAs contain constitutive transport elements (CTE) that are recognized by cellular proteins to facilitate nuclear export of incompletely spliced viral mRNAs. Other retroviruses encode Rev/Rex proteins that act in *trans* to regulate nuclear export of unspliced or incompletely spliced mRNAs required for expression of structural and enzymatic proteins as well as progeny viral RNA genomes (10). The Rev/Rex RNA export pathway has been best characterized in HIV-1 (24). After entering the nucleus, the HIV-1 Rev protein binds to a specific *cis*-acting element, termed the Rev response element (RRE), within the viral pre-mRNA (9, 68), multimerizes (50, 69) and then facilitates nuclear export of incompletely spliced viral mRNA via the Crm1 nuclear export pathway (17, 18). Discrete functional domains within Rev mediate nuclear localization, RRE binding and multimerization, and

nuclear export. The HIV-1 RRE is a highly structured RNA located at the junction between the SU (gp120) and TM (gp41) domains of the *env* gene (12, 32, 68). Biochemical and biophysical experiments have implicated specific stem-loop structures in Rev binding and multimerization (2, 30, 38). Because the Rev-dependent export pathway plays an essential role in HIV-1 replication, disruption of the Rev-RRE interactions is an attractive target for design of effective antiviral therapies.

All lentiviruses utilize the Rev-dependent, Crm-1-mediated export pathway for expression of incompletely spiced mRNAs. There is no conservation among the lentiviral RREs at the RNA sequence level; however, the RRE regions of several primate and non-primate lentiviral genomes map to the SU/TM junction in *env* gene (12) (32, 55, 64). In addition, computational analyses suggest that several of the lentiviral RREs may share RNA secondary structural elements (34). Equine infectious anemia virus (EIAV) is one of the most divergent members of the subfamily (16). EIAV Rev (E-Rev) is functionally homologous with other lentivirus Revs and utilizes the Crm-1 pathway for export of incompletely spliced mRNAs; yet EIAV differs from most lentiviruses in structural and functional features of Rev and RRE. EIAV therefore offers an opportunity for comparative analysis of the molecular interactions important in regulation of lentiviral gene expression. This approach that may identify highly conserved RNA-protein interactions that could be targeted in novel anti-lentiviral therapies.

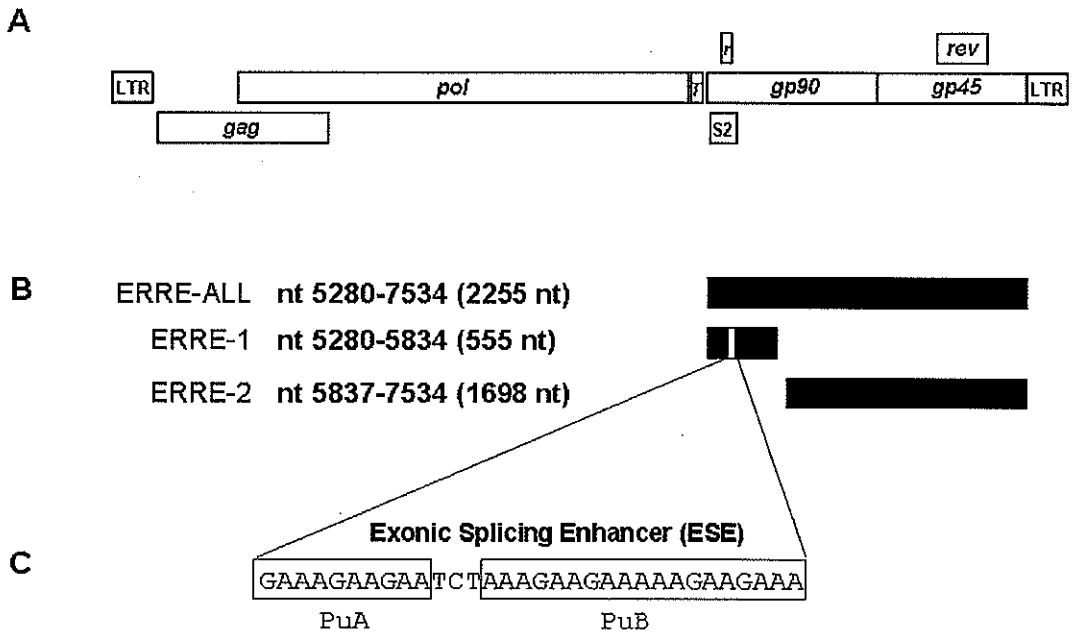


Figure 3.1. Organization of EIAV genome and location of EIAV RRE sequences. **(A)** Schematic view of the EIAV genome showing the locations of open reading frames. **(B)** Location of RRE regions in the EIAV genome. ERRE-All, ERRE-1 and ERRE-2 refer to regions defined by Belshan et al. (4), with numbering according to Kawakami et al. (29). **(C)** The exonic splicing enhancer (ESE) sequence in ERRE-1. Boxed sequences represent two purine-rich sequence stretches (PuA and PuB) previously reported to interact with both EIAV Rev and SF2-ASF (5, 8).

EIAV Rev is a 165 amino acid protein translated from exons 3 and 4 of a multiply spliced, four-exon, bicistronic mRNA that also encodes the *trans*-activating protein, Tat (Fig. 3.1). The leucine-rich nuclear export signal (NES) in E-Rev is similar to other viral and cellular export proteins that interact with the Crm1, but is atypical in the spacing of the leucine residues within the NES (18). The E-Rev RNA-binding domain is bipartite, comprised of two short arginine-rich motifs (ARMs) separated by 79 amino acids in the

primary sequence (33). It not clear how the two domains cooperate to form a complex with the RRE, but a theoretical structural model of the E-Rev protein places the ARMs in close proximity within the three dimensional structure, suggesting they could form a single RNA binding interface within the Rev-RRE complex (25, 33). In addition to promoting nuclear export of incompletely spliced RNA, E-Rev also regulates inclusion of exon 3 in the multiply spliced, bicistronic RNA: in the presence of Rev, exon 3 is skipped, resulting in a three-exon, monocistronic mRNA encoding only Tat (5, 21). Exon 3 is flanked by a suboptimal splice acceptor and contains a purine rich, exonic splicing enhancer (ESE) required for exon three inclusion (35). ESEs typically are purine rich sequences embedded within alternatively spliced exons that bind cellular SR proteins and recruit essential splicing factors to suboptimal splice sites, resulting in exon inclusion of alternatively spliced exons. It is thought that Rev-mediated skipping of exon 3 is a consequence of either Rev-SR protein, or Rev-RNA interactions that disrupt ESE-SR protein interactions (5, 8, 19, 35).

The EIAV RRE (ERRE) differs from other lentiviral RREs with respect to location and function. The ERRE is located in a 555 nt region near the 5′ end of *env*, which spans exon 3 of the bicistronic Tat-Rev mRNA (4, 5, 19, 39). A 57 nt sequence encompassing the ESE within exon 3 was shown to bind GST-Rev and to act as a functional RRE in a heterologous nuclear export assay system (5); however, nuclear export activity of the 57 nt "minimal ERRE" was reduced as compared to the full-length ERRE (4). Mutational analyses of ERRE demonstrated that the purine rich sequences within exon 3 function as both an ESE and an RRE (4, 8, 35). The ERRE thus plays an important role in the complex interactions between viral pre-mRNAs, the viral Rev protein, and cellular splicing factors to mediate alternative splicing and regulation of viral gene expression (3, 8, 19, 35).

In this paper, we investigate the RNA structure of the EIAV RRE and its interactions with the Rev protein. We propose an RNA secondary structure model for the essential RRE in EIAV, based on a combination of secondary structure prediction and chemical probing

experiments. We present the first detailed *in vitro* footprinting analysis of EIAV Rev-RRE complexes, and identify two distinct domains within the essential RRE that undergo significant structural transitions upon Rev binding. Computational analyses revealed an RNA secondary structural motif within the high affinity Rev-binding sites of both HIV-1 and EIAV that is present within the mapped RREs of four additional lentiviruses. The identification of a conserved recognition element for lentiviral Rev-RRE interactions lays the groundwork for more detailed comparative analyses of lentiviral Rev-RRE interactions.

MATERIALS AND METHODS

Preparation of ERRE-1 RNA and purified EIAV Rev protein

The EIAV ERRE-1 (corresponding nts 5280 to 5843 of our "standard" wildtype EIAV strain, MA-1, GenBank accession no. M58039) was amplified by PCR from pERRE-All (4, 5, 29) using a 5′ primer containing a T7 promoter site. The PCR product was purified using QIAquick PCR purification columns (QIAGEN, Valencia, CA), and RNA was generated by *in vitro* transcription (T7-MEGAscript; Ambion, Austin, TX). Transcribed RNA was purified using MEGAClear kits (Ambion, Austin, TX), denatured at 90°C for 2 min and annealed by slow cooling. Ethanol precipitation was performed to remove salts and concentrate ERRE-1 RNA. Concentrated RNA was stored at -80℃. EIAV Rev protein was expressed as an MBP-ERev fusion protein and purified as described previously (33).

Chemical probing of RRE RNA secondary structure and footprinting Rev-RRE complexes

Prior to chemical probing or footprinting experiments, ERRE-1 in RNA storage buffer (10mM Tris-HCl, pH 7.5) was pre-incubated at 42℃ for 15 min. To generate

unmodified and modified unbound ERRE-1 samples for RNA secondary structure probing experiments, two aliquots (each containing 35 pmol) of ERRE-1 were added to RNA binding buffer (10 mM HEPES-KOH, pH 7.5, 100 mM KCl, 1 mM $MgCl_2$, and 0.5 mM EDTA). To generate ERev-ERRE-1 protein-RNA complexes for footprinting experiments, samples containing one to thirty-fold molar excess of purified MBP-ERev fusion protein were incubated with pre-folded ERRE-1 RNA (35pmol) in binding buffer in a total volume of 87.5 μℓ for 20 min on ice. Samples containing folded ERRE-1 RNA alone or RNA-protein complexes were incubated with 10.5 μℓ 880 mM dimethylsulfate (DMS) (Sigma-Aldrich, St. Louis, MO) or 7 μℓ 720 mM kethoxal (ICN, Costa Mesa, CA) or 16 μℓ hydroxyl radical probing mixture (4 μℓ of 50mM $Fe(NH_4)_2(SO_4)_2 \cdot 6H_2O$, 4 μℓ of 100mM EDTA, 4 μℓ of 250mM ascorbic acid, and 4 μℓ of 2.5% hydrogen peroxide) for 10 min at room temperature. DMS modification reactions were stopped by addition of 59.3 μℓ of DMS stop buffer (1M Tris-HCl pH 7.5, 0.1M EDTA pH 8.0, and 1M 2-mercaptoethanol). For the kethoxal probing, 8.2 μℓ of 0.5 M potassium borate was added for stabilizing kethoxal. Hydroxyl radical probing reactions were quenched by the addition of 92.8 μℓ of 1M thiourea. After ethanol precipitation of RNA or RNA-protein complexes, 3 phenol and 2 chloroform extractions were performed to purify RNA from RNA-protein complexes. RNA was concentrated and washed using 100% and 70% ethanol precipitations. Finally, RNA was resuspended in 35 μℓ of water (for DMS and hydroxyl radical modifications) or 35 μℓ of 40mM potassium borate for kethoxal modification. Additional details are provided in (47).

Primer extension analysis of chemically-modified EIAV RRE sequences

To identify positions of chemically modified nucleotides in ERRE-1, primer extension analysis was used using 5 different oligonucleotide primers (Table 3.1), essentially as previously described (47, 48).

Nitrocellulose filter binding assays

Nitrocellulose filter binding assays were carried out using purified ^{32}P-labelled RNAs corresponding to ERRE-1 (555 nt; 5280 to 5834) or subfragment of ERRE-1 (123 nt; 5443 to 5465) using standard procedures. Binding affinities were calculated using Dynafit software (31).

Sequences of gp90 (SU) variants and information content analysis

The EIAV genomic RNA sequences used in the present study were originally collected for analyzing sequence variants in the EIAV Env SU protein by Mealey, et al. (46) and deposited in the NCBI GenBank in two segments (5′ and 3′). From a total of 284 EIAV sequences used in Mealey, et al., those that overlapped the ERRE-1 region were collected. After removal of sequences with deletions or premature stop codons, a total of 258 sequences (126 corresponding to the 5′ fragment and 132 corresponding to the 3′ fragment of SU), corresponding to 139 complete Env gene variant sequences, remained. The concept and methods for analyzing of information content are described elsewhere (7, 51, 59). Briefly, collections of the 5′ or 3′ fragment sequences were aligned using CLUSTALW (http://www.ebi.ac.uk/clustalw/) and information content was calculated according to the following equation:

$$I_i = \sum_j q_{ij} \log_2 \frac{q_{ij}}{p_j}$$

where I_i is the information content for the nucleotide position i in the alignment, the index j sums over all possible nucleotides (A, T, G, and C), q_{ij} represents the observed frequency of

nucleotide j at position i and p_j represents the expected frequency value, which is 0.25. The calculated information content at each nucleotide position was plotted using Microsoft Excel.

RNA secondary structure prediction

Different methods for RNA secondary structure prediction have been reviewed recently (40, 44). In this work, Mfold (70) was used to predict the lowest free energy secondary structure of ERRE-1, (http://frontend.bioinfo.rpi.edu/applications/mfold/cgi-bin/rna-form1.cgi) using only the standard EIAV sequence as input (i.e., with no experimental constraints). To model the secondary structure of ERRE-1, using a single sequence of ERRE-1 and experimental constraints derived from the results of chemical probing experiments, four different methods, Mfold, Sfold (13) (http://sfold.wadsworth.org), RNAfold (23) (http://www.tbi.univie.ac.at/~ivo/RNA/), and RNAStructure (41)(http://rna.urmc.rochester.edu/rnastructure.html) were used. Experimental constraints data were preprocessed to generate appropriately formatted input files for each of the four programs. A fifth method, RNAalifold, generates an optimal RNA secondary structure based on calculation of the minimum free energy structure, and a partition function and base-pairing probability matrix derived from an multiple sequence alignment (22) (http://www.tbi.univie.ac.at/~ivo/RNA/). Input for RNAalifold consisted of an alignment of 140 ERRE-1 variant sequences generated using CLUSTALW (63) (http://www.ebi.ac.uk/clustalw/). The resulting alignment and experimental constraint input was used to determine an optimal RNA secondary structure for ERRE-1. RNA secondary structures were drawn using PSEUDOVIEWER2 (20) (http://wilab.inha.ac.kr/pseudoviewer2/).

Identification of a conserved RNA structural motif within lentiviral genomes

RNAstructure Dynalign software (43) (http://rna.urmc.rochester.edu/rnastructure.html) was used to test whether sequences corresponding to the high affinity Rev binding sites of EIAV and HIV-1 have the capacity to form similar RNA secondary structures. Sequences of 100 nt regions encompassing the HIV-1 stem-loop IIB region (nt 8-107) and the EIAV RBR-2 region (nt 371-470) were compared, resulting in two very similar ensembles of predicted secondary structures. RNA motif models based on common features of the HIV-1 and EIAV RRE structures were generated using RNAMotif (37) (http://www.scripps.edu/mb/case/casegr-sh-3.5.html) and used to scan the complete genomic sequences of ten different lentiviruses (GenBank accession no. M15654;HIV-1, NC_001450; EIAV, NC_001722; HIV-2, NC_001549; SIV, NC_001452; VV, NC_001463; CAEV, NC_001511; OLV, NC_001482; FIV, NC_001413; BIV, NC_001654; JDV) and rabbit endogenous lentivirus type K (RELIK) (28).

RESULT

Probing the RNA secondary structure of ERRE-1

To identify structural features of the EIAV genomic RNA essential for Rev responsiveness, we focused on a 555 nucleotide region (nt 5280-5834), referred to as ERRE-1 (Fig. 3.1B) (4). Initial mapping of *cis*-acting RRE regions in the EIAV genome had identified two elements, one near each end of the Env gene, that provide partial RRE function (39), and subsequent studies demonstrated that several distinct regions of the EIAV genome can contribute to RRE function (4, 19). The 555 nt ERRE-1 region was shown to function as a "minimal" RRE, retaining 60% of wildtype function *in vivo,* by Belshan, et al. (4). ERRE-1 encompasses exon 3 of the bicistronic that encodes both Env and Rev. It also contains an ESE (Fig. 3.1C) for Rev-mediated alternative mRNA splicing (5, 19), and is

required for Rev-mediated nuclear export of partially spliced mRNAs (4, 5). Both EIAV Rev and the cellular protein, SF2/ASF, a member of the SR family of splicing factors, have been shown to interact with the ESE within ERRE-1 (5, 8, 33). Thus, the ERRE-1 sequence was chosen for RNA secondary structure analysis and as substrate for ERev-RRE footprinting experiments.

The secondary structure of ERRE-1 was analyzed using a combination of computational and experimental approaches. First, the lowest free energy structure of ERRE-1 was predicted by computational method (Fig. 3.2A). The structure was generated by Mfold (70) with default parameters. Also we experimentally interrogated the structure by analyzing the accessibility of ribonucleotides in the folded RNA to single-strand specific chemical probes (kethoxal and DMS), thus identifying regions not involved in base-pairing (14). Chemically-modified nucleotides in ERRE-1 were identified by primer extension analysis using 5 different primers (Table 3.1). Experimental data from chemical probing experiments were integrated into computational predictions of secondary structure of the ERRE-1 using several different algorithms designed to incorporate experimental constraints from chemical probing assays.

Primers	Nucleotide sequence
Primer 90	TTTTTCTGACTGTTGGG
Primer 185	TCTTGGTCTCTTGCTTC
Primer 291	CCAAAGTATTCCTCCAG
Primer 389	CCCCAGCATTCTATAGC
Primer 485	GCTTCTAATAATGTAGC
Primer 555	TCCCCAATATTCCGCTGTGT

Table 3.1. A primer list for the primer extension analysis for ERRE-1

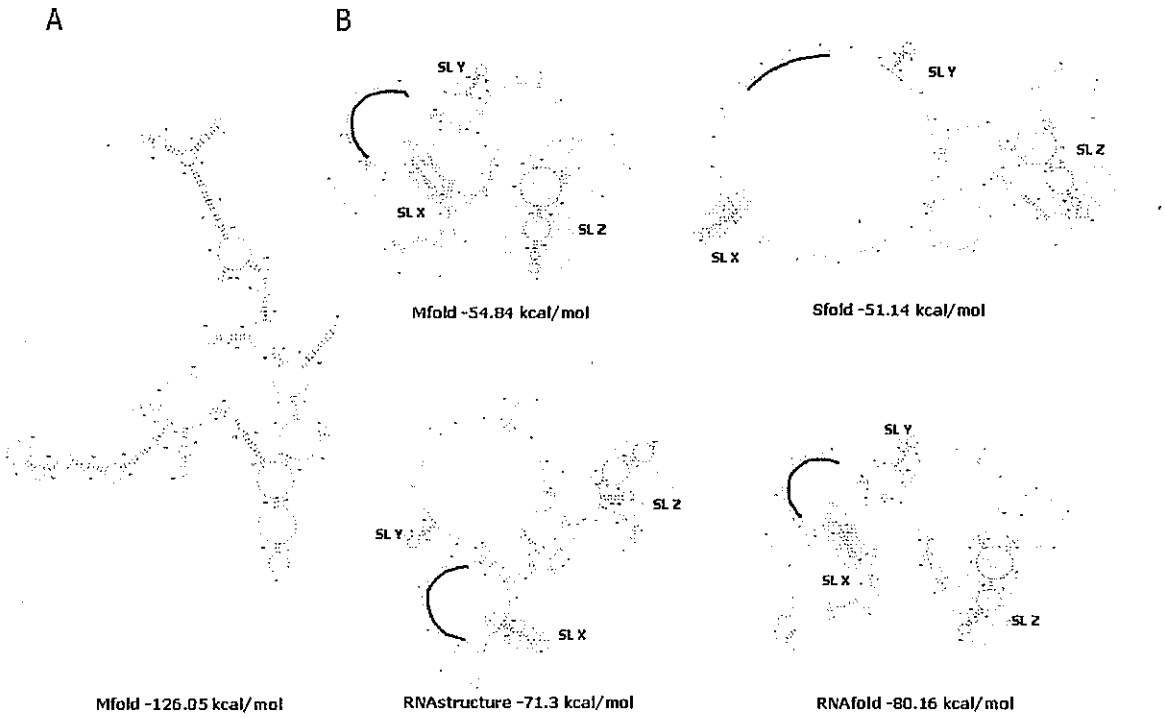


Figure 3.2. RNA secondary structural models for ERRE-1 **(A)** The lowest free energy secondary structure of ERRE-1, predicted by Mfold without incorporatating experimental constraints (70). **(B)** Lowest free energy RNA secondary structure models of ERRE-1, generated by for different algorithms, all using chemical probing results as experimental constraint input: Mfold, Sfold (13), RNAStructure (41), and RNAfold (23). SL-X, -Y and -Z, stem-loop structures common to all four models, are highlighted (see Materials and Methods for details).

In Figure 3.2B, four different secondary structure models were generated by Mfold (70), Sfold (13), RNAstructure (42) and RNAfold (23), with experimental constraints. Although the latter four models differ in detail, they are in similar in the overall topology: all four models share a set of 3 stem-loop structures (SL-X, -Y, and -Z) and in every case, the ESE is

located within a large loop (heavy line in Fig. 3.2B). Notably, with the exception of SL-X, none of these shared features is also found in the structure generated without experimental constraints (Fig. 3.2A).

To further refine and validate the RRE secondary structure models, we used available sequence information from EIAV variants to perform covariation analyses, using RNAalifold software (22). RNAalifold can incorporate covariation information from a collection of aligned RNA sequences, in addition to experimental constraint data, as input. Thus, the algorithm determines a consensus RNA secondary structure based on thermodynamic considerations and the evaluates the compatibility of observed sequence covariations with that secondary structure. Experimental constraints are used in the final step to identify an optimized RNA secondary structure. A total of 140 different EIAV sequences (the standard or "wildtype" sequence from strain MA-1 strain and 139 SU variant sequences (46) that overlap ERRE-1) were used to generate a multiple sequence alignment for computing ribonucleotide covariation frequencies within ERRE-1. The secondary structure generated by RNAalifold is shown in Figure 3.3.

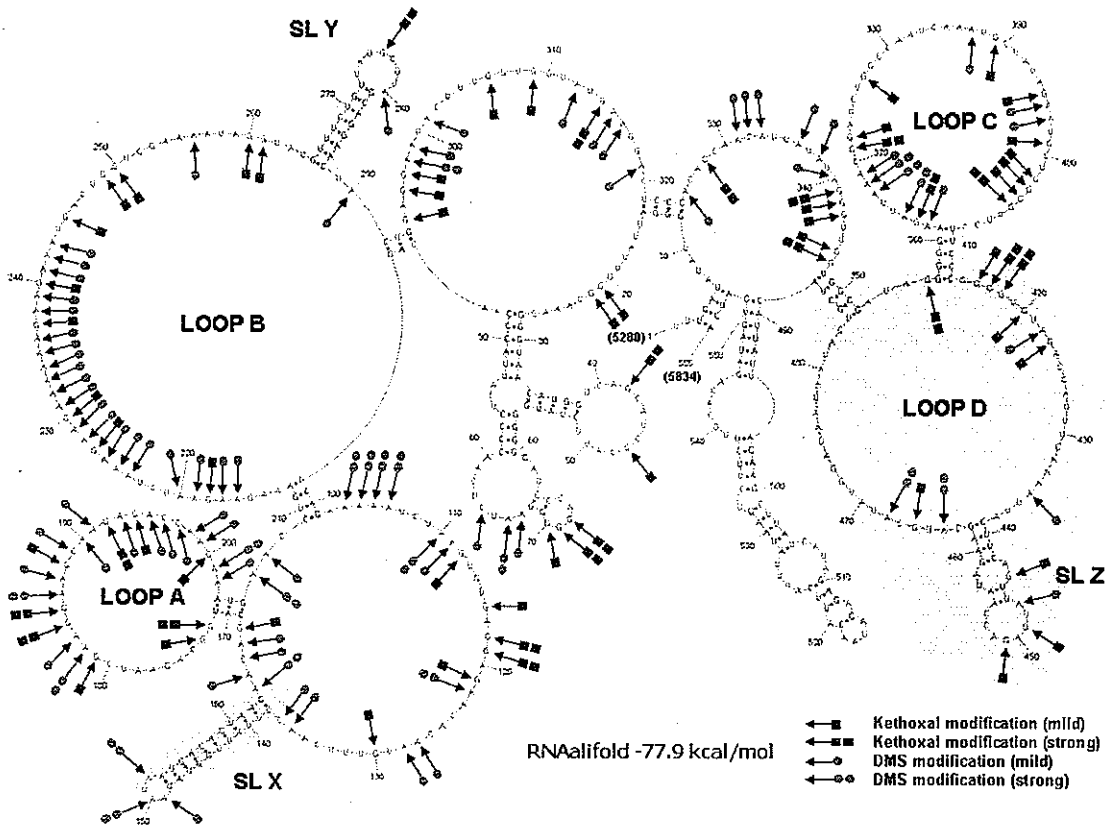


Figure 3.3. Chemical probing results mapped onto the RNA secondary structure of ERRE-1. RNAalifod (22) was used to generate an optimized RNA secondary structure of ERRE-1, based on a combination of thermodynamic considerations, experimental constraints, and sequence co-variation information derived from multiple sequence alignment of a collection of 140 ERRE-1 sequence variants. Arrowheads denote ribonucleotides modified by chemical probing reagents: kethoxal (red squares) and DMS (green circles), with the relative extent of modification represented by either two (strong) or one (weak) symbol. SL-X, -Y and -Z are stem-loop structures also shown in Figure 3.2.

Arrows indicate ribonucleotides for which chemical probing experiments indicated accessibility to modification by kethoxal (red rectangles) or DMS (green circles). The number of rectangles or circles corresponds to the intensity of the cleavage band at that position, with more symbols indicating a higher probability of "single-strandedness." The overall topology of this secondary structure of ERRE-1 is very similar to the secondary structures presented in Figure 3.2. In the optimized model shown in Figure 3.3, 373 of 555 ribonucleotides in ERRE-1 participate in the base-pairing and 182 are located in single-stranded regions. The estimated free energy for this structure, based on a combination of thermodynamic considerations, chemical probing results and co-variation analyses, is -77.90 kcal/mol. Notably, the ESE and the previously identified EIAV Rev binding region (5, 8, 33), are both located within the single-stranded loop B. Several structural features, including loop B, and stem-loop regions SL-X and SL-Y, are the same all five models shown in Figure 3.2B and Figure 3.3. The inclusion of covariation information results in one significant difference: in SL-Z, the base-paired region between nt 425-428 and nt 470-473 (Fig. 3.2) is converted to a single-stranded region in the optimized RNA secondary structure model, creating a new loop (loop D, Fig. 3.3).

EIAV Rev "footprints" two regions within ERRE-1

Previous experiments had implicated the purine-rich ESE within ERRE-1 as a primary binding site for EIAV Rev protein (5, 8). To obtain detailed information regarding the interaction of EIAV Rev with ERRE-1 sequences and structural motifs, we performed RNA "footprinting" experiments, using chemical probing and primer extension analysis (11, 47). Three different chemical reagents were used to compare the accessibilities of ribonucleotides in ERRE-1 to modification in the presence or absence of bound E-Rev. Hydroxyl radicals were used to monitor cleavages in the sugar-phosphate backbone, and

kethoxal (which modifies N1 and N2 of guanines) and DMS (which methylates N1 of adenines and N3 of cytosines) were employed as base-specific probes. The folded ERRE-1 RNA alone or Rev-ERRE-1 complexes were subjected to modification by each chemical reagent. Primer extension by reverse transcriptase was used to identify the chemically-modified ribonucleotides in the ERRE-1 sequence. In ladders of ^{32}P-labelled primer extension products visualized by denaturing polyacrylamide gel electrophoresis, sites of modification correspond to a "stop" or enhanced band located one position downstream (3′) of the band corresponding to the modified ribonucleotide site.

A summary of the footprinting results is shown in Figure 3.4, and representative footprinting gels are shown in Figures 3.5 and 3.6. Two distinct regions within ERRE-1 are "footprinted" by Rev (highlighted in yellow, Figure 3.4). Rev binding region-1 (RBR-1) is ~90 nt long (nt 170-260) and encompasses the ESE, including both purine-rich regions PuA and PuB (5, 8). A second domain, RBR-2, is ~110 nucleotides long (nt 360-470), and represents a newly-identified Rev interaction domain. Positions with enhanced kethoxal and DMS reactivity (circled residues) in ERev-ERRE-1 complexes, compared with the unbound RNA, are located primarily in single-stranded regions of the ERRE-1 secondary structure. Most regions protected from hydroxyl radical cleavage are also in single-stranded loops The experiments illustrated in Figures 3.5 and 3.6, together with many similar experiments using different primers (Table 3.1) to probe the entire ERRE-1 sequence were used to generate Figure 3.4, which summarizes the reproducible patterns of significant differences in chemical reactivity of the ERRE upon Rev binding.

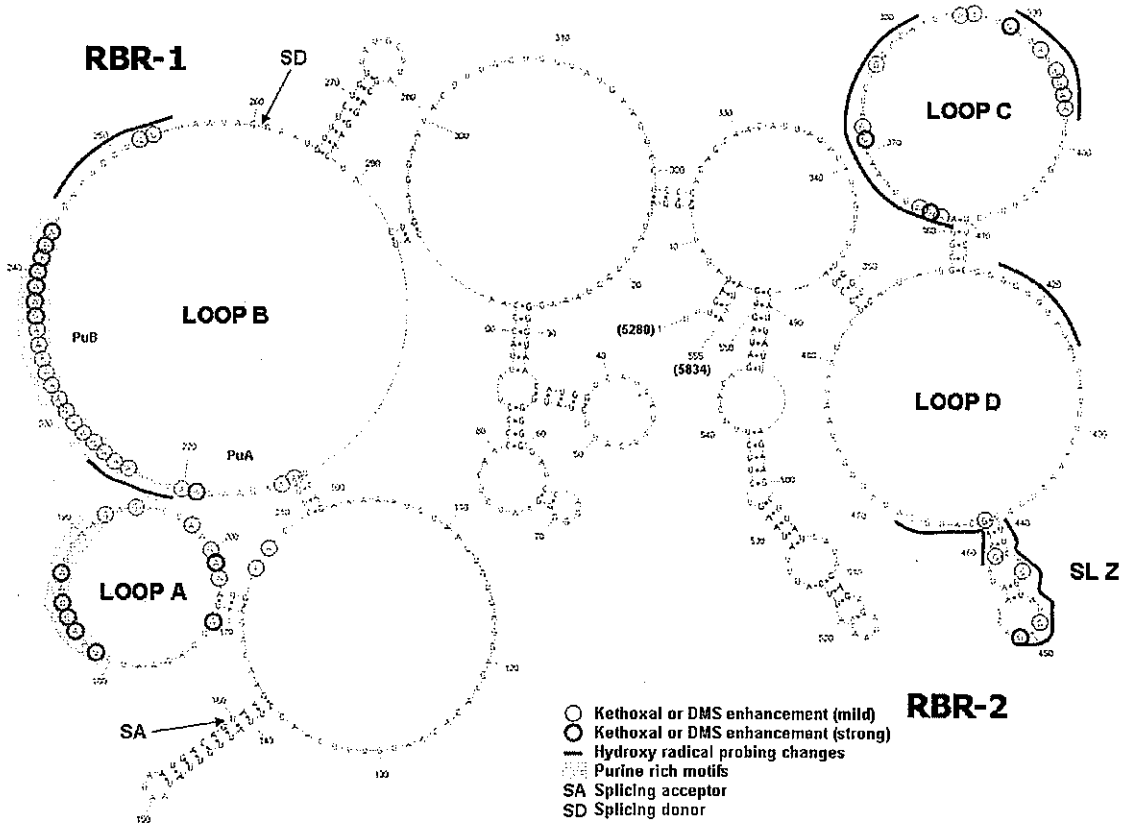


Figure 3.4. Two distinct regions of ERRE-1 undergo structural transitions in the presence of bound EIAV Rev protein. Consensus chemical modification patterns, based on at least 3 experiments in which several different primers were used to probe the complete ERRE-1 RNA, are mapped onto the RNA secondary structure model shown in Figure 3.3. Ribonucleotides that consistently display enhanced modification with either kethoxal or DMS upon Rev binding are circled: bold circle (strong) and thin circle (mild). Regions protected from hydroxyl radical cleavage in the presence of Rev are denoted by a thick line. Purine-rich motifs are highlighted in green.

RBR-1: Rev binding induces structural changes both within the ESE in ERRE-1 and in adjacent single-stranded regions

Within RBR-1, corresponding to the 5′ terminal region of ERRE-1, several purine-rich regions and GAR (guanine-adenine-purine) motifs have been identified as binding sites for the host splicing factor, SF2/ASF (19, 67). The purine-rich regions PuA and PuB, in particular, have been shown to bind both EIAV Rev and SF2/ASF in electrophoretic mobility shift assays (5, 8, 19, 33). To obtain more detailed footprinting information for this region, we systematically analyzed changes in the relative extent of ribonucleotide modification using titration experiments, in which ERev-ERRE-1 complexes were formed with increasing amounts of purified EIAV Rev protein (Fig. 3.5A, B).

Within loop B, numerous changes in ribonucleotide accessibilities were detected with increased amounts of EIAV Rev bound, especially in the PuA and PuB regions (Fig. 3.5). In the PuA, enhanced reactivity was seen for 4 out of 9 nucleotides and in PuB, all 20 purine residues showed enhanced reactivity, with both kethoxal and DMS (Fig. 3.5). Two regions of Rev-mediated protection from hydroxyl radical cleavage, which are not as strong as ones in the loop C and D, (see below) were also observed in loop B, one located in the sequence between PuA and PuB and another immediately downstream from PuB (nt 245-255, Fig 3.4). In addition to the two purine-rich motifs within the ESE, a third purine-rich motif, located in loop A, was strongly affected by Rev binding. Enhanced reactivity was seen for 7 out of 15 nucleotides within this motif and for 4 additional purine residues near the stem at the base of loop A (Fig. 3.5)

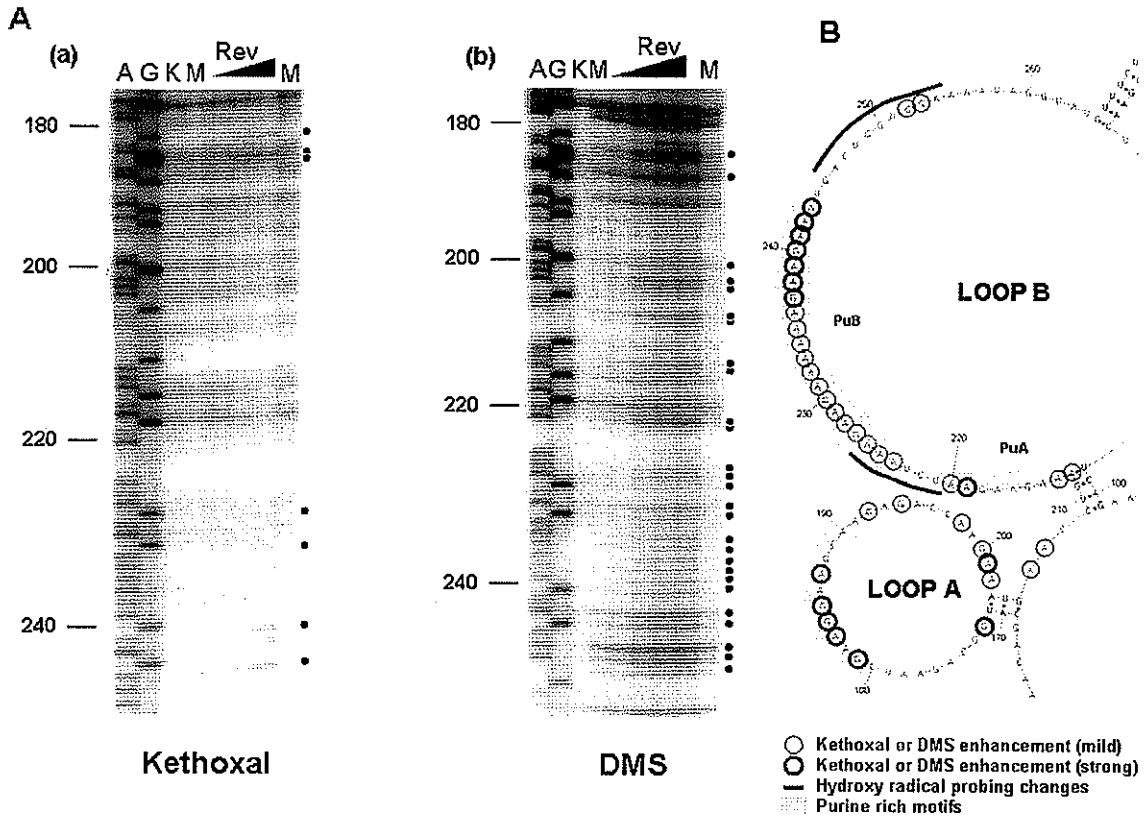


Figure 3.5. Representative footprinting results for RBR-1 **(A)** Representative gels from primer extension analysis of chemical probing experiments using kethoxal **(a)** and DMS **(b).** Similar experiments were performed using the hydroxyl radical cleavage reagent, Fe-EDTA (data not shown). Circled ribonucleotides denote positions with enhanced reactivity in the presence of bound EIAV Rev protein ("footprints"). Lanes A & G, Dideoxy sequencing markers; lane K, control, unmodified ERRE-1, in the absence of Rev; lane M, ERRE-1 modified in the absence of Rev; (lanes ◢) ERRE-1 modified in the presence of increasing amounts of Rev protein (1~30 fold molar excess). **(B)** Consensus chemical modification patterns in RBR-1, based on several experiments similar to those illustrated in part **(A),** are mapped onto the corresponding portion of the RNA secondary structure of ERRE-1 (from Fig. 3.3). Ribonucleotides that consistently display enhanced modification with either kethoxal or DMS upon Rev binding are circled: bold circle (strong) and thin circle (mild). Regions protected from hydroxyl radical cleavage in the presence of Rev are denoted by a thick line. Purine-rich motifs are highlighted in green.

Although the region corresponding to loop A, formed by nt 172 to 202, has not previously been implicated in EIAV Rev binding, our footprinting analyses revealed significantly enhanced reactivities of ribonucleotides in this loop. Therefore, we conclude that within RBR-1, a purine-rich motif in loop A, in addition to a region within loop B that encompasses the previously identified purine-rich PuA and PuB motifs in the ESE, undergoes significant structural rearrangement upon Rev binding.

RBR-2: A newly identified structured region in ERRE-1 also undergoes conformational changes upon Rev binding

By monitoring nucleotide accessibility changes in response to Rev binding across the entire ERRE-1 sequence, we were able to identify RBR-2, a region that has not been previously implicated in Rev binding, located approximately 100 nt 3′ to RBR-1. Hydroxyl radical probing identified several regions within loops C and D that are strongly protected from hydroxyl radical cleavage upon EAIV Rev binding (Fig. 3.6). Enhanced DMS and kethoxal reactivities were also observed for several positions within loops C and D. Within RBR-2, most ribonucleotides involved EIAV Rev binding are located within these large single-stranded regions. However, one stem-loop region, SL-Z, is strongly footprinted by Rev. SL-Z contains two protected stretches (nts 439-451 and 459-468) and several enhanced nucleotide reactivities (e.g., nts 451 and 459) (Fig. 3.6).

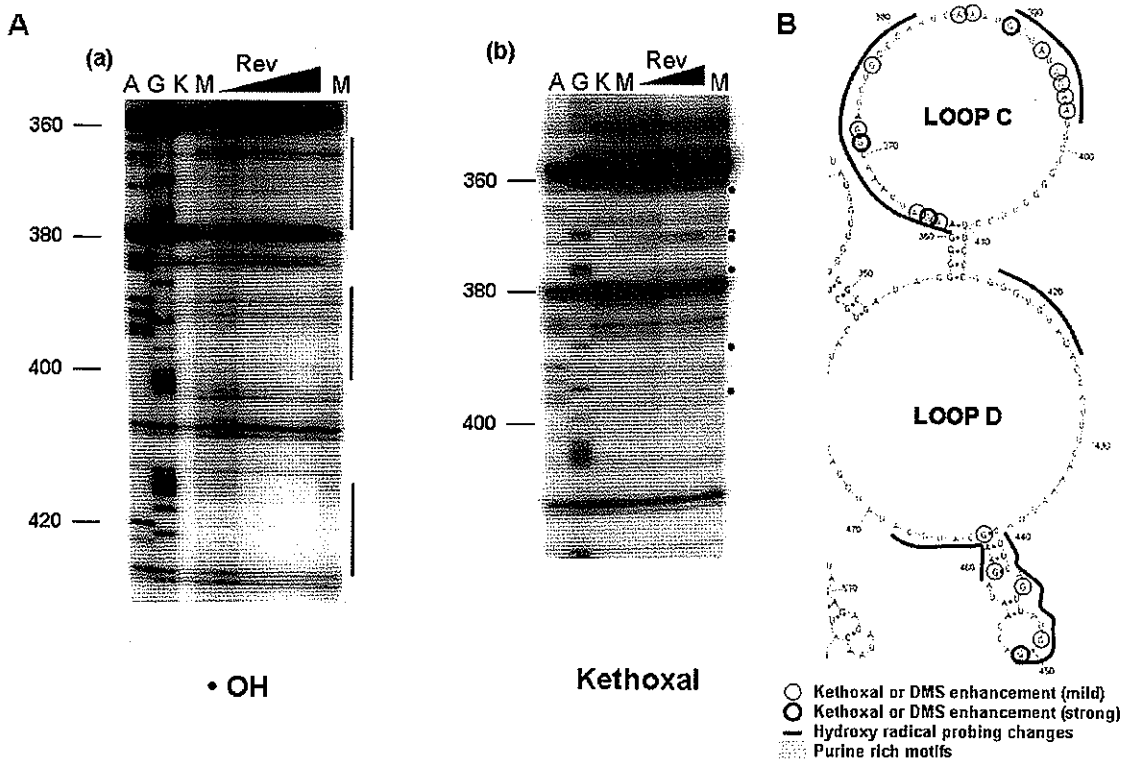


Figure 3.6. Representative Rev footprinting results in RBR-2 **(A)** Representative gels from primer extension analysis of chemical probing experiments using hydroxyl radical cleavage reagent, Fe-EDTA (•OH) **(a)** and DMS **(b).** Similar experiments were performed using kethoxal (data not shown). Circled ribonucleotides denote positions with enhanced reactivity in the presence of bound EIAV Rev protein ("footprints"). Lanes A & G, Dideoxy sequencing markers; lane K, control, unmodified ERRE-1, in the absence of Rev; lane M, ERRE-1 modified in the absence of Rev; (lanes ◢) ERRE-1 modified in the presence of increasing amounts of Rev protein (1~30 molar excess). **(B)** Consensus chemical modification patterns in RBR-2, based on several experiments similar to those illustrated in part **(A),** are mapped onto the corresponding portion of the RNA secondary structure of ERRE-1 (from Fig. 3.3). Ribonucleotides that consistently display enhanced modification with either kethoxal or DMS upon Rev binding are circled: bold circle (strong) and thin circle (mild). Regions protected from hydroxyl radical cleavage in the presence of Rev are denoted by a thick line. Purine-rich motifs are highlighted in green.

The discovery of interactions between EIAV Rev and RBR-2 is consistent with previous reports that the complete 555 nt ERRE-1 has more functional activity *in vivo* than several shorter ERRE-derived constructs that encompass the ESE region (5). We hypothesize that Rev-binding sites within RBR-2 are, in fact, the functional elements "missing" in shorter constructs that retain less activity than the complete ERRE-1 *in vivo*. Nitrocellulose filter binding experiments support this interpretation: the binding affinity of purified ERev for ERRE-1 (555 nt) is ~20nM, whereas a 123 nt fragment of ERRE-1 (nts 164-286) that lacks RBR-2 has much lower affinity, ~5μM, *in vitro* (data not shown), suggesting that RBR-2 is required for high affinity binding of EIAV Rev to ERRE-1.

The RNA sequence of ERRE-1 is conserved in variant EIAV sequences

An analysis of 178 HIV-1 variant sequences revealed that RNA sequences corresponding to the RRE are highly conserved as a result of evolutionary pressure for maintenance of both: i) the protein sequence encoded by the gp160 gene (within which the HIV-1 RRE is embedded), and ii) the RNA sequence and secondary structure of the RRE (51). To investigate the potential conservation of the ERRE-1 sequences at the RNA sequence level, we evaluated the information content in a group of variant EIAV sequences. Information content analysis is widely used to quantify sequence conservation in nucleic acid or protein sequences, (e.g., (51, 57, 58, 62)). Sequences of the EIAV gp90 gene, collected for analysis of envelope SU protein variants by Mealey, et al. (46) were aligned and information content was calculated as described in Materials and Methods. The plot in Figure 3.7 shows the distribution of information content values for ribonucleotide positions in the gp90 gene. The highest possible value for information content is 2, corresponding to cases in which all ribonucleotides at a particular position in the alignment are identical. Lines representing the previously described hypervariable regions in gp90 (46), the EIAV Rev exon 1 (61) coding

region and the Rev-footprinted regions determined in this study are shown. This analysis shows that, except for one segment corresponding to a known hypervariable region of gp90, the entire ERRE-1 region is highly conserved relative to the rest of gp90 gene. Further, the two major Rev-footprinted regions (RBR-1 and RBR-2) are located within this conserved region. Taken together, these results suggest that the conservation of ERRE-1 sequences results from evolutionary pressure for conservation of not only the protein sequence encoded by the gp90 gene, but also the RNA sequence and, potentially, RNA secondary structural features of ERRE-1 required for Rev binding.

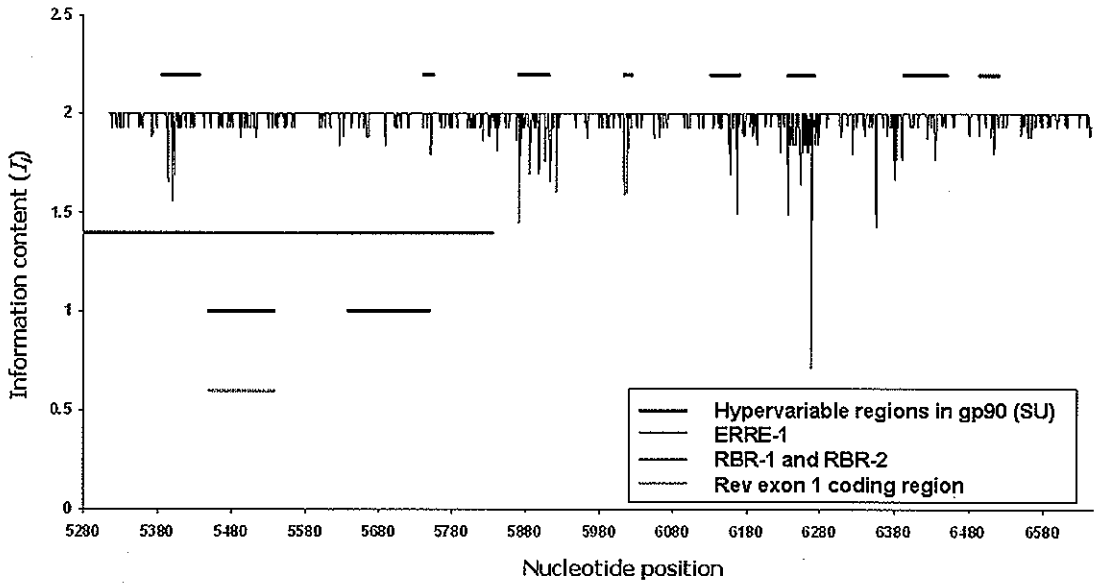


Figure 3.7. Conservation of RNA sequences in the gp90 (SU) gene of EIAV. Conservation of RNA sequences in EIAV Env gene (gp90) was assessed by evaluating information content at each nucleotide position in a CLUSTAL-W generated multiple sequence alignment of 140 gp90 sequence variants (see Materials and Methods for details). Information content (I) is plotted against nucleotide position, numbered from the first ribonucleotide in ERRE-1. The gp90 gene begins at position 35. The maximum information content value is 2, which corresponds to 100% conservation at a particular position in this alignment. The locations of gp90

hypervariable regions, identified in a previous analysis of SU variants (46), are indicated by blue lines above the graph. Colored horizontal lines indicate the locations of (pink), EIAV Rev binding regions 1 and 2 (RBR-1 & RBR-2) (red), and EIAV Rev exon 1 (green).

The high affinity Rev-binding sites of EIAV-1 and EIAV form a conserved RNA structural motif found in the RRE regions of diverse lentiviruses

To explore the possibility that RNA secondary structural elements required for Rev binding in EIAV might also be found in other lentiviruses, we first asked whether RBR-1 and RBR-2, identified in this study, have any predicted RNA structural features in common with one another or with the previously identified high affinity Rev binding site in HIV-1 RRE (stem-loop IIB) (9, 65). Pairwise RNA sequence and secondary structure alignments performed using RNAstructure Dynalign software (43) failed to identify significant RNA structural similarities in EIAV RBR-1 and RBR-2 (data not shown). In striking contrast, the stem-loop IIB region of HIV-1 and RBR-2 of EIAV, which encompasses a high affinity Rev binding site in EIAV, have potential to form a very similar ensemble of secondary structures (two examples are shown in Fig. 3.8A).

The unexpected RNA structural similarities detected through pairwise analysis of the HIV-1 and EIAV Rev-binding domains prompted us to generate a computational RNA structural motif model based on shared features of the two regions, using RNAMotif (37) (http://www.scripps.edu/mb/case/casegr-sh-3.5.html). The resulting RNA motif model was used to scan complete genomic sequences of ten different lentiviruses (see Materials and Methods). Figure 3.8B illustrates nine of the conserved RNA motifs identified in this analysis. In all except three cases (Visna virus, CAEV and BIV), the RNA motif lies within the Env gene. Both EIAV and HIV-1 have two copies of the RNA motif within the mapped RRE. In four other cases, a single copy of the motif occurs either within or near (< 250 nt

from) the proposed boundaries of RRE regions mapped in previous studies. The striking conservation of this RNA structural motif, together with its occurrence in or near within the known RRE regions of eight different lentiviruses and rabbit endogenous lentivirus type K (RELIK), suggests that it plays an important role in Rev-RRE recognition, not only in HIV-1 and EIAV, but potentially in all lentiviruses.

DISCUSSION

The Rev-responsive element (RRE) is an essential *cis*-acting RNA element recognized by the multifunctional Rev/Rex proteins of lentiviruses, including HIV-1 and EIAV. Rev-RRE interactions are involved in the regulation of viral gene expression and genomic replication, playing a critical role in the export of incompletely spliced or unspliced lentiviral mRNAs from the nucleus to the cytoplasm of infected host cells (24, 53). Because of their central role in viral replication, a better understanding of the molecular mechanisms that regulate Rev binding to RREs could give rise to new therapeutic approaches for treating for lentiviral diseases.

Organization of Functional Domains in the EIAV Rev Protein

Previous work has shown that the organization of functional domains within the sequences of the EIAV Rev protein and its HIV-1 counterpart differ significantly (18, 21, 33), despite the fact that activation domain is functionally interchangeable between EIAV and HIV-1 (18). In EIAV Rev, exon 1 appears to be non-essential, and exon 2 contains all functional domains identified to date: an N-terminal nuclear export signal (NES), a large central region that contains an arginine-rich motif (ARM) required for RNA binding, a non-essential region, and C-terminal region that contains overlapping motifs required for nuclear localization (NLS) and RNA binding (ARM) (18, 21, 33). In HIV-1 Rev, a single RNA

binding domain is located in the N-terminal half of the protein (53), whereas in EIAV Rev, the RNA binding domain is bipartite, composed of two ARMs separated by 79 amino acids in the linear amino acid sequence (33). One objective of the current study was to investigate whether these differences in the RNA binding domains of the two Rev proteins reflect differences in RNA sequences or structural features they recognize within the RREs of HIV-1 and EIAV.

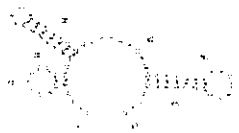
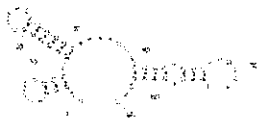
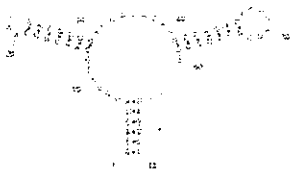
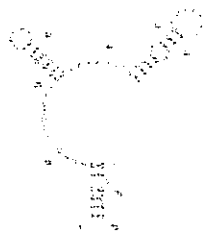
Mapping Rev Responsive Elements in Lentiviral Genomes

In most lentiviruses, including HIV-1, HIV-2, SIV, VV, and CAEV, the RRE is located near the HIV-1 protease cleavage site in the Env gene, between SU and TM proteins (12, 32, 55, 64, 68). In case of FIV, the RRE is located near the 3′ end of Env gene and in HTLV-1, the RRE in located within the 3′ LTR (10, 52, 56). In contrast, in EIAV, two RREs were roughly mapped in the Env gene (19, 39), and quantitative assays for Rev nuclear export activity showed that a 555 nt segment, ERRE-1, can provide RRE function *in vivo* (4).

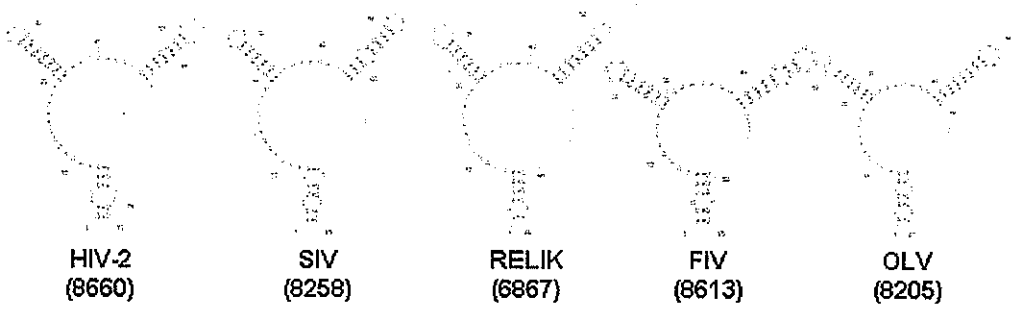
Several groups have attempted to identify stable secondary structural elements within the EIAV RRE using computational approaches. Martarano et al., (39) searched unsuccessfully for stable secondary structures analogous to the RREs within the 3′ halves of Env genes in other retroviruses. In that study, analyses of two experimentally mapped RREs (one near each end of the Env gene) failed to identify similarities in the sequence or predicted structures of the two putative RREs (39). More extensive computational analyses aimed at identifying highly structured regions (HSRs) in the EIAV Env gene were carried out by Lesnik et al., (34), who identified two HSRs that appeared to coincide with RREs proposed by Gontarek, et al. (19). The two HSRs identified in EIAV had relatively unstable predicted structures (-0.28 and -0.26 kcal/mol nt) compared with those found in other retroviral RREs, such as those of HIV-1 and SIV (-0.42 and -0.46 kcal/mol nt) (34). One of the HSRs

identified in the Lesnik study appears to roughly co-localize with RBR-1 and the ESE, although its predicted structure and precise location differ.

A.

	HIV-1	EIAV
Structure #1		
Structure #11		

B.



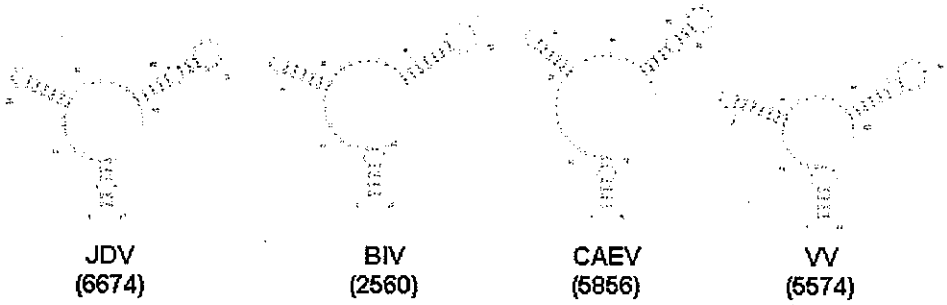


Figure 3.8. An RNA structural motif identified in the high affinity Rev-binding sites of HIV-1 and EIAV is conserved within or near the RRE regions or Env gene of diverse lentiviruses. **(A)** Two representative structures (#1 and #11) among 20 identified ensemble structures, were represented. 100 nt HIV-1 and EIAV binding region was extracted and structurally aligned (see Materials and Methods) using Dynalign software (43). **(B)** A conserved RNA structural motif was identified in the genomic RNA sequences of eight different lentiviruses, HIV-2, SIV (simian immunodeficiency virus), RELIK (rabbit endogenous lentivirus type K), FIV (feline immunodeficiency virus), OLV (ovine lentivirus), JDV (jembrana disease virus), BIV (bovine immunodeficiency virus), CAEV (caprine arthritis-encephalitis virus), and Visna (visna virus). An identified RNA structural motif (structure #11) based on shared RNA secondary structural features of high affinity Rev binding sites of from EIAV and HIV-1, was used to scan the complete genomic RNA sequences of eight different lentiviruses.

An Experimentally Validated RNA Structure for the EIAV RRE

This study represents the first high-resolution analysis of EIAV RRE secondary structure and Rev-RRE interactions. Using a combination of computational and experimental approaches, we generated an optimized RNA secondary structure model for the EIAV ERRE-1. In the context of this structure, individual ribonucleotides and secondary structural elements that are protected from chemical cleavage or undergo structural transitions upon Rev binding were identified using detailed RNA footprinting experiments.

One striking characteristic of our proposed ERRE-1 structure is the relative paucity of stable stem structures and abundance of single-stranded loops, several of which are unusually large compared with loops in known RNA structures. The calculated free energy of the optimized ERRE-1 secondary structure, based on a combination of thermodynamic calculations, chemical probing data and ribonucleotide co-variation frequencies using RNAalifold is -77.9 kcal/mol (Figure 3.3), a value similar to that obtained using

RNAStructure -71.3 kcal/mol (Fig. 3.2D). These values are significantly higher than free energies of alternative secondary structures predicted using computational methods that do not incorporate experimental constraints or available phylogenetic information, e.g., -126.05 kcal/mol, using Mfold (Fig. 3.2A). Incorporating experimental constraints has been shown to result in significant improvement in the fidelity of RNA structure predictions using RNAStructure software, for a variety of RNAs (41, 42). Thus, we believe the structural model for ERRE-1 shown in Figure 3.2A, based solely on thermodynamic calculations, is less likely to be physiologically relevant than structures presented in Figures 3.2B and Figure 3.3, which may reflect features of the ensemble of ERRE-1 structures that exist *in vivo*.

Two Distinct Rev Binding Domains within the RRE of EIAV

The ERRE-1 footprinting experiments provided a detailed view of the interaction between EIAV Rev and RBR-1, which contains the ESE region previously implicated in Rev binding. Within loop B, changes in chemical reactivities of every purine residue in PuA and several As in PuB were detected, consistent with previous studies reporting decreased Rev binding when mutations converting several GAA motifs into GCAs were introduced within PuB (5). Similarly, EIAV Rev binding activity was virtually abolished when a GAA sequence was changed to GAU in PuB (8). Both previous studies reported that mutations in PuA had little effect on the binding affinity of EIAV Rev, again consistent with the footprinting results reported here: only a few changes in ribonucleotide accessibilities in this region were detected upon Rev binding. In contrast, Rev binding resulted in significant changes in ribonucleotide accessibilities in a third purine-rich motif, located within loop A (Fig. 3.3), including enhanced chemical reactivities at several purines within the loop. Thus, in RBR-1, structural changes that occur upon Rev binding include enhanced nucleotide

reactivities within loop A, in addition to the previously mapped Rev-binding region in loop B.

RBR-2, a novel Rev-binding region within ERRE-1, was identified as second Rev-mediated footprint, encompassing relatively long tracts of residues protected from hydroxyl radical cleavage and several individual ribonucleotides with enhanced reactivity to kethoxal and DMS. Quantitative *in vitro* binding assays showed that RBR-2 is required for high affinity binding of EIAV Rev: ERRE-1 fragments containing both RBR-1 and RBR-2 have much higher affinity for EIAV Rev than those containing only RBR-1.

For RBR-2, all computational methods produced the same RNA secondary structure. The Rev-induced footprint in SL-Z (nt 439-462) in RBR-2 is especially interesting because the topology of this stem-loop structure is very similar to that of stem-loop IIB in HIV-I, which corresponds to the primary high-affinity Rev binding site in the HIV-1 RRE (9, 65). Although the sequences of the two structures differ, they share a strong protection pattern on opposite strands in the stem, characteristic of a "5′ stagger" of 7 nucleotides. A 5′ stagger in the footprinting protection pattern on complementary strands of an RNA duplex suggests interaction of the bound protein within the major groove of the RNA (as in the interaction between ribosomal protein S7 and the 3′ domain of 16S rRNA in *E. coli.* (54)). This similarity supports the hypothesis that EIAV Rev binds within major groove of the RNA stem-loop structure, SL-Z, in RBR-2 (Fig. 3.5C), in a manner analogous to that observed for HIV-1 peptides bound to RNA oligonucleotides corresponding to stem-loop IIB (2).

In HIV-1, an extended RRE sequence that includes structured regions surrounding the stem-loop IIB binding site has been shown to enhance Rev binding and promote its multimerization on HIV-1 RNAs (27, 38). We have observed multimerization of EIAV Rev on fragments of ERRE-1 using *in vitro* binding assays such as EMSA and UV-crosslinking (Lee pub & unpublished data), but we are not aware of any direct evidence for a functional

role of EIAV Rev multimerization *in vivo*. Further investigation will be required to elucidate the functional relationships among RBR-1, RBR-2 and the full length EIAV RRE.

RNA Structural Motifs in Lentiviral RREs may also Regulate Alternative mRNA Splicing

In the EIAV secondary structure proposed in this study, there are two stem-loop structures near the splicing donor and acceptor sites of EIAV Rev exon 1. The 3′ splicing acceptor site of Rev exon 1 is located within SL-X and the 5′ splicing donor site is near SL-Y. Both of these stem-loop structures were predicted by every secondary structure prediction method used in this study (Figs. 3.2, 3.3 and data not shown). In addition, both stem-loop structures are well conserved in EIAV variant sequences. Two compensatory mutations observed in SL-X maintain the stem structure: i) 146U-154G to 146C-156G and ii) 148U-152A to 148U-152G. In SL-Y, 3 compensatory mutations: i) 272U-281G to 272C-281G, ii) 270U-283A to 270U-283G, and iii)267U-285A to 267U-285G are observed in the stem.

The importance of specific RNA secondary structures in regulating of pre-mRNA splicing events is well established (6) and there are several cases in which stem-loop structures near splicing donor and acceptor sites are known to play a role in alternative splicing (15, 36, 45, 60). In HIV-1 pre-mRNA, stem-loop structures have been identified near splicing acceptor sites and are proposed mediate splicing events (26). Stem-loop structure 2 (SLS2), encompassing the 3′ splice acceptor site (A3) for the HIV-1 Tat exon is well conserved in HIV-1 variants and in related SIVcpz strains. One cis-regulatory element required for splicing, a polypyrimidine track (PPT) is located in the SLS2 stem. By designing specific mutations that either changed the base-pairing potential of this region or altered its pyrimidine composition, or both, Jacquenet et al, demonstrated that both the sequence of the PPT and the thermodynamic stability of the helical stem in SLS2 affect splicing efficiency

(26). Similarly, both the configuration and nucleotide composition of a *cis*-acting intronic stem-loop motif near the 5′ splice were shown to be important for regulating alternative splicing of the HIV-1 Tat exon (45).

In EIAV, the SL-X loop of ERRE-1 includes two cis-regulatory elements for splicing, an AG consensus sequence and weak polypyrimidine track (PPT), similar to that in the A3 splicing site in HIV-1 tat exon (26). Interestingly, in the absence of the ESE sequence in exon 1, virtually all multiply-spliced EIAV mRNAs do not include exon 1 (35). We hypothesize, therefore, that the masking of 3′ splicing sites due to formation of stem-loop structures identified here, in conjunction with the weak PPT, could prevent the spliceosome from recognizing the 3′ splicing site in the absence of splicing factors like SF2/ASF, an idea that has been suggested previously (19). Further analysis will be required to elucidate the complex interactions among *cis*-acting RNA regulatory elements and *trans*-acting proteins within ERRE-1 in directing alternative splicing of exon 1 in EIAV.

As in other lentiviruses, the EIAV RRE overlaps the Env gene (gp90). ERRE-1 differs from other lentiviral RREs, however, in that it also overlaps exon 1 of the Rev gene. As discussed above, alternative splicing events in EIAV appear to be regulated both by the EIAV Rev protein and by cellular splicing factors such as SF2/ASF (5, 8, 19, 35, 39). In particular, Rev-RRE interactions likely affect the ratio of partially-spliced to fully-spliced cytoplasmic EIAV mRNAs. Partially-spliced viral mRNAs are essential for the late stages of EIAV replication and production of only fully-sliced mRNAs, generated by the host's cellular splicing machinery, can prevent viral replication (3, 8, 19, 35). Direct competition between EIAV Rev and SF2/ASF for binding the RRE has been suggested as one potential mechanism for regulating the course of EIAV infection (5, 19, 35), an idea supported by footprinting results showing that SF2/ASF and EIAV Rev bind overlapping elements within ERRE-1 (Park *et al.*, unpublished data).

A Conserved RNA Structural Motif for Rev Recognition in both EIAV and HIV-1

To examine the relationship between conservation of the ERRE-1 sequence and potential conservation of its RNA secondary structure, we assessed ribonucleotide conservation of the RRE sequence using classical information theory. Measurement of the information content of nucleotide or protein sequences has been widely used to evaluate the conservation of certain motifs, such as transcription factor binding sites in DNA, splice sites in pre-mRNAs, and functional motifs in proteins (1, 49, 66). Our analysis of 258 EIAV sequence variants showed that ERRE-1 region is highly conserved compared with other regions within the gp90 gene. In EIAV, the gp90 gene overlaps both ERRE-1 and the non-essential exon 1 of the Rev gene, making it difficult to clearly delineate the relationship between specific sequence variants and the functions of the overlapping genes. However, constant and hypervariable regions of the gp90 gene were previously identified using this approach (46) and even within the context of the conserved region of gp90, the 555 nt ERRE-1 is more highly conserved at the ribonucleotide sequence level than expected for a non-functional region (Figure 3.7). Finally, it is intriguing that two protein encoding regions (a gp90 constant region and exon 1 of Rev), as well as three *cis*-acting RNA elements recognized by regulatory proteins (an ESE, RBR-1 and RBR-2), are all located within the most highly conserved region of the ERRE-1 sequence, suggesting that it has experienced selective pressure at both the amino acid and ribonucleotide sequence levels.

To investigate whether the RBR-2 domain might have RNA structural features in common with the high affinity Rev binding stem-loop IIB region of HIV-1, we performed pairwise comparisons using Dynalign software (43). We found that the two RRE regions can form very similar ensembles of secondary structures (Fig. 3.8A and data not shown). Because these predictions were generated using fragments of the two RRE sequences, we tested whether such similar structures could be detected in the context of the complete genomic

RNA sequences of the two lentiviruses. Using a computational RNA motif model based on common structural elements identified in the HIV-1 and EIAV Rev binding sites, a shared RNA motif was detected within the RREs of HIV-1 and EIAV. Strikingly, the same motif was found within or very near the proposed RREs in six out of ten lentiviruses examined, suggesting that it could play a role in Rev-RRE recognition in diverse lentiviruses (Fig. 3.8). This hypothesis is supported by an unexpected similarity in the binding mode of HIV-1 and EIAV Rev to this conserved RNA motif. Results of our detailed footprinting experiments suggest that EIAV Rev binds in the major groove of the double-stranded RNA stem corresponding to the high affinity Rev binding site in ERRE-1, as is the case for HIV-1 Rev binding to stem-loop IIB in the HIV-1 RRE. Recently, the first endogenous lentivurses were discovered and characterized by Katzourakik et al. (28). Although the viruses are more than 7 million years old, the viruses still have many lentiviral features like genomic structures and regulation proteins like tat and rev. Interestingly, when we did the same search using RELIK entire sequence, we found the common secondary structural motif in RELIK Env gene region (Fig 3.8B).

EIAV is genetically the simplest of the exogenous lentiviruses, has the fewest number of genes, and lacks a vif protein. Phlyogenetic analyses suggest that the ancient endogenous rabbit lentivirus clusters more closely with EIAV than with the other exogenous lentivirus (28). EIAV therefore offers an opportunity for comparative analysis of the molecular interactions important in regulation of lentiviral gene expression that may identify highly conserved interactions that could be targeted in novel anti-lentiviral therapies.

REFERENCES

1. **Bairoch, A.** 1991. PROSITE: a dictionary of sites and patterns in proteins. Nucleic Acids Res **19 Suppl:**2241-5.

2. **Battiste, J. L., H. Mao, N. S. Rao, R. Tan, D. R. Muhandiram, L. E. Kay, A. D. Frankel, and J. R. Williamson.** 1996. Alpha helix-RNA major groove recognition in an HIV-1 rev peptide-RRE RNA complex. Science **273:**1547-51.

3. **Belshan, M., P. Baccam, J. L. Oaks, B. A. Sponseller, S. C. Murphy, J. Cornette, and S. Carpenter.** 2001. Genetic and biological variation in equine infectious anemia virus Rev correlates with variable stages of clinical disease in an experimentally infected pony. Virology **279:**185-200.

4. **Belshan, M., M. E. Harris, A. E. Shoemaker, T. J. Hope, and S. Carpenter.** 1998. Biological characterization of Rev variation in equine infectious anemia virus. J Virol **72:**4421-6.

5. **Belshan, M., G. S. Park, P. Bilodeau, C. M. Stoltzfus, and S. Carpenter.** 2000. Binding of equine infectious anemia virus rev to an exon splicing enhancer mediates alternative splicing and nuclear export of viral mRNAs. Mol Cell Biol **20:**3550-7.

6. **Buratti, E., and F. E. Baralle.** 2004. Influence of RNA secondary structure on the pre-mRNA splicing process. Mol Cell Biol **24:**10505-14.

7. **Chaitin, G. J.** 1982. Algorithmic information theory. Encyclopedia of Statistical Sciences **1:**38-41.

8. **Chung, H., and D. Derse.** 2001. Binding sites for Rev and ASF/SF2 map to a 55-nucleotide purine-rich exonic element in equine infectious anemia virus RNA. J Biol Chem **276:**18960-7.

9. **Cook, K. S., G. J. Fisk, J. Hauber, N. Usman, T. J. Daly, and J. R. Rusche.** 1991. Characterization of HIV-1 REV protein: binding stoichiometry and minimal RNA substrate. Nucleic Acids Res **19:**1577-83.

10. **Cullen, B. R.** 1992. Mechanism of action of regulatory proteins encoded by complex retroviruses. Microbiol Rev **56:**375-94.

11. **Culver, G. M., and H. F. Noller.** 2000. Directed hydroxyl radical probing of RNA from iron(II) tethered to proteins in ribonucleoprotein complexes. Methods Enzymol **318:**461-75.

12. **Dillon, P. J., P. Nelbock, A. Perkins, and C. A. Rosen.** 1990. Function of the human immunodeficiency virus types 1 and 2 Rev proteins is dependent on their ability to interact with a structured region present in env gene mRNA. J Virol **64:**4428-37.

13. **Ding, Y., C. Y. Chan, and C. E. Lawrence.** 2004. Sfold web server for statistical folding and rational design of nucleic acids. Nucleic Acids Res **32:**W135-41.

14. **Ehresmann, C., F. Baudin, M. Mougel, P. Romby, J. P. Ebel, and B. Ehresmann.** 1987. Probing the structure of RNAs in solution. Nucleic Acids Res **15:**9109-28.

15. **Estes, P. A., N. E. Cooke, and S. A. Liebhaber.** 1992. A native RNA secondary structure controls alternative splice-site selection and generates two human growth hormone isoforms. J Biol Chem **267:**14902-8.

16. **Foley, B. T.** 2000. An overview of the molecular phylogeny of lentiviruses. HIV Sequence Compendium 2000:35-43.

17. **Fridell, R. A., H. P. Bogerd, and B. R. Cullen.** 1996. Nuclear export of late HIV-1 mRNAs occurs via a cellular protein export pathway. Proc Natl Acad Sci U S A **93:**4421-4.

18. **Fridell, R. A., K. M. Partin, S. Carpenter, and B. R. Cullen.** 1993. Identification of the activation domain of equine infectious anemia virus rev. J Virol **67:**7317-23.

19. **Gontarek, R. R., and D. Derse.** 1996. Interactions among SR proteins, an exonic splicing enhancer, and a lentivirus Rev protein regulate alternative splicing. Mol Cell Biol **16:**2325-31.

20. **Han, K., and Y. Byun.** 2003. PSEUDOVIEWER2: Visualization of RNA pseudoknots of any type. Nucleic Acids Res **31:**3432-40.

21. **Harris, M. E., R. R. Gontarek, D. Derse, and T. J. Hope.** 1998. Differential requirements for alternative splicing and nuclear export functions of equine infectious anemia virus Rev protein. Mol Cell Biol **18:**3889-99.

22. **Hofacker, I. L., M. Fekete, and P. F. Stadler.** 2002. Secondary structure prediction for aligned RNA sequences. J Mol Biol **319:**1059-66.

23. **Hofacker, I. L., W. Fontana, P. F. Stadler, S. Bonhoeffer, M. Tacker, and P. Schuster.** 1994. Fast Folding and Comparison of RNA Secondary Structures. Monatshefte f. Chemie **125:**167-188.

24. **Hope, T. J.** 1999. The ins and outs of HIV Rev. Arch Biochem Biophys **365:**186-91.

25. **Ihm, Y., O. W. Sparks, J. H. Lee, W. Wannemuehler, M. Terribilini, H. Cao, C. Z. Wang, S. Carpenter, K.-H. Ho, and D. Dobbs.** 2007. Structural model of the Rev regulatory protein from Equine Infectious Anemia Virus (EIAV). In preparation.

26. **Jacquenet, S., D. Ropers, P. S. Bilodeau, L. Damier, A. Mougin, C. M. Stoltzfus, and C. Branlant.** 2001. Conserved stem-loop structures in the HIV-1 RNA region containing the A3 3′ splice site and its cis-regulatory element: possible involvement in RNA splicing. Nucleic Acids Res **29:**464-78.

27. **Jain, C., and J. G. Belasco.** 2001. Structural model for the cooperative assembly of HIV-1 Rev multimers on the RRE as deduced from analysis of assembly-defective mutants. Mol Cell 7:603-14.

28. **Katzourakis, A., M. Tristem, O. G. Pybus, and R. J. Gifford.** 2007. Discovery and analysis of the first endogenous lentivirus. Proc Natl Acad Sci U S A **104:**6261-5.

29. **Kawakami, T., L. Sherman, J. Dahlberg, A. Gazit, A. Yaniv, S. R. Tronick, and S. A. Aaronson.** 1987. Nucleotide sequence analysis of equine infectious anemia virus proviral DNA. Virology **158:**300-12.

30. **Kjems, J., M. Brown, D. D. Chang, and P. A. Sharp.** 1991. Structural analysis of the interaction between the human immunodeficiency virus Rev protein and the Rev response element. Proc Natl Acad Sci U S A **88:**683-7.

31. **Kuzmic, P.** 1996. Program DYNAFIT for the analysis of enzyme kinetic data: application to HIV proteinase. Anal Biochem **237:**260-73.

32. **Le, S. Y., M. H. Malim, B. R. Cullen, and J. V. Maizel.** 1990. A highly conserved RNA folding region coincident with the Rev response element of primate immunodeficiency viruses. Nucleic Acids Res **18:**1613-23.

33. **Lee, J. H., S. C. Murphy, M. Belshan, W. O. Sparks, Y. Wannemuehler, S. Liu, T. J. Hope, D. Dobbs, and S. Carpenter.** 2006. Characterization of functional domains of equine infectious anemia virus Rev suggests a bipartite RNA-binding domain. J Virol **80:**3844-52.

34. **Lesnik, E. A., R. Sampath, and D. J. Ecker.** 2002. Rev response elements (RRE) in lentiviruses: an RNAMotif algorithm-based strategy for RRE prediction. Med Res Rev **22:**617-36.

35. **Liao, H. J., C. C. Baker, G. L. Princler, and D. Derse.** 2004. cis-Acting and trans-acting modulation of equine infectious anemia virus alternative RNA splicing. Virology **323:**131-40.

36. **Libri, D., A. Piseri, and M. Y. Fiszman.** 1991. Tissue-specific splicing in vivo of the beta-tropomyosin gene: dependence on an RNA secondary structure. Science **252:**1842-5.

37. **Macke, T. J., D. J. Ecker, R. R. Gutell, D. Gautheret, D. A. Case, and R. Sampath.** 2001. RNAMotif, an RNA secondary structure definition and search algorithm. Nucleic Acids Res **29:**4724-35.

38. **Mann, D. A., I. Mikaelian, R. W. Zemmel, S. M. Green, A. D. Lowe, T. Kimura, M. Singh, P. J. Butler, M. J. Gait, and J. Karn.** 1994. A molecular rheostat. Co-operative rev binding to stem I of the rev-response element modulates human immunodeficiency virus type-1 late gene expression. J Mol Biol **241:**193-207.

39. **Martarano, L., R. Stephens, N. Rice, and D. Derse.** 1994. Equine infectious anemia virus trans-regulatory protein Rev controls viral mRNA stability, accumulation, and alternative splicing. J Virol **68:**3102-11.

40. **Mathews, D. H.** 2006. Revolutions in RNA secondary structure prediction. J Mol Biol **359:**526-32.

41. **Mathews, D. H.** 2004. Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. Rna **10:**1178-90.

42. **Mathews, D. H., M. D. Disney, J. L. Childs, S. J. Schroeder, M. Zuker, and D. H. Turner.** 2004. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. Proc Natl Acad Sci U S A **101:**7287-92.

43. **Mathews, D. H., and D. H. Turner.** 2002. Dynalign: an algorithm for finding the secondary structure common to two RNA sequences. J Mol Biol **317:**191-203.

44. **Mathews, D. H., and D. H. Turner.** 2006. Prediction of RNA secondary structure by free energy minimization. Curr Opin Struct Biol **16:**270-8.

45. **McAlinden, A., N. Havlioglu, L. Liang, S. R. Davies, and L. J. Sandell.** 2005. Alternative splicing of type II procollagen exon 2 is regulated by the combination of a weak 5′ splice site and an adjacent intronic stem-loop cis element. J Biol Chem **280:**32700-11.

46. **Mealey, R. H., S. R. Leib, S. L. Pownder, and T. C. McGuire.** 2004. Adaptive immunity is the primary force driving selection of equine infectious anemia virus envelope SU variants during acute infection. J Virol **78:**9295-305.

47. **Merryman, C. N., H. F.** 1998. Footprinting and modification-interference analysis of binding sites on RNA. Oxford University Press, New York.

48. **Moazed, D., S. Stern, and H. F. Noller.** 1986. Rapid chemical probing of conformation in 16 S ribosomal RNA and 30 S ribosomal subunits using primer extension. J Mol Biol **187:**399-416.

49. **Mount, S. M., C. Burks, G. Hertz, G. D. Stormo, O. White, and C. Fields.** 1992. Splicing signals in Drosophila: intron size, information content, and consensus sequences. Nucleic Acids Res **20:**4255-62.

50. **Olsen, H. S., A. W. Cochrane, P. J. Dillon, C. M. Nalin, and C. A. Rosen.** 1990. Interaction of the human immunodeficiency virus type 1 Rev protein with a structured region in env mRNA is dependent on multimer formation mediated through a basic stretch of amino acids. Genes Dev **4:**1357-64.

51. **Peleg, O., S. Brunak, E. N. Trifonov, E. Nevo, and A. Bolshoy.** 2002. RNA secondary structure and squence conservation in C1 region of human immunodeficiency virus type 1 env gene. AIDS Res Hum Retroviruses **18:**867-78.

52. **Phillips, T. R., C. Lamont, D. A. Konings, B. L. Shacklett, C. A. Hamson, P. A. Luciw, and J. H. Elder.** 1992. Identification of the Rev transactivation and Rev-responsive elements of feline immunodeficiency virus. J Virol **66:**5464-71.

53. **Pollard, V. W., and M. H. Malim.** 1998. The HIV-1 Rev protein. Annu Rev Microbiol **52:**491-532.

54. **Powers, T., and H. F. Noller.** 1995. Hydroxyl radical footprinting of ribosomal proteins on 16S rRNA. Rna **1:**194-209.

55. **Saltarelli, M., G. Querat, D. A. Konings, R. Vigne, and J. E. Clements.** 1990. Nucleotide sequence and transcriptional analysis of molecular clones of CAEV which generate infectious virus. Virology **179:**347-64.

56. **Saltarelli, M. J., R. Schoborg, G. N. Pavlakis, and J. E. Clements.** 1994. Identification of the caprine arthritis encephalitis virus Rev protein and its cis-acting Rev-responsive element. Virology **199:**47-55.

57. **Schneider, T. D., and G. D. Stormo.** 1989. Excess information at bacteriophage T7 genomic promoters detected by a random cloning technique. Nucleic Acids Res **17:**659-74.

58. **Schneider, T. D., G. D. Stormo, L. Gold, and A. Ehrenfeucht.** 1986. Information content of binding sites on nucleotide sequences. J Mol Biol **188:**415-31.

59. **Shannon, C. E.** 1948. A mathematical theory of communication. The Bell System Technical Journal **27:**379-423, 623-656.

60. **Singh, N. N., R. N. Singh, and E. J. Androphy.** 2007. Modulating role of RNA structure in alternative splicing of a critical exon in the spinal muscular atrophy genes. Nucleic Acids Res **35:**371-89.

61. **Stephens, R. M., D. Derse, and N. R. Rice.** 1990. Cloning and characterization of cDNAs encoding equine infectious anemia virus tat and putative Rev proteins. J Virol **64:**3716-25.

62. **Stephens, R. M., and T. D. Schneider.** 1992. Features of spliceosome evolution and function inferred from an analysis of the information at human splice sites. J Mol Biol **228:**1124-36.

63. **Thompson, J. D., D. G. Higgins, and T. J. Gibson.** 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res **22:**4673-80.

64. **Tiley, L. S., and B. R. Cullen.** 1992. Structural and functional analysis of the visna virus Rev-response element. J Virol **66:**3609-15.

65. **Tiley, L. S., M. H. Malim, H. K. Tewary, P. G. Stockley, and B. R. Cullen.** 1992. Identification of a high-affinity RNA-binding site for the human immunodeficiency virus type 1 Rev protein. Proc Natl Acad Sci U S A **89:**758-62.

66. **Workman, C. T., and G. D. Stormo.** 2000. ANN-Spec: a method for discovering transcription factor binding sites with improved specificity. Pac Symp Biocomput:467-78.

67. **Xu, R., J. Teng, and T. A. Cooper.** 1993. The cardiac troponin T alternative exon contains a novel purine-rich positive splicing element. Mol Cell Biol **13:**3660-74.

68. **Zapp, M. L., and M. R. Green.** 1989. Sequence-specific RNA binding by the HIV-1 Rev protein. Nature **342:**714-6.

69. **Zapp, M. L., T. J. Hope, T. G. Parslow, and M. R. Green.** 1991. Oligomerization and RNA binding domains of the type 1 human immunodeficiency virus Rev protein: a dual function for an arginine-rich binding motif. Proc Natl Acad Sci U S A **88:**7734-8.

70. **Zuker, M.** 2003. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res **31:**3406-15.

CHAPTER 4. A SINGLE AMINO ACID DIFFERENCE WITHIN THE A-2 DOMAIN OF TWO NATURALLY OCCURRING EQUINE MHC CLASS I MOLECULES ALTERS THE RECOGNITION OF GAG AND REV EPITOPES BY EQUINE INFECTIOUS ANEMIA VIRUS-SPECIFIC CTL

A paper published in *Journal of Immunology*

Robert H. Mealey, Jae-Hyung Lee, Steven R. Leib, Matt H. Littke and Travis C. McGuire

ABSTRACT

Although CTL are critical for control of lentiviruses, including equine infectious anemia virus, relatively little is known regarding the MHC class I molecules that present important epitopes to equine infectious anemia virus-specific CTL. The equine class I molecule 7-6 is associated with the equine leukocyte Ag (ELA)-A1 haplotype and presents the Env-RW12 and Gag-GW12 CTL epitopes. Some ELA-A1 target cells present both epitopes, whereas others are not recognized by Gag-GW12-specific CTL, suggesting that the ELA-A1 haplotype comprises functionally distinct alleles. The Rev-QW11 CTL epitope is also ELA-A1-restricted, but the molecule that presents Rev-QW11 is unknown. To determine whether functionally distinct class I molecules present ELA-A1-restricted CTL epitopes, we sequenced and expressed MHC class I genes from three ELA-A1 horses. Two horses had the *7-6* allele, which when expressed, presented Env-RW12, Gag-GW12, and Rev-QW11 to CTL. The other horse had a distinct allele, designated *141*, encoding a molecule that differed from 7-6 by a single amino acid within the α-2 domain. This substitution did not affect recognition of Env-RW12, but resulted in more efficient recognition of Rev-QW11.

Significantly, CTL recognition of Gag-GW12 was abrogated, despite Gag-GW12 binding to 141. Molecular modeling suggested that conformational changes in the 141/Gag-GW12 complex led to a loss of TCR recognition. These results confirmed that the ELA-A1 haplotype is comprised of functionally distinct alleles, and demonstrated for the first time that naturally occurring MHC class I molecules that vary by only a single amino acid can result in significantly different patterns of epitope recognition by lentivirus-specific CTL.

INTRODUCTION

Infections by lentiviruses induce virus-specific CTL responses that are critical for control of viral load and clinical disease. Specifically, Env, Gag, and Rev proteins are important targets for CTL in HIV-1-infected individuals (1, 2, 3, 4). Sustained HIV-1 Gag-specific CTL responses are associated with very low numbers of infected $CD4^+$ T cells and stable $CD4^+$ T cell counts in long-term nonprogressors, whereas loss of Gag-specific CTL is associated with clinical progression to AIDS (5). Moreover, HIV-1 Gag-specific CTL responses and frequency are inversely correlated with viral load (6, 7, 8), and high frequencies of Gag-specific CTL are significantly associated with slower disease progression (9). In addition, high levels of HIV-1 Env- and Gag-specific memory CTL are strongly associated with low viral load and lack of disease in long-term nonprogressors (10), and CTL responses directed against the early expressed viral regulatory protein Rev inversely correlate with rapid HIV-1 disease progression (11, 12).

Equine infectious anemia virus (EIAV) is a macrophage-tropic lentivirus that causes persistent infections in horses worldwide. In contrast to HIV-1 infection however, most EIAV-infected horses eventually control plasma viremia and clinical disease and remain life-long inapparently infected carriers (13, 14, 15). The initial as well as the long-term control of viremia and clinical disease is the result of adaptive immune responses, including

neutralizing Ab and importantly, CTL (16, 17, 18, 19, 20, 21, 22). Due to these robust viral-specific immune responses that contain viral replication, EIAV infection in horses is a unique and useful model system for the study of lentiviral immune control.

As in HIV-1, Env, Gag, and Rev proteins are important CTL targets in EIAV-infected horses. EIAV Env- and Gag-specific CTL are detected during acute and inapparent infection (17, 23, 24), with Gag-specific CTL responses occurring in the majority of horses tested (25, 26). Although EIAV Env-specific CTL can be immunodominant and may be important in early viral control, viral escape can limit the effectiveness of CTL directed against variable Env epitopes, and Gag-specific CTL frequency can correlate with the ability to control viral load and clinical disease (27, 28). Epitope clusters occur within EIAV Gag proteins that are recognized by CTL (including high-avidity CTL) in horses with disparate MHC class I (MHC I) haplotypes and are likely important in control of clinical disease (29, 30). Additionally, moderate and high-avidity Rev-specific CTL are associated with control of viremia and disease in EIAV-infected nonprogressor horses (27). Significantly, no viral escape from high-avidity Rev-specific CTL has been observed. The Rev epitope recognized by these CTL is highly conserved among other strains of EIAV, and independent studies by other investigators show that this sequence does not change during long-term EIAV infection (31, 32).

Given that CTL epitopes in Env, Gag, and Rev are critical for lentiviral immune control, factors that affect the presentation and recognition of these peptides by CTL in a population of infected individuals are important considerations for vaccine development. One of the most important factors is MHC I polymorphism, because CTL recognize peptide epitopes only in the context of allelic forms of MHC I molecules. The CTL TCR binds to a cell surface MHC I complex consisting of the MHC I molecule, a peptide processed from a viral protein, and β_2-microglobulin (β_2m) (33). The HLA classical MHC I genes are the products of three loci, and as of July 2006, there were 478 HLA-A, 805 HLA-B, and 256

HLA-C named alleles (http://www.ebi.ac.uk/imgt/hla/intro.html) (34). MHC I polymorphism occurs primarily in the residues that form the peptide-binding domain, therefore influencing the types of peptides that are selectively bound. Similarities in peptide-binding specificity between human MHC I molecules have been identified based on the structure of peptide-binding pockets, peptide-binding assays, and analysis of motifs, and sets of molecules with similar specificities are called supertypes (35, 36, 37, 38). Despite the fact that different MHC I molecules can belong to a supertype and have similar peptide-binding specificities, differences in only a few residues among class I molecules can significantly affect the recognition of the MHC I-peptide complex by CTL. Saturation mutagenesis of a murine MHC I molecule has shown that a single amino acid substitution in the α-1 domain can result in more effective killing of lymphocytic choriomeningitis virus-infected target cells by lymphocytic choriomeningitis virus-specific CTL (39). Importantly, the HLA-B*35 subtypes HLA-B*3503 and HLA-B*3502 (B*35-Px genotype) correlate more strongly with rapid progression to AIDS than does the HLA-B*3501 subtype (B*35-PY genotype), which varies from HLA-B*3503 and HLA-B*3502 by only one and two amino acids, respectively, in the peptide-binding domain (40). It is presumed that this effect is due to differential presentation of epitopes to HIV-1-specific CTL, and it has been demonstrated that a higher frequency of Gag-specific CTL correlate with lower viral loads in individuals with the B*35-PY genotype, whereas no significant relationship exists between CTL activity and viral load in the B*35-Px group (41).

Compared with humans, less is known regarding the numbers and locus assignments of equine MHC I alleles. Serology has been the most widely used MHC I typing method in horses, with 17 equine leukocyte Ag (ELA)-A haplotypes defined (42, 43, 44, 45). More recently, classical MHC I alleles have been identified and sequenced, but it has not been possible to assign alleles to loci (46, 47, 48, 49, 50, 51, 52). Recent work indicates that up to three or four (or more) classical MHC I loci exist, and it is likely that the number of loci is

variable and dependent on haplotype (48, 52). Regardless, we have identified the classical equine MHC I molecule 7-6, which is associated with the serologically defined ELA-A1 haplotype (51). The 7-6 molecule presents Env and Gag epitopes (Env-RW12 and Gag-GW12) to distinct populations of EIAV-specific CTL (27, 51). Target cells from some ELA-A1 horses present both epitopes, whereas target cells from other ELA-A1 horses present only Env-RW12 (51), suggesting that subtypes that comprise functionally distinct alleles exist within the ELA-A1 haplotype. In addition, a CTL epitope in EIAV Rev (Rev-QW11) is also restricted by the ELA-A1 haplotype (27), but the molecule that presents Rev-QW11 has not been identified. The purpose of the present study was to determine whether different horses sharing the ELA-A1 haplotype used functionally distinct MHC I molecules to present ELA-A1-restricted CTL epitopes. To test this hypothesis, we sequenced and expressed MHC I genes from three different horses with the ELA-A1 haplotype and determined how these different molecules affected the recognition of important Env, Gag, and Rev epitopes by EIAV-specific CTL.

MATERIALS AND METHODS

Horses

Arabian horses A2140, A2150, and A2152 were used in this study. A2152 is a noninfected 7-year-old breeding stallion in which the 7-6 MHC I allele was initially identified (51). A2140 and A2150 are 8- and 7-year-old mares, respectively, that have been infected with $EIAV_{WSU5}$ for 7 and 6 years, respectively (27). The ELA-A haplotypes were determined serologically by lymphocyte microcytotoxicity (42, 53, 54) using reagents provided by Dr. E. Bailey (University of Kentucky, Lexington, KY). All experiments

involving horses were approved by the Washington State University Institutional Animal Care and Use Committee.

Identification of equine MHC I alleles

The full-length *141* gene (GenBank accession no. AY374512) from A2150 was cloned and sequenced as previously described (51). Briefly, PBMC were isolated and cultured for 48 h in RPMI 1640 medium with 10% FBS, 2 mM L-glutamine, and 10 μg/ml gentamicin. Con A (40 μg/ml) was added to up-regulate MHC I expression. mRNA isolated from these cells was used for first-strand synthesis, which was primed with 1 μg of oligo$(dT)_{12-18}$ and 50 ng of random hexamers. This reaction was incubated at an initial temperature of 37°C for 15 min, followed by an additional hour at 49°C. The second-strand reaction was incubated at 16°C for 2 h using *Escherichia coli* DNA ligase, polymerase, and RNase H. cDNA was blunt-ended with T4 DNA polymerase, *Eco*RI adapters were added with T4 ligase, and the adapters were phosphorylated with T4 polynucleotide kinase. The cDNA was then size fractionated, and cDNA between 2 and 3 kb was ligated into EcoRI-digested and dephosphorylated pcDNA3 vector (Invitrogen Life Technologies). *E. coli* electrocompetent cells (Invitrogen Life Technologies) were transformed with this ligation mixture resulting in a cDNA library. Clones were selected from the library by colony-lift hybridization with a ^{32}P-labeled *Hind*III-*Xba*I fragment from horse MHC I gene *8/9* (provided by Dr. D. Antczak, Cornell University, Ithaca, NY) (46). Positive colonies were isolated, and those with inserts of the correct size were sequenced. Sequencing was performed at the Laboratory for Biotechnology and Bioanalysis (Washington State University) using dye-labeled dideoxynucleotide-cycle sequencing with an ABI 377 automated sequencer.

The presence of the *7-6* allele (GenBank accession no. AY225155) was confirmed in A2140 using a RT-PCR as previously described (48) with modifications. Briefly, total RNA was isolated from 2 x 10^7 equine kidney (EK) cells with an RNeasy mini kit (Qiagen). Ten microliters of RNA and a SuperScript One-Step RT-PCR kit (Invitrogen Life Technologies) was then used with 1 μM each of forward (5'-ATG ATG CCC CCA ACC TTC-3') and reverse (5'-TGA ACA AAT CTT GCA TCA CTT G-3') primers. The reaction conditions were 45°C for 50 min, followed by 94°C for 2 min, then 35 cycles of 94°C for 30 s, 53°C for 30 s, and 72°C for 1 min. The 1116-bp product was gel isolated and extracted (Qiagen), then used for TOPO TA cloning into the pCR4-TOPO vector for sequencing (Invitrogen Life Technologies). Isolated colony minipreps were digested and electrophoresed to screen for inserts and those with inserts of correct size were sequenced by the Laboratory for Biotechnology and Bioanalysis using dye-labeled dideoxynucleotide-cycle sequencing with an ABI 377 automated sequencer.

Although we independently named the alleles in this study, official Immuno Polymorphism Database nomenclature (http://www.ebi.ac.uk/ipd/) will soon be available for equine MHC alleles (55).

Expression of equine MHC I alleles

Two retroviral vectors were constructed as described (51) using the plasmid pLXSN (provided by Dr. A. Dusty Miller, Fred Hutchinson Cancer Research Center, Seattle, WA). The MHC I genes *7-6* and *141* were PCR amplified and ligated into the cloning site of pLXSN downstream of the Moloney murine sarcoma virus long terminal repeat and upstream of the neomycin phosphotransferase gene, which was under the control of the SV40 early promoter (56). The sequences of the inserts and flanking plasmid DNA were determined. To generate vector-producing cell lines, published procedures were used (51, 57). Briefly, an

amphotropic packaging cell line, PA317 (CRL-9078; American Type Culture Collection) was transfected with each plasmid. Supernatant from PA317 cells was used to transduce amphotropic PG13 packaging cells (CRL-10686; American Type Culture Collection), which were then selected using 750 μg/ml G-418 sulfate (Invitrogen Life Technologies). Vectors were harvested from the selected PG13 cells and used to transduce CTL target cells. EK cells and human mutant B lymphoblastoid 721.221 cells (58) were transduced with the retroviral vectors expressing *7-6* and *141*, pulsed with peptides, and used as CTL targets (51). The 721.221 cells (obtained from Dr. A. Sette, La Jolla Institute for Allergy and Immunology, San Diego, CA) express human β_2m, but not HLA-A, -B, or -C class I molecules (58).

PBMC stimulations and CTL assays

PBMC stimulations and CTL assays were performed as described (17, 27, 28, 51) with modifications. Briefly, PBMC were isolated from A2140 and A2150 and stimulated with peptide-pulsed autologous monocytes. EK target cells from A2140 and A2150 and mixed-breed pony H585 were established from kidney tissue obtained by biopsy (17). For stimulation with peptides, 2 μM Env-RW12, Gag-GW12, or Rev-QW11 was added to PBMC in 10% FBS. Peptide and PBMC were incubated for 2 h at 37°C with occasional mixing before centrifugation at 250 x g for 10 min. PBMC were resuspended to 2 x 10^6/ml in RPMI 1640 medium with 10% FBS, 20 mM HEPES, 10 μg/ml gentamicin, and 10 μM 2-ME. One milliliter was added to each well of a 24-well plate and incubated for 1 wk at 37°C before use in CTL assays. CTL activity was measured using a ^{51}Cr release assay with a 17-h incubation period using EK target cells (17, 27, 28). In addition, human B lymphoblastoid 721.221 target cells transduced with retroviral vectors expressing equine MHC I genes *7-6* or *141* as described above were used in assays with a 5-h incubation period (51). The shorter incubation period was used because 721.221 target cells are less hardy than EK target cells

and they develop spontaneous lysis sooner (51). Target cells were pulsed with various amounts of peptide Env-RW12, Gag-GW12, and Rev-QW11 as indicated in the figures. The formula, percent-specific lysis = $[(E - S)/(M - S)] \times 100$, was used, where E is the mean of three test wells, S is the mean spontaneous release from three target cell wells without effector cells, and M is the mean maximal release from three target cell wells with 2% Triton X-100 in distilled water. The E:T ratio was 20:1 or 50:1 as indicated in the figures, and each well contained ~30,000 target cells. The 50:1 E:T ratio was used to confirm 20:1 E:T ratio results. Comparisons were only made between assays that used the same E:T ratio. Previous work indicates that these E:T ratios yield consistent results corresponding to the log portion of the killing curve, and that the 50:1 E:T ratio results in the highest percent-specific lysis (17, 19, 24, 25, 26, 27, 28, 29, 30, 51, 57, 59, 60, 61, 62). Assays with similar constant E:T ratios have been used by others (63). Only assays with a spontaneous target cell lysis of <30% were used. The SE of percent-specific lysis was calculated using a formula that accounts for the variability of E, S, and M (64). Significant lysis was defined as the percent-specific lysis of peptide-pulsed target cells that was >10% and also >3 SE above the nonpulsed target cells or above target cells transduced with control vectors and pulsed with the relevant peptide. For comparisons of CTL recognition efficiency, the peptide concentration that resulted in 50% maximal target cell-specific lysis (EC_{50}) was used. The EC_{50} was calculated after transforming the percent-specific lysis data to percent-maximal lysis (with the lowest percent-specific lysis value set to 0% and the highest percent-specific lysis value set to 100%) and fitting the curve with nonlinear regression using GraphPad Prism version 3.03 (GraphPad). This is an established method to measure and compare CTL recognition efficiency and avidity (27, 30, 65, 66). All CTL assays were performed at least twice, and the results were consistent in each case.

Live cell peptide-binding assay

Peptide binding to equine MHC I molecules 7-6 and 141 was measured as previously described (51), with slight modifications, using the chloramine-T method (67, 68). Briefly, 721.221 cells transduced with MHC I gene *7-6, 141*, or with a retroviral vector that did not express an equine gene, were preincubated with human β_2m. The cells were washed and resuspended in RPMI 1640 containing β_2m, EDTA, PMSF, and *Nα-p-* tosyl-L-lysine chloromethyl ketone. One hundred microliters containing 2 x 10^6 cells plus 1 µl containing 1.5 x 10^5 cpm of ^{125}I -labeled Env-RW12 was incubated for 4 h at 22°C in wells of a 96-well U-bottom plate. Cells were then washed three times with serum-free medium, centrifuged through calf serum to remove any remaining unbound radiolabeled peptide, and then washed a final time. The radioactivity of the cell pellet was counted with a gamma scintillation counter (Packard Instrument). Competitive inhibition assays were performed twice in triplicate by adding 10 µl containing sufficient unlabeled peptide competitors to result in final concentrations of 1–1000 nM to the initial mixture of cells before addition of radiolabeled Env-RW12 peptide. Competing peptides were unlabeled Env-RW12, Gag-GW12, Rev-QW11, and control peptide 1b4a (VRVEDVTNTAEY), which does not inhibit the binding of Env-RW12 to 7-6 (51). For each competing peptide, the concentration resulting in 50% inhibition of radiolabeled Env-RW12 binding (IC_{50}) was calculated by fitting the curve with nonlinear regression using GraphPad Prism version 3.03. All peptides used in binding assays had 95% purity and were synthesized by Sigma-Genosys.

Molecular modeling

Because crystal structures of equine MHC I molecules are not available, three-dimensional computer models of 7-6 and 141 were generated based on known structures of human MHC I molecules using MODELLER 8v2 (http://www.salilab.org/modeler) (69).

Templates for modeling were chosen based on PSI-BLAST searches of the Brookhaven Protein Data Bank database. The 1XR9A structure was used for 7-6 and the 1ZSDA structure (70) was used for 141. The models were verified using VERIFY3D (http://nihserver.mbi.ucla.edu/Verify_3D) (71). Models of the Env-RW12, Gag-GW12, and Rev-QW11 peptides were also generated. To predict the side chain conformations for each peptide, SCWRL3.0 (http://dunbrack.fccc.edu/SCWRL3.php) (72) was used. Viral peptides bound to human MHC I molecules served as templates and were chosen from the Brookhaven Protein Data Bank database (1ZHKC for Env-RW12 and Gag-GW12, and 1ZSDC for Rev-QW11). The 1ZSDC structure is an 11-mer peptide (70), like Rev-QW11. Because no structures of 12-mer peptides bound to MHC I molecules were available, the first residue (L) of 1ZHKC, a 13-mer peptide (73), was eliminated before use as the backbone template for Env-RW12 and Gag-GW12. To model the binding of the three peptides to 7-6 and 141, the docking program FTDock2.0 (http://www.bmm.icnet.uk/docking) (74, 75, 76) was used, and the docking score (RPScore) for each complex was determined (76). The interacting residues for each complex were identified, and the interactions by category (hydrophobic, salt bridges, repulsive charged, hydrogen bonds, and aromatic stacking) between atoms of contact residues were determined using the STING Millennium Suite program (http://trantor.bioc.columbia.edu/SMS/index_m.html) (77, 78). Finally, the models were visualized using the PyMOL Molecular Graphics System (DeLano Scientific, http://www.pymol.org).

RESULTS

Target cells from A2140 and A2152 pulsed with Env-RW12, Gag-GW12, and Rev-QW11 were recognized differently by CTL than target cells from A2150

Horses A2140, A2152, and A2150 all had the ELA-A1 haplotype as determined serologically, inheriting ELA-A1 from unrelated dams 172, 162, and 169, respectively (Table 4.1).

Horse	ELA-A Haplotype	Dam	Dam ELA-A Haplotype[a]	Sire	Sire ELA-A Haplotype
A2140	A1/w11	172	A1	Sire B	A4/w11
A2152	A1/A4	162	A1	Sire B	A4/w11
A2150	A1/w11	169	A1/A4	Sire B	A4/w11

Table 4.1. Pedigree of ELA-A1 horses

[a] ELA-A haplotypes were determined serologically by lymphocyte microcytotoxicity. It was not known whether horses with one haplotype were homozygous or heterozygous with a haplotype not recognized by available antisera.

Our previous work indicates that EK target cells from A2140 and A2152 present both the Env-RW12 (RVEDVTNTAEYW) and Gag-GW12 (GSQKLTTGNCNW) epitopes to Env-RW12- and Gag-GW12-specific A2140 CTL (27, 51). Although EK target cells from A2150 present Env-RW12 to Env-RW12-specific A2140 CTL, they do not present Gag-GW12 to Gag-GW12-specific A2140 CTL (51). Because the 7-6 MHC I molecule identified in A2152 presents both Env-RW12 and Gag-GW12 (51), it was likely that a different MHC I molecule presented Env-RW12 but not Gag-GW12 in A2150.

The Rev-QW11 (QAEVLQERLEW) epitope is recognized by CTL from A2150 (27). Although EK target cells from all three horses were capable of presenting Rev-QW11 to Rev-QW11-specific A2150 CTL, these CTL recognized A2150 EK targets more efficiently than A2140 and A2152 targets (Fig. 4.1a). This observation suggested that the MHC I molecule presenting Rev-QW11 in A2150 was different from the one presenting Rev-QW11 in A2140 and A2152.

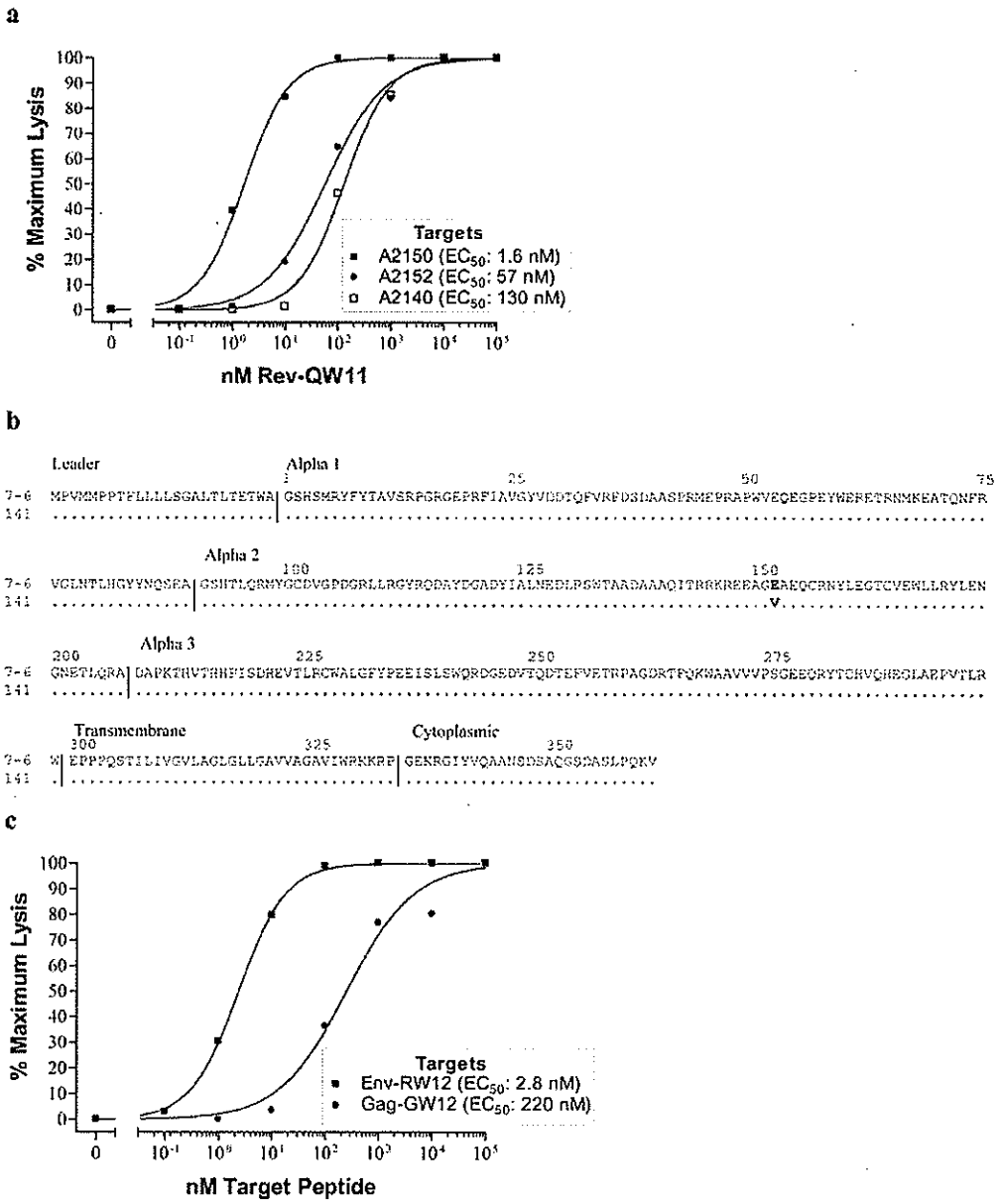


Figure 4.1. Differential CTL recognition efficiencies and sequences of MHC I alleles. **(a)** CTL recognized Rev-QW11-pulsed A2150 EK target cells more efficiently than A2152 and A2140 EK target cells. A2150 CTL were stimulated with Rev-QW11 peptide and percent-specific lysis was determined on A2150, A2152, and A2140 EK targets pulsed with increasing concentrations of Rev-QW11 peptide. E:T cell ratio was 20:1. EC_{50}, Peptide concentration resulting in 50% maximal-specific lysis. Actual minimum and maximum percent-specific lysis for A2150, A2152, and A2140 targets was 7.1 and 49.3, 4.9 and 32.5, and 2.7 and 34.7, respectively. **(b)** Equine MHC I molecules 7-6 and 141 differed in only one amino acid in the α-2 domain. Amino acid sequences of 7-6 and 141 are shown with domains (46, 49) indicated. The E->V substitution at position 152 is shown in bold. **(c)** CTL recognized Env-RW12 more efficiently than Gag-GW12 when presented by equine MHC I molecule 7-6. A2140 CTL were stimulated with Env-RW12 or Gag-GW12 peptides and percent-specific lysis was determined on 7-6-transduced 721.221 cells pulsed with increasing concentrations of Env-RW12 or Gag-GW12 peptides. E:T cell ratios were 20:1. Actual minimum and maximum percent-specific lysis for Env-RW12 and Gag-GW12 targets was 0 and 43.8, and 2.1 and 21.9, respectively.

Identification of MHC class I alleles in A2140 and A2150

Because the 7-6 molecule that presents Env-RW12 and Gag-GW12 occurs in A2152 (51), and because Env-RW12- and Gag-GW12-specific CTL display similar recognition of A2152 and A2140 EK target cells (51), it was hypothesized that A2140 also possessed the *7-6* allele. Therefore, RT-PCR was used to amplify MHC I genes from A2140 PBMC. Cloning and sequencing confirmed the presence of the *7-6* allele in A2140. Of the 21 isolates processed, 5 were copies of a pseudogene, 3 were other pseudogenes, 6 were copies of a classical gene, and 7 were other classical genes, which included *7-6* (data not shown).

Because previous work indicates that A2150 EK target cells are recognized differently by Env-RW12- and Gag-GW12-specific CTL (51), and because Rev-QW11-specific CTL recognized A2150 EK target cells more efficiently than A2140 and A2152

targets, it was hypothesized that a MHC I molecule distinct from 7-6 presented these epitopes in horse A2150. A previous study identified partial sequences for three MHC I alleles in A2150, one of which, designated *141*, shared the *7-6* sequence except for a single amino acid difference encoded at position 152 (E→V) in the α-2 domain (48). Due to its sequence similarity to 7-6, it was of interest to determine whether 141 had similar functional characteristics. To obtain the full-length *141* gene for expression, a cDNA library from A2150 was screened for MHC I genes and *141* was subsequently identified by sequencing (Fig. 4.1b).

Previous work indicated that the *141* allele is not present in A2152 and that the *7-6* allele is not present in A2150 (48). In addition, allele *141* was not identified among sequenced MHC I clones in A2140 (data not shown). Therefore, the 141 class I molecule was a likely candidate for presenting Env-RW12 and Rev-QW11 but not Gag-GW12 in horse A2150.

CTL recognized Env-RW12 more efficiently than Gag-GW12 when presented by the 7-6 molecule, and also recognized Rev-QW11 presented by 7-6

PBMC from A2140 were stimulated separately with Env-RW12 and Gag-GW12 peptides, then assayed for CTL activity on 7-6-transduced human 721.221 target cells that were pulsed with increasing concentrations of Env-RW12 and Gag-GW12 peptides, respectively. Human mutant B lymphoblastoid 721.221 cells express human β_2m, but not HLA-A, -B, or -C class I molecules (58). Env-RW12 was recognized by CTL more efficiently than Gag-GW12, with 50% maximal lysis (EC_{50}) of 2.8 nM for Env-RW12 vs 220 nM for Gag-GW12 (Fig. 4.1c).

Initial assays indicated that recognition of *7-6*-transduced 721.221 cells pulsed with Rev-QW11 by A2150 Rev-QW11-specific CTL was equivocal (Fig. 4.2a). Because it was

possible that the absence of equine β_2m on 721.221 cells contributed to poor recognition of Rev-QW11 by CTL, ELA-A mismatched EK cells from pony H585 (ELA-A6) were transduced with *7-6*, pulsed with Rev-QW11, and used as targets. Results indicated that in addition to Env-RW12 and Gag-GW12, 7-6 also presented Rev-QW11 (Fig. 4.2b).

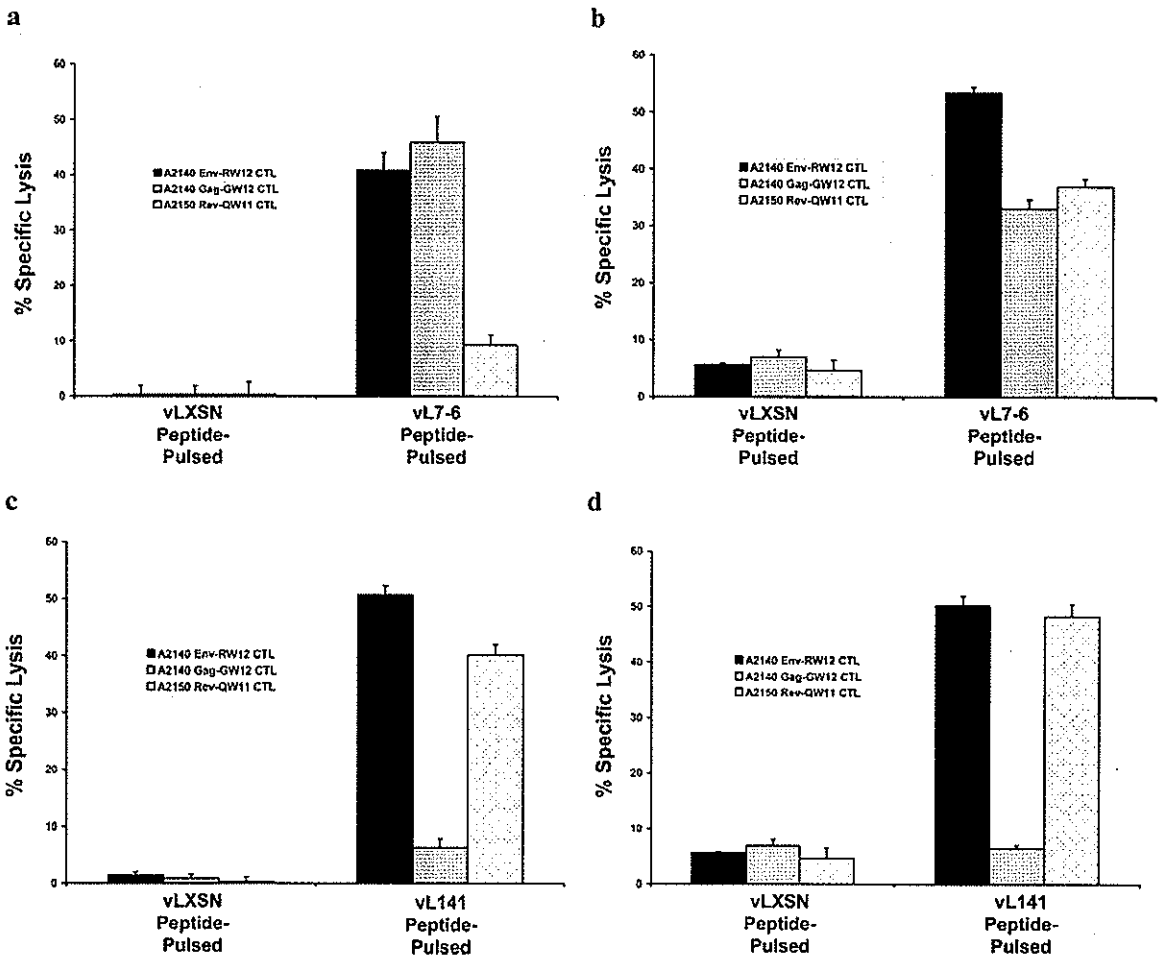


Figure 4.2. Presentation of CTL epitopes by equine MHC I molecules 7-6 and 141. **(a)** and **(b),** MHC I molecule 7-6 presented Env-RW12, Gag-GW12, and Rev-QW11 to CTL. A2140 Env-RW12-, A2140 Gag-GW12-, and A2150 Rev-QW11-stimulated CTL on *7-6*-transduced 721.221 cells **(a)** pulsed with 104 nM of the corresponding peptide and on *7-6*-transduced H585 EK cells **(b)** pulsed with 104 nM of the corresponding

peptide. **(c)** and **(d)**, CTL recognized Env-RW12 and Rev-QW11 when presented by equine MHC I molecule 141, but did not recognize Gag-GW12-pulsed targets expressing 141. A2140 Env-RW12-, A2140 Gag-GW12-, and A2150 Rev-QW11-stimulated CTL on *141*-transduced 721.221 cells **(c)** pulsed with 10^4 nM of the corresponding peptide and on *141*-transduced H585 EK cells **(d)** pulsed with 10^4 nM of the corresponding peptide. a–d, vLXSN, empty vector. Error bars are SE for the assay shown, derived as described in Materials and Methods. E:T ratio is 50:1.

The 141 molecule presented Env-RW12 and Rev-QW11 to CTL, but CTL failed to recognize Gag-GW12 on 141-expressing target cells

A retroviral vector containing the *141* gene was constructed and used to transduce 721.221 cells and H585 EK cells. Rev-QW11-specific CTL from A2150 recognized both *141* -transduced 721.221 and *141*-transduced H585 EK target cells pulsed with the Rev-QW11 peptide (Fig. 4.2, c and d). Similarly, Env-RW12-specific CTL from A2140 recognized both 141-transduced 721.221 and *141*-transduced H585 EK target cells pulsed with Env-RW12 (Fig. 4.2, c and d). In contrast, Gag-GW12-specific CTL from A2140 failed to recognize both 141-transduced 721.221 and *141*-transduced H585 EK target cells pulsed with Gag-GW12 (Fig. 4.2, c and d).

CTL recognized Env-RW12 with similar efficiency when presented by the 7-6 and 141 molecules, whereas CTL recognized Rev-QW11 more efficiently when presented by 141

PBMC from horse A2140 were stimulated with Env-RW12, then assayed for CTL activity on *7-6*- and *141*-transduced 721.221 target cells that were pulsed with increasing concentrations of Env-RW12. The efficiency of Env-RW12 recognition on *7-6*-transduced targets (EC_{50}: 15 nM) was similar to that on *141*-transduced targets (EC_{50}: 8.8 nM) (Fig. 4.3a).

Because *7-6*-transduced 721.221 cells presented Rev-QW11 poorly, MHC I-mismatched H585 EK target cells were used to compare the efficiency of Rev-QW11 recognition when presented by the 7-6 and 141 molecules. Following Rev-QW11 stimulation, A2150 CTL recognized Rev-QW11-pulsed *141*-transduced H585 EK targets more efficiently (EC_{50}: 0.68 nM) than Rev-QW11-pulsed *7-6*-transduced H585 EK targets (EC_{50}: 40 nM) (Fig. 4.3b). These results were consistent with those obtained using A2140, A2152, and A2150 EK targets cells (Fig. 4.1a), and confirmed that in addition to abrogating the recognition of Gag-GW12, the single $^{152}E \rightarrow V$ amino acid substitution in class I molecule 141 increased the recognition efficiency of Rev-QW11 by CTL.

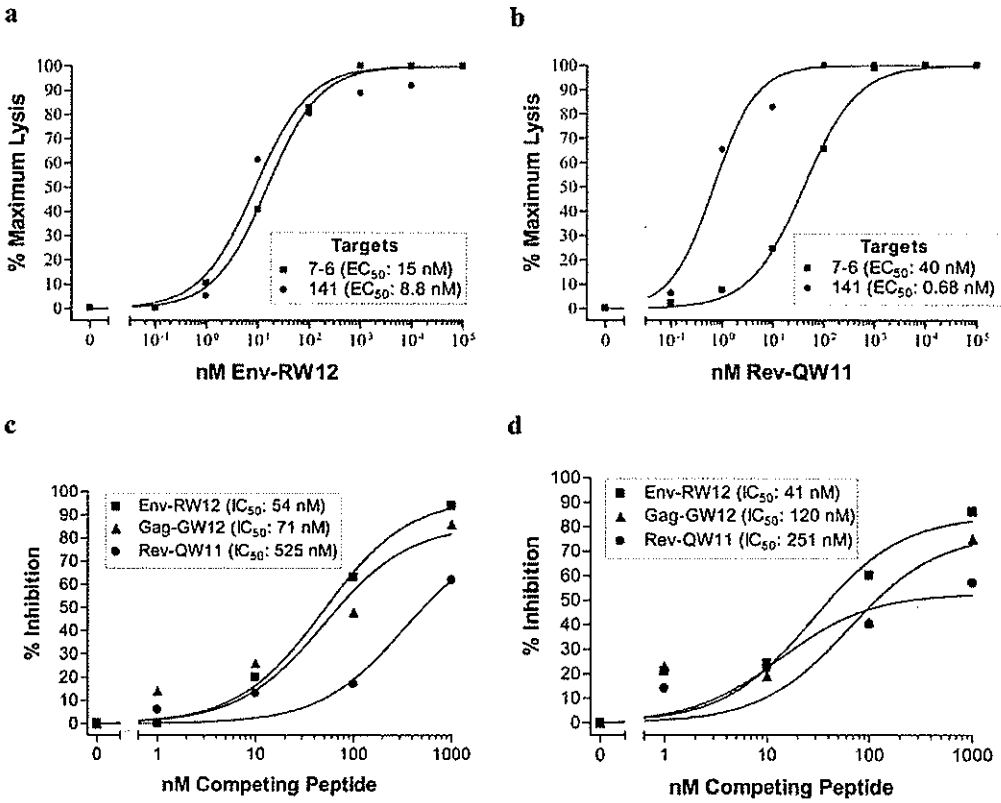


Figure 4.3. Env-RW12- and Rev-QW11-specific CTL recognition efficiencies and competitive peptide-binding inhibition. **(a)** CTL recognized Env-RW12 with similar efficiency when presented by equine MHC I molecules 7-6 and 141. A2140 CTL were stimulated with Env-RW12 peptide and percent-specific lysis was determined on *7-6*- and *141*-transduced 721.221 cells pulsed with increasing concentrations of Env-RW12 peptide. E:T cell ratios were 20:1. EC_{50}, peptide concentration resulted in 50% maximal-specific lysis. Actual minimum and maximum percent-specific lysis for 7-6 and 141 targets was 0 and 63.5, and 0 and 49.1, respectively. **(b)** CTL recognized Rev-QW11 more efficiently when presented by equine MHC I molecule 141 than when presented by 7-6. A2150 CTL were stimulated with Rev-QW11 peptide and percent-specific lysis was determined on 7-6- and 141-transduced H585 EK cells pulsed with increasing concentrations of Rev-QW11 peptide. E:T cell ratios were 20:1. Actual minimum and maximum percent-specific lysis for 7-6 and 141 targets was 4.0 and 38.1, and 7.8 and 53.4, respectively. c and d, Unlabeled Env-RW12, Gag-GW12, and Rev-QW11 inhibited ^{125}I-labeled Env-RW12 binding to live 721.221 cells expressing MHC I molecule 7-6 **(c)**, and live 721.221 cells expressing MHC I molecule 141 **(d)**. IC_{50}, peptide concentration resulted in 50% inhibition of radiolabeled Env-RW12 binding.

Env-RW12 and Gag-GW12 bound to 7-6 and 141 with higher affinity than Rev-QW11

Live-cell Env-RW12 peptide-binding inhibition experiments were performed to determine the relative binding affinities of the three peptides to 7-6 and 141. The Env-RW12 peptide was chosen for ^{125}I labeling because CTL recognized Env-RW12 on 7-6- and 141-expressing targets with similar efficiency (Fig. 4.3a), suggesting that Env-RW12 bound 7-6 and 141 with similar affinity. Moreover, Env-RW12 was the only peptide with a tyrosine residue, necessary for the chloramine-T method used for ^{125}I labeling (68). The relative binding affinities of each peptide were then determined by using unlabeled peptides in

competitive binding inhibition assays. Binding of ^{125}I -labeled Env-RW12 to 7-6 was more efficiently inhibited by unlabeled Env-RW12 (IC_{50}: 54 nM) than by Gag-GW12 (IC_{50}: 71 nM) or Rev-QW11 (IC_{50}: 525 nM) (Fig. 4.3c). For 7-6 binding, the IC_{50} of the negative control peptide 1b4a was >1000 nM (percent inhibition caused by 1000 nM was 27%; data not shown). These results were in agreement with the CTL recognition data.

Binding of ^{125}I -labeled Env-RW12 to 141 was efficiently inhibited by unlabeled Env-RW12 (IC_{50}: 41 nM) (Fig. 4.3d), consistent with the observation for 7-6. Surprisingly, binding of Env-RW12 to 141 was also inhibited by Gag-GW12 (IC_{50}: 120 nM). Thus, the lack of Gag-GW12-specific CTL recognition of 141-expressing target cells was not due to the inability of Gag-GW12 to bind 141. However, the IC_{50} values indicated that Gag-GW12 bound 141 with lower affinity than 7-6 (120 nM vs 71 nM). Also unexpectedly, Rev-QW11 bound to 141 with lower affinity (IC_{50}: 251 nM) than Gag-GW12. Consistent with the CTL results however, Rev-QW11 bound to 141 with higher affinity than it did to 7-6 (251 nM vs 525 nM). For 141 binding, the IC_{50} of the negative control peptide 1b4a was >1000 nM (percent inhibition caused by 1000 nM was 33%; data not shown). Taken together, these experiments suggested that differences in MHC/peptide-binding affinity were not sufficient to explain the differential CTL recognition of Gag-GW12 and Rev-QW11 on 7-6- and 141-expressing target cells.

Computer modeling and docking of peptides with the 7-6 and 141 molecules

Because crystal structures are not available, three-dimensional molecular modeling was performed to determine the possible structural and functional effects of the ^{152}E→V substitution. Computer models were generated for 7-6 and 141 and the Env-RW12, Gag-GW12, and Rev-QW11 peptides. A docking algorithm was then used to dock each of the three peptides with 7-6 and 141, and docking scores for each complex were calculated.

Modeling each of the MHC-peptide complexes indicated that amino acid position 152 occurred in the α-2 helix of 7-6 and 141, within the wall of the peptide-binding cleft, and it appeared that the conformations of the bound peptides were affected differently (Fig. 4.4). Modeling suggested that the peptides bound 7-6 and 141 in a bulged conformation, with the first and last residues of each peptide anchored in the binding clefts (Fig. 4.5). The conformation of Env-RW12 was similar when bound to 7-6 and 141 (Fig. 4.5a), whereas Rev-QW11 was slightly more bulged when bound to 7-6, presumably because of the W11 residue binding less deeply in the peptide-binding cleft of 7-6 (Fig. 4.5b). Interestingly, the bound conformation of Gag-GW12 was shifted and more sharply bulged in the 141 complex, apparently because the G1 residue of Gag-GW12 bound less deeply in the cleft of 141 as compared with 7-6 (Fig. 4.5c). Based on an analysis of pair potentials at the interface (76), the docking algorithm predicted the 7-6/Env-RW12 complex as the most favorable (highest RPScore docking score), followed by the 141/Rev-QW11 and 141/Env-RW12 complexes (Table 4.2). The 7-6/Gag-GW12 and 7-6/Rev-QW11 complexes were less favorable and the 141/Gag-GW12 complex was the least favorable, although the docking score differences between these latter three complexes were not profound (Table 4.2). Based on the MHC/peptide-docking models, the 7-6/Env-RW12 complex had a greater number of hydrophobic, salt bridge, hydrogen bond, and aromatic stacking interactions than did the 141/Env-RW12 complex (Table 4.2 and Fig. 4.6). For the 7-6/Rev-QW11 and 141/Rev-QW11 complexes, there were more hydrophobic, hydrogen bond, and aromatic stacking interactions for 7-6/Rev-QW11, but 141/Rev-QW11 had a greater number of salt bridges (Table 4.2 and Fig. 4.6). Importantly, the 7-6/Rev-QW11 complex had two destabilizing repulsive charged interactions that were absent in the 141/Rev-QW11 complex. For the 7-6/Gag-GW12 and 141/Gag-GW12 complexes, the two salt bridges in 7-6/Gag-GW12 were absent in 141/Gag-GW12, and 141/Gag-GW12 had one less hydrogen bond (Table 4.2 and Fig. 4.6). The salt bridges present in 7-6/Gag-GW12 but absent in 141/Gag-GW12 occurred

between the ^{69}E residue of 7-6 and the K4 residue of Gag-GW12 (Fig. 4.6, c and d, and Fig. 4.7, a and b). In addition, the G1 residue of Gag-GW12 had more interactions with residues of 7-6 (^{7}Y: two hydrophobic, one hydrogen bond; ^{99}Y: two hydrophobic; ^{159}Y: three hydrophobic, two hydrogen bonds) than it did with residues of 141 (^{159}Y: three hydrophobic) (Fig. 4.6, c and d, and Fig. 4.7, c and d). Based on the numbers of interactions between contact residues in the docking models for each of the six MHC/peptide complexes, the first two residues and W12 were probable anchor residues for Env-RW12 and Gag-GW12, whereas the first three residues and W11 were probable anchor residues for Rev-QW11 (Fig. 4.6). In general, the molecular modeling results supported the experimental data and suggested specific mechanisms for the observed differences in peptide binding and CTL recognition.

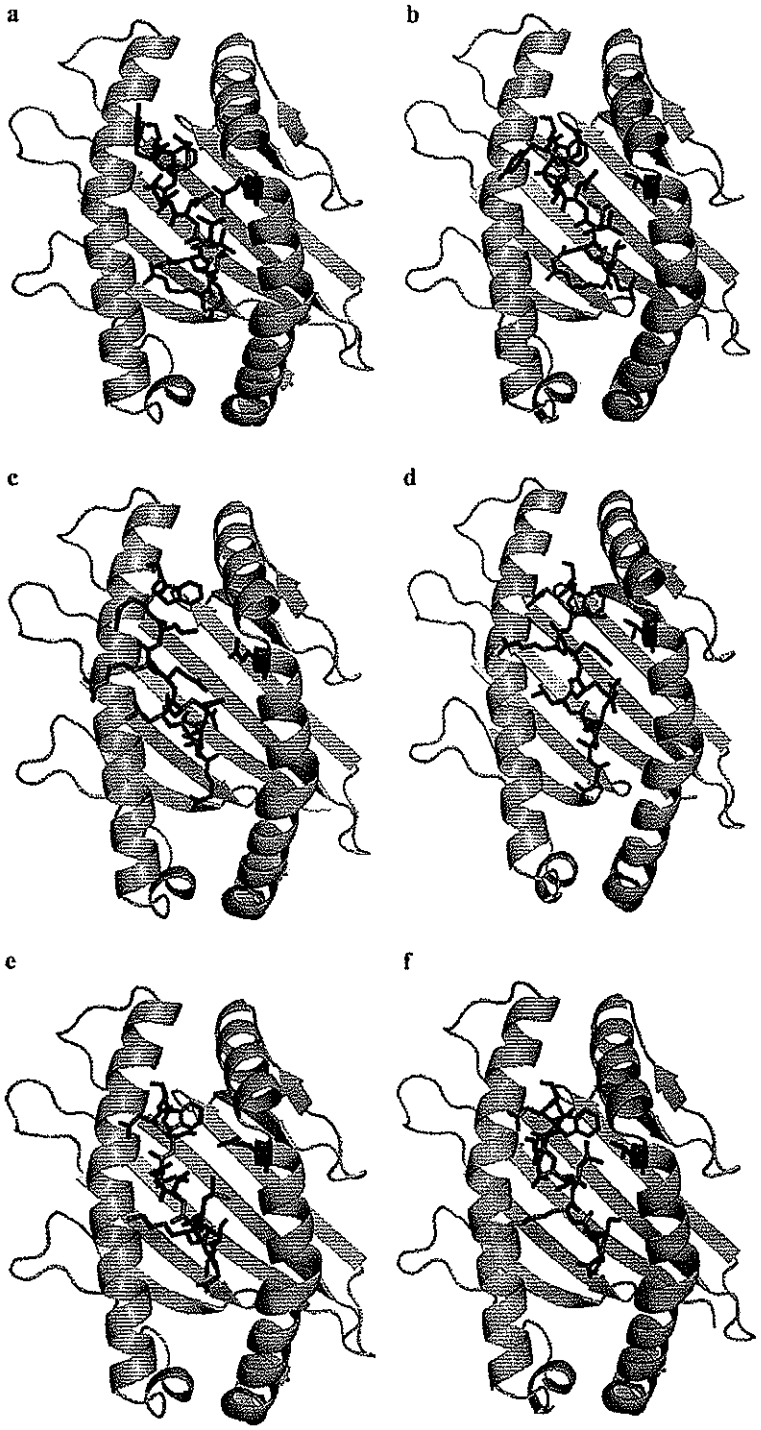


Figure 4.4. Molecular modeling of the Env-RW12, Rev-QW11, and Gag-GW12 peptides bound to MHC I molecules 7-6 and 141. Top views of Env-RW12 bound to 7-6 **(a)** and 141 **(b)**, Rev-QW11 bound to 7-6

(c) and 141 (d), and Gag-GW12 bound to 7-6 (e) and 141 (f). The binding clefts of 7-6 and 141 are shown in green, the peptides in red, and the ^{152}E (V) residue in blue. The first residue of the bound peptides is oriented down.

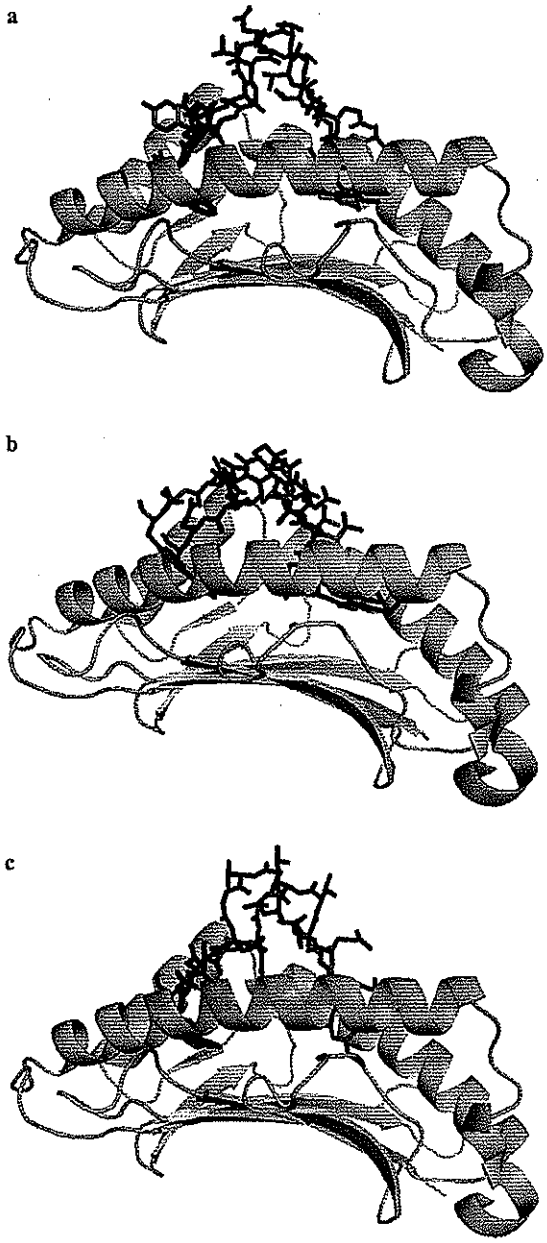


Figure 4.5. Molecular modeling suggested that the Env-RW12, Rev-QW11, and Gag-GW12 peptides bound to MHC I molecules 7-6 and 141 in bulged conformations, and that Env-RW12 (a) and Rev-QW11 (b)

maintained similar conformations when bound to 7-6 and 141, whereas the conformation of Gag-GW12 (c) changed when bound to 141. For each peptide, the conformation when bound to 7-6 is shown in red, and the conformation when bound to 141 is shown in blue. The first residue of the bound peptides is oriented to the right.

	7-6	141
Docking score[a]		
Env-RW12	6.35	4.66
Gag-GW12	2.96	2.25
Rev-QW11	2.84	4.89
Interaction category		
Hydrophobic		
Env-RW12	55	33
Gag-GW12	26	26
Rev-QW11	24	20
Salt bridges (attractive charged)		
Env-RW12	10	5
Gag-GW12	2	0
Rev-QW11	8	11
Repulsive charged		
Env-RW12	3	3
Gag-GW12	0	0
Rev-QW11	2	0
Hydrogen bonds	1	
Env-RW12	2	1
Gag-GW12	7	6
Rev-QW11	6	5
Aromatic stacking		
Env-RW12	6	5
Gag-GW12	2	2
Rev-QW11	2	1

Table 4.2. Docking scores and numbers of interactions (by category) among atoms of contact residues for the Env-RW12, Gag-GW12, and Rev-QW11 peptides complexed with the 7-6 and 141 MHC class I molecules

[a]The RPScores were calculated for each complex using the FTDock2.0 program (69).

[b] The number of interactions by category between atoms of contact residues for each complex was determined using the STING Millenium Suite program (71, 72).

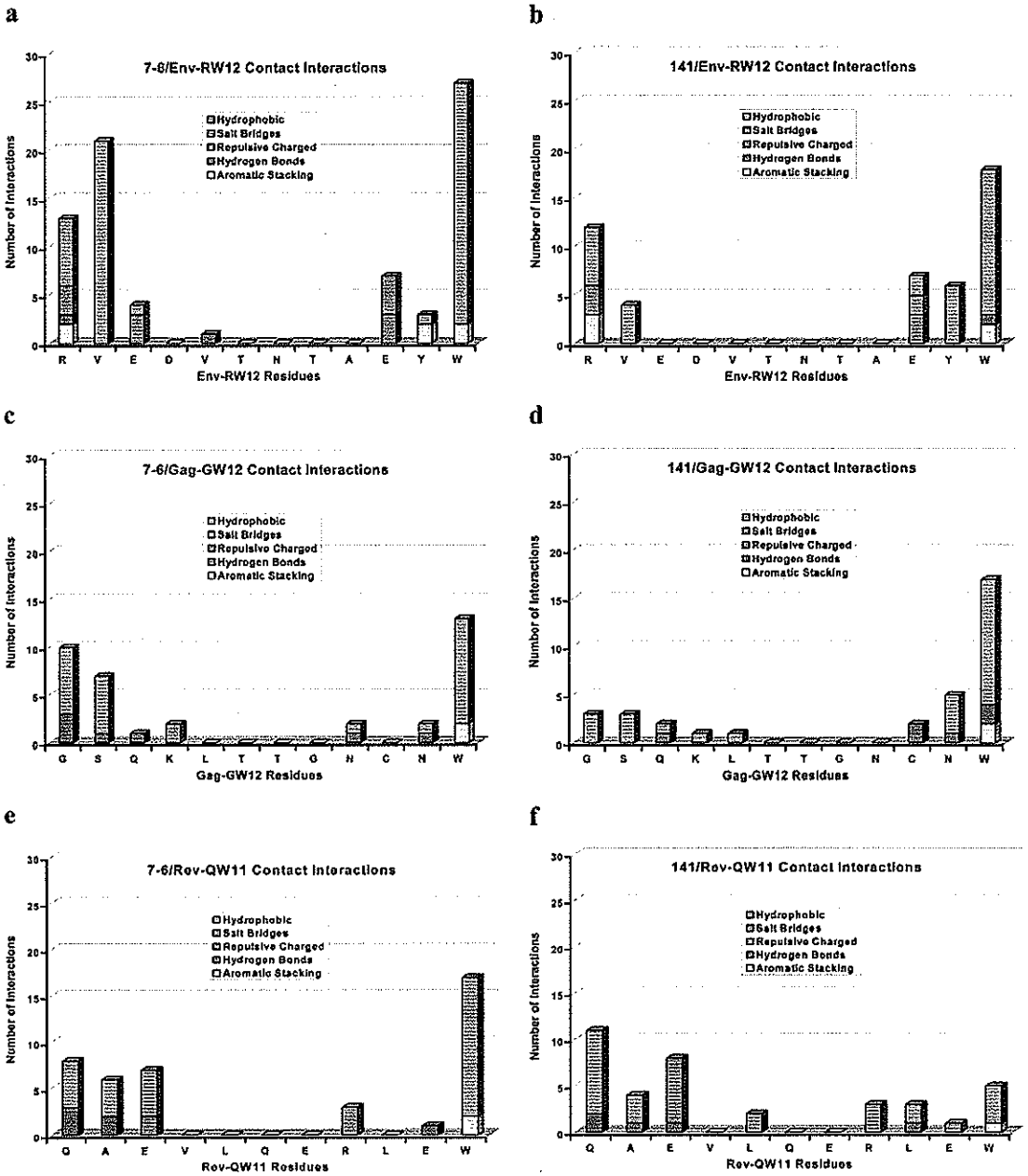


Figure 4.6. Numbers of interactions by category between each residue of Env-RW12 and its contact residues in the 7-6 (**a**) and 141 (**b**) complex, each residue of Gag-GW12 and its contact residues in the 7-6 (**c**) and 141 (**d**) complex, and each residue of Rev-QW11 and its contact residues in the 7-6 (**e**) and 141 complex (**f**).

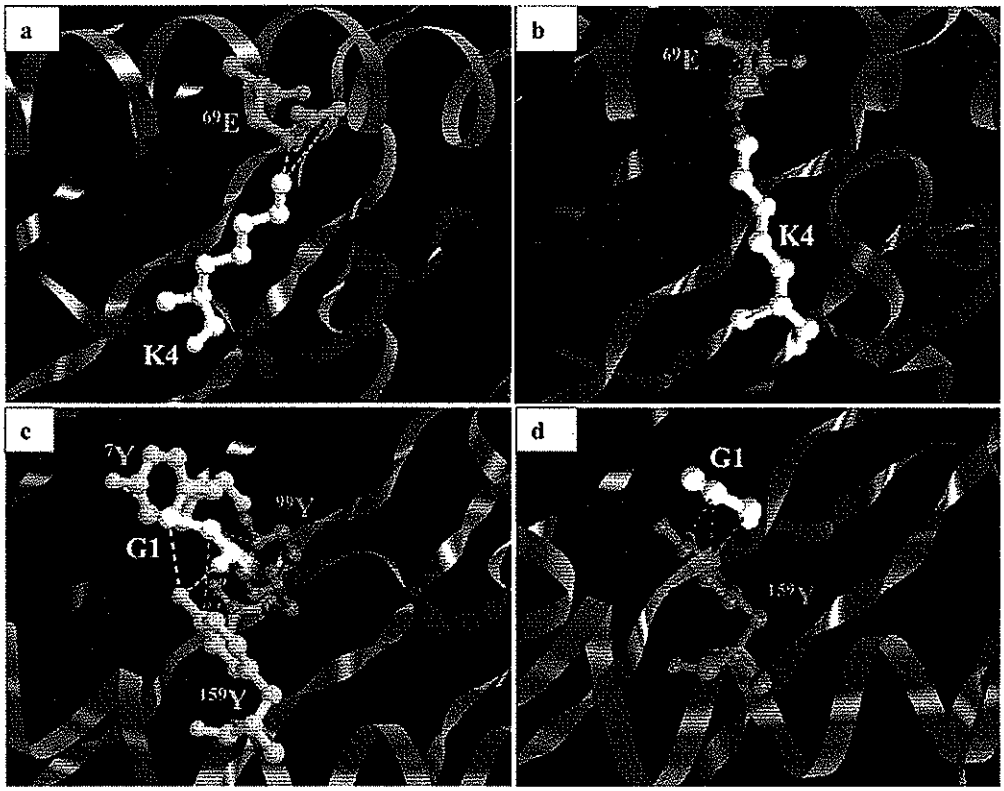


Figure 4.7. Interactions between contact residues of Gag-GW12 and the 7-6 and 141 MHC I molecules. **(a)** Two salt bridges (bright green dotted lines) between 69E (bright green) of 7-6 and the K4 residue (white) of Gag-GW12 (red ribbon). 69E protrudes from the α-1 helix (turquoise coiled ribbon) of 7-6. The other green and blue ribbons represent the floor of the peptide-binding cleft. **(b)** A single hydrophobic interaction (purple dotted line) between 69E (purple) of 141 and the K4 residue (white) of Gag-GW12 (red ribbon). Other designations are the same as in Fig. 4.1a. **(c)** Seven hydrophobic interactions (purple dotted lines) and three hydrogen bonds (tan dotted lines) among 7Y (tan), 99Y (purple), and 159Y (tan) of 7-6 and the G1 residue (white) of Gag-GW12 (red ribbon). 7Y and 99Y protrude from the floor of the peptide-binding cleft (blue and green ribbons), and 159Y protrudes from the α-2 helix (green coiled ribbon) of 7-6. **(d)** Three hydrophobic interactions (purple dotted lines) between 159Y (purple) of 141 and the G1 residue (white) of Gag-GW12 (red

ribbon). Other designations are the same as in Fig. 4.1c. Images a–d were generated with the STING Millennium Suite (77,78), based on the docking models.

DISCUSSION

This study provided the first cloning, sequencing, and expressing of functionally distinct equine MHC I alleles within an ELA-A haplotype as defined by peptide-specific CTL, and demonstrated for the first time that a single residue difference between two naturally occurring MHC I molecules affected the recognition of Gag and Rev epitopes by lentiviral-specific CTL. This was accomplished using standard ^{51}Cr release assays to assess the ability of EIAV Env-, Gag-, and Rev-specific CTL to recognize the corresponding peptides on human 721.221 and heterologous EK target cells transduced with retroviral vectors expressing two distinct MHC I genes derived from three horses with the ELA-A1 haplotype. When compared with the 7-6 molecule, the single $^{152}E \rightarrow V$ substitution in the α-2 domain of the 141 molecule enhanced CTL recognition of the Rev-QW11 epitope, but abolished CTL recognition of the Gag-GW12 epitope. Recognition of the Env-RW12 epitope by CTL was not affected.

The hypothesis that the ELA-A1 haplotype was comprised of functionally distinct alleles was based on the initial observations that target cells from ELA-A1 horses A2140 and A2152 were recognized by both Env-RW12- and Gag-GW12-specific CTL whereas target cells from ELA-A1 horse A2150 were not recognized by Gag-GW12-specific CTL, and that Rev-QW11-specific CTL recognized A2150 target cells more efficiently than A2140 and A2152 target cells. Although the latter observation was consistent with the conclusion that the 7-6 MHC I molecule was used by A2140 and A2152 to present these epitopes while the 141 molecule was used by A2150, the efficiency of Rev-QW11-specific CTL recognition of A2140 and A2152 target cells was not exactly the same. The reason for the slightly greater

recognition efficiency of the A2152 targets is not known. However, in addition to inheriting ELA-A1 haplotypes from different dams, A2140 and A2150 inherited different ELA haplotypes from the sire. It was therefore possible that the heterogeneous complement of other class I molecules on the respective target cells affected the expression of 7-6, or otherwise decreased peptide binding by 7-6 through competition or other unknown mechanisms (79, 80). Nonetheless, later experiments confirmed the hypothesis.

The loss of Gag-GW12 recognition by CTL was the most striking result of the single amino acid difference between the 7-6 and 141 molecules. Live-cell peptide-binding inhibition assays indicated that Gag-GW12 was able to bind 141, albeit with 1.7 times lower affinity than 7-6. Although this lower binding affinity could have contributed to the inability of Gag-GW12-specific CTL to lyse 141-expressing target cells, loss of TCR recognition of the 141/Gag-GW12 complex was probably more important. Others have shown that a single residue difference between two MHC class I molecules can dictate marked conformational differences in the same bound peptide, directly affecting TCR recognition by CTL (81). Although crystal structures are necessary for confirmation, molecular modeling suggested that the absence of salt bridges, along with the paucity of interactions between the G1 residue (a probable anchor residue) of Gag-GW12 and the 141 molecule, resulted in the observed lower-affinity binding, lower calculated docking score, and a more sharply protruding Gag-GW12 bound confirmation as compared with the 7-6/Gag-GW12 complex. These factors could have lowered TCR affinity for the 141/Gag-GW12 complex such that 141-restricted killing by Gag-GW12-specific CTL no longer occurred.

For the 7-6 molecule, Env-RW12 bound with only 1.3 times higher affinity than Gag-GW12, yet Env-RW12 CTL recognized 7-6 targets with 78 times greater efficiency than Gag-GW12 CTL. Therefore, the difference in affinity for the 7-6 MHC class I molecule between Env-RW12 and Gag-GW12 did not account for the difference in CTL-recognition efficiency, again suggesting that TCR affinity for the MHC/peptide complex played a major

role in the discrepancy. Earlier work yielded similar results and conclusions (27, 51). Although the observed difference in 7-6 binding affinity for Env-RW12 and Gag-GW12 was small, molecular modeling provided some possible explanations for the difference. Specifically, the 7-6/Gag-GW12 complex had fewer hydrophobic interactions, fewer salt bridges, and a lower docking score than the 7-6/Env-RW12 complex.

Despite the detrimental effects on Gag-GW12-specific CTL recognition, the $^{152}E \rightarrow V$ change in 141 enhanced recognition of Rev-QW11 by CTL. Although CTL recognized Rev-QW11 on *7-6*-transduced MHC I-mismatched EK cell targets, *7-6*-transduced human 721.221 cells presented Rev-QW11 poorly. This type of differential recognition (equine vs human targets) was not observed for the other peptides, suggesting that human β_2m did not effectively stabilize the 7-6/Rev-QW11 complex. This explanation is supported by the observation that human and murine β_2m have different murine MHC I-stabilizing effects (82). It is not known why the presentation of the other peptides was not negatively affected by human β_2m. Binding inhibition assays indicated that Rev-QW11 bound to 7-6 with 2.1 times lower affinity than to 141, and this inefficient binding could have made the stabilizing effects of autologous β_2m more important for the 7-6/Rev-QW11 complex. Based on the molecular modeling results, the 7-6/Rev-QW11 complex had the second lowest docking score overall, and had fewer salt bridges and more destabilizing repulsive-charged interactions than did the 141/Rev-QW11 complex. In addition, and likely because of the reasons just listed, Rev-QW11 bulged out from the 7-6 binding cleft more than it did from 141. This conformational change in bound Rev-QW11 could have negatively affected TCR recognition of the 7-6/Rev-QW11 complex by Rev-QW11-specific CTL. Although crystal structures are needed, modeling provided plausible mechanisms for the inefficient binding of Rev-QW11 to 7-6, and for the inefficient Rev-QW11-specific CTL recognition of 7-6-expressing target cells.

Given the efficient Rev-QW11-specific CTL recognition of 141-expressing target cells and the absence of Gag-GW12-specific CTL recognition of 141-expressing target cells,

the observation that Rev-QW11 bound 141 with half the affinity of Gag-GW12 was quite unexpected. Despite the lower MHC binding affinity, the 141-bound conformation of Rev-QW11 must have been efficiently recognized by the TCR of Rev-QW11-specific CTL. In contrast, the TCR of Gag-GW12-specific CTL probably could not bind the sharply protruding and highly bulged conformation of Gag-GW12 in the 141/Gag-GW12 complex.

The CTL epitopes (Env-RW12, Gag-GW12, and Rev-QW11) evaluated in this study were similar in that all three have a large aromatic tryptophan at the C terminus. Additionally, both the N-terminal and C-terminal residues are required for MHC binding and/or CTL recognition for all three epitopes (27, 51). For Env-RW12, the V2 residue is also a probable anchor residue based on 7-6 binding inhibition assays using peptides with amino acid substitutions (51). These observations are consistent with the molecular modeling results obtained in the present study. Analysis of the hydrophilic and hydrophobic amino acid residues for each of the three peptides indicate that Env-RW12 and Rev-QW11 differ from Gag-GW12 at positions 1, 2, 8, and 9. Both Env-RW12 and Rev-QW11 have hydrophilic residues at positions 1 and 8, whereas Gag-GW12 has hydrophobic residues at these positions. At positions 2 and 9, both Env-RW12 and Rev-QW11 have hydrophobic residues, whereas Gag-GW12 has hydrophilic residues at these positions. The differences in these residues, which included probable N-terminal anchor residues, likely contributed to the differences in MHC/peptide complex conformations that allowed CTL TCR recognition of all three peptides when presented by 7-6, but recognition of only Env-RW12 and Rev-QW11 when presented by 141.

Interestingly, all three peptides in this study were longer than the 8–10 aa generally considered optimal for MHC I binding. However, MHC I molecules can bind peptides as long as 14 aa in a bulged conformation and elicit dominant CTL responses (83). Importantly, the BZLF1 protein of EBV includes three completely overlapping CTL epitopes of 9, 11, and 13 aa in length (84). Although all three peptides bind well to the HLA-B*3501 molecule, the

CTL response in individuals with this allele is directed exclusively toward the 11-mer epitope (84). Of particular interest, individuals with the B*3503 allele, which differs from B*3501 by a single amino acid in the F pocket of the peptide-binding cleft, do not mount CTL responses to these peptides because they do not bind B*3503. However, individuals with B*3508, which differs from B*3501 by a single amino acid in the D pocket of the peptide-binding cleft, develop CTL responses to the 13-mer epitope (84). The crystal structures indicate that the 13-mer binds both B*3508 and B*3501 in a centrally bulged conformation with the N and C termini anchored in the A and F pockets of the peptide-binding cleft (73). The differential CTL response is due to a broader peptide-binding cleft in B*3508, since the narrower binding cleft of the B*3501-peptide complex interacts poorly with the dominant TCR (73). Crystal structures will be required to confirm that the peptides in the present study bind the 7-6 and 141 molecules in a bulged conformation, and whether or not similar mechanisms are involved in the differential recognition by Env-RW12-, Gag-GW12-, and Rev-QW11-specific CTL.

Although crystal structures of equine MHC I molecules are lacking, the sequences of the two equine class I alleles presented here are surprisingly similar to human and murine class I alleles, and many of the residues forming the binding pockets A–F (in human and murine class I molecules) are shared (85, 86, 87). This suggests that the structure of equine MHC I molecules is similar to that of the mouse and human. In the absence of crystallography, molecular modeling provided important insights into the differential recognition of 7-6- and 141-expressing targets by EIAV Gag-GW12- and Rev-QW11-specific CTL. For example, the charged ^{152}E to hydrophobic ^{152}V substitution in the 141 molecule likely affected stabilizing salt bridges for Gag-GW12 binding, as seen when the 13-mer BZLF1 EBV peptide binds to B*3508 and B*3501, which differ only at residue position 156 (charged ^{156}R→hydrophobic ^{156}L) (73). If the modeling is correct, the absence of these

stabilizing salt bridges contributed to the conformational change leading to loss of TCR recognition of the 141/Gag-GW12 complex.

This study confirms that a single amino acid difference between naturally occurring MHC I molecules can result in the loss of Gag-specific CTL recognition and enhanced (or diminished) efficiency of Rev-specific CTL recognition. The CTL, peptide-binding, and molecular modeling observations in this study support and suggest molecular mechanisms for the observed differences in disease progression and CTL responses in HIV-1-infected individuals of the B*35-Px and B*35-PY MHC I genotypes, because these genotypes also only differ by as few as one amino acid in the peptide-binding domain (40, 41). The implications of the observed differences in CTL recognition of important EIAV epitopes due to a single amino acid difference between otherwise identical MHC I molecules are important for designing protective lentivirus-specific CTL-inducing vaccines and understanding differential lentivirus disease progression in individuals within a population.

ACKNOWLEDGEMENTS

The important technical assistance of Emma Karel and Lori Fuller is acknowledged. We also thank Dr. Susan Carpenter for helpful discussions and advice.

DISCLOSURES

The authors have no financial conflict of interest.

REFERENCES

1. **Addo, M. M., M. Altfeld, E. S. Rosenberg, R. L. Eldridge, M. N. Philips, K. Habeeb, A. Khatri, C. Brander, G. K. Robbins, G. P. Mazzara, et al.** 2001 The HIV-1 regulatory proteins tat and rev are frequently targeted by cytotoxic T lymphocytes derived from HIV-1-infected individuals. Proc. Natl. Acad. Sci. USA **98:** 1781-1786.

2. **Betts, M. R., J. F. Krowka, T. B. Kepler, M. Davidian, C. Christopherson, S. Kwok, L. Louie, J. Eron, H. Sheppard, J. A. Frelinger.** 1999. Human immunodeficiency virus type 1-specific cytotoxic T lymphocyte activity is inversely correlated with HIV type 1 viral load in HIV type 1-infected long-term survivors. AIDS Res. Hum. Retroviruses **15:** 1219-1228.

3. **Borrow, P., H. Lewicki, B. H. Hahn, G. M. Shaw, M. B. Oldstone.** 1994. Virus-specific $CD8^f$ cytotoxic T-lymphocyte activity associated with control of viremia in primary human immunodeficiency virus type 1 infection. J. Virol. **68:** 6103-6110.

4. **Gea-Banacloche, J. C., S. A. Migueles, L. Martino, W. L. Shupert, A. C. McNeil, M. S. Sabbaghian, L. Ehler, C. Prussin, R. Stevens, L. Lambert, et al.** 2000. Maintenance of large numbers of virus-specific $CD8^+$ T cells in HIV-infected progressors and long-term nonprogressors. J. Immunol. **165:** 1082-1092.

5. **Klein, M. R., C. A. Van Baalen, A. M. Holwerda, S. R. Kerkhof Garde, R. J. Bende, I. P. Keet, J. K. Eeftinck Schattenkerk, A. D. Osterhaus, H. Schuitemaker, F. Miedema.** 1995. Kinetics of Gag-specific cytotoxic T lymphocyte responses during the clinical course of HIV-1 infection: a longitudinal analysis of rapid progressors and long-term asymptomatics. J. Exp. Med. **181:** 1365-1372.

6. **Goulder, P. J., M. A. Altfeld, E. S. Rosenberg, T. Nguyen, Y. Tang, R. L. Eldridge, M. M. Addo, S. He, J. S. Mukherjee, M. N. Phillips, et al.** 2001. Substantial

differences in specificity of HIV-specific cytotoxic T cells in acute and chronic HIV infection. J. Exp. Med. **193:** 181-194.

7. **Novitsky, V., P. Gilbert, T. Peter, M. F. McLane, S. Gaolekwe, N. Rybak, I. Thior, T. Ndung'u, R. Marlink, T. H. Lee, M. Essex.** 2003. Association between virus-specific T-cell responses and plasma viral load in human immunodeficiency virus type 1 subtype C infection. J. Virol. 77: 882-890.

8. **Ogg, G. S., X. Jin, S. Bonhoeffer, P. R. Dunbar, M. A. Nowak, S. Monard, J. P. Segal, Y. Cao, S. L. Rowland-Jones, V. Cerundolo, et al.** 1998. Quantitation of HIV-1-specific cytotoxic T lymphocytes and plasma load of viral RNA. Science **279:** 2103-2106.

9. **Ogg, G. S., S. Kostense, M. R. Klein, S. Jurriaans, D. Hamann, A. J. McMichael, F. Miedema.** 1999. Longitudinal phenotypic analysis of human immunodeficiency virus type 1-specific cytotoxic T lymphocytes: correlation with disease progression. J. Virol. **73:** 9153-9160.

10. **Rinaldo, C., X. L. Huang, Z. F. Fan, M. Ding, L. Beltz, A. Logar, D. Panicali, G. Mazzara, J. Liebmann, M. Cottrill, et al.** 1995. High levels of anti-human immunodeficiency virus type 1 (HIV-1) memory cytotoxic T-lymphocyte activity and low viral load are associated with lack of disease in HIV-1-infected long-term nonprogressors. J. Virol. **69:** 5838-5842.

11. **Gruters, R. A., C. A. van Baalen, A. D. Osterhaus.** 2002. The advantage of early recognition of HIV-infected cells by cytotoxic T-lymphocytes. Vaccine **20:** 2011-2015.

12. **Van Baalen, C. A., O. Pontesilli, R. C. Huisman, A. M. Geretti, M. R. Klein, F. de Wolf, F. Miedema, R. A. Gruters, A. D. Osterhaus.** 1997. Human immunodeficiency virus type 1 Rev- and Tat-specific cytotoxic T lymphocyte frequencies inversely correlate with rapid progression to AIDS. J. Gen. Virol. **78:** 1913-1918.

13. **Cheevers, W. P., T. C. McGuire.** 1985. Equine infectious anemia virus: immunopathogenesis and persistence. Rev. Infect. Dis. **7:** 83-88.

14. **McGuire, T. C., K. I. O'Rourke, L. E. Perryman.** 1990. Immunopathogenesis of equine infectious anemia lentivirus disease. Dev. Biol. Stand. **72:** 31-37.

15. **Sellon, D. C., F. J. Fuller, T. C. McGuire.** 1994. The immunopathogenesis of equine infectious anemia virus. Virus Res. **32:** 111-138.

16. **Kono, Y., K. Hirasawa, Y. Fukunaga, T. Taniguchi.** 1976. Recrudescence of equine infectious anemia by treatment with immunosuppressive drugs. Natl. Inst. Anim. Health Q. Tokyo **16:** 8-15.

17. **McGuire, T. C., D. B. Tumas, K. M. Byrne, M. T. Hines, S. R. Leib, A. L. Brassfield, K. I. O'Rourke, L. E. Perryman.** 1994. Major histocompatibility complex-restricted $CD8^+$ cytotoxic T lymphocytes from horses with equine infectious anemia virus recognize env and gag/PR proteins. J. Virol. **68:** 1459-1467.

18. **McGuire, T. C., D. G. Fraser, R. H. Mealey.** 2002. Cytotoxic T lymphocytes and neutralizing antibody in the control of equine infectious anemia virus. Viral Immunol. **15:** 521-531.

19. **Mealey, R. H., D. G. Fraser, J. L. Oaks, G. H. Cantor, T. C. McGuire.** 2001. Immune reconstitution prevents continuous equine infectious anemia virus replication in an Arabian foal with severe combined immunodeficiency: lessons for control of lentiviruses. Clin. Immunol. **101:** 237-247.

20. **Mealey, R. H., S. R. Leib, S. L. Pownder, T. C. McGuire.** 2004. Adaptive immunity is the primary force driving selection of equine infectious anemia virus envelope SU variants during acute infection. J. Virol. **78:** 9295-9305.

21. **Perryman, L. E., K. I. O'Rourke, T. C. McGuire.** 1988. Immune responses are required to terminate viremia in equine infectious anemia lentivirus infection. J. Virol. **62:** 3073-3076.

22. **Tumas, D. B., M. T. Hines, L. E. Perryman, W. C. Davis, T. C. McGuire.** 1994. Corticosteroid immunosuppression and monoclonal antibody-mediated $CD5^+$ T lymphocyte depletion in normal and equine infectious anaemia virus-carrier horses. J. Gen. Virol. **75:** 959-968.

23. **Hammond, S. A., S. J. Cook, D. L. Lichtenstein, C. J. Issel, R. C. Montelaro.** 1997. Maturation of the cellular and humoral immune responses to persistent infection in horses by equine infectious anemia virus is a complex and lengthy process. J. Virol. **71:** 3840-3852.

24. **McGuire, T. C., W. Zhang, M. T. Hines, P. J. Henney, K. M. Byrne.** 1997. Frequency of memory cytotoxic T lymphocytes to equine infectious anemia virus proteins in blood from carrier horses. Virology **238:** 85-93.

25. **McGuire, T. C., S. R. Leib, S. M. Lonning, W. Zhang, K. M. Byrne, R. H. Mealey.** 2000. Equine infectious anaemia virus proteins with epitopes most frequently recognized by cytotoxic T lymphocytes from infected horses. J. Gen. Virol. **81:** 2735-2739.

26. **Zhang, W., S. M. Lonning, T. C. McGuire.** 1998. Gag protein epitopes recognized by ELA-A-restricted cytotoxic T lymphocytes from horses with long-term equine infectious anemia virus infection. J. Virol. **72:** 9612-9620.

27. **Mealey, R. H., B. Zhang, S. R. Leib, M. H. Littke, T. C. McGuire.** 2003. Epitope specificity is critical for high and moderate avidity cytotoxic T lymphocytes associated with control of viral load and clinical disease in horses with equine infectious anemia virus. Virology **313:** 537-552.

28. **Mealey, R. H., A. Sharif, S. A. Ellis, M. H. Littke, S. R. Leib, T. C. McGuire.** 2005. Early detection of dominant env-specific and subdominant gag-specific $CD8^+$ lymphocytes in equine infectious anemia virus-infected horses using major histocompatibility complex class I/peptide tetrameric complexes. Virology **339:** 110-126.

29. **Chung, C., R. H. Mealey, T. C. McGuire.** 2004. CTL from EIAV carrier horses with diverse MHC class I alleles recognize epitope clusters in gag matrix and capsid proteins. Virology **327:** 144-154.

30. **Chung, C., R. H. Mealey, T. C. McGuire.** 2005. Evaluation of high functional avidity CTL to gag epitope clusters in EIAV carrier horses. Virology **342:** 228-239.

31. **Belshan, M., P. Baccam, J. L. Oaks, B. A. Sponseller, S. C. Murphy, J. Cornette, S. Carpenter.** 2001. Genetic and biological variation in equine infectious anemia virus rev correlates with variable stages of clinical disease in an experimentally infected pony. Virology **279:** 185-200.

32. **Leroux, C., C. J. Issel, R. C. Montelaro.** 1997. Novel and dynamic evolution of equine infectious anemia virus genomic quasispecies associated with sequential disease cycles in an experimentally infected pony. J. Virol. **71:** 9627-9639.

33. **Germain, R. N., D. H. Margulies.** 1993. The biochemistry and cell biology of antigen processing and presentation. Annu. Rev. Immunol. **11:** 403-450.

34. **Robinson, J., M. J. Waller, P. Parham, N. de Groot, R. Bontrop, L. J. Kennedy, P. Stoehr, S. G. Marsh.** 2003. IMGT/HLA and IMGT/MHC: sequence databases for the study of the major histocompatibility complex. Nucleic Acids Res. **31:** 311-314.

35. **Sette, A., J. Sidney.** 1998. HLA supertypes and supermotifs: a functional perspective on HLA polymorphism. Curr. Opin. Immunol. **10:** 478-482.

36. **Sette, A., J. Sidney.** 1999. Nine major HLA class I supertypes account for the vast preponderance of HLA-A and -B polymorphism. Immunogenetics **50:** 201-212.

37. **Sidney, J., H. M. Grey, R. T. Kubo, A. Sette.** 1996. Practical, biochemical and evolutionary implications of the discovery of HLA class I supermotifs. Immunol. Today **17:** 261-266.

38. **Sidney, J., S. Southwood, A. Sette.** 2005. Classification of A1- and A24-supertype molecules by analysis of their MHC-peptide binding repertoires. Immunogenetics **57:** 393-408.

39. **Muller, D., K. Pederson, R. Murray, J. A. Frelinger.** 1991. A single amino acid substitution in an MHC class I molecule allows heteroclitic recognition by lymphocytic choriomeningitis virus-specific cytotoxic T lymphocytes. J. Immunol. **147:** 1392-1397.

40. **Gao, X., G. W. Nelson, P. Karacki, M. P. Martin, J. Phair, R. Kaslow, J. J. Goedert, S. Buchbinder, K. Hoots, D. Vlahov, et al.** 2001. Effect of a single amino acid change in MHC class I molecules on the rate of progression to AIDS. N. Engl. J. Med. **344:** 1668-1675.

41. **Jin, X., X. Gao, M. Ramanathan, Jr, G. R. Deschenes, G. W. Nelson, S. J. O'Brien, J. J. Goedert, D. D. Ho, T. R. O'Brien, M. Carrington.** 2002. Human immunodeficiency virus type 1 (HIV-1)-specific $CD8^+$-T-cell responses for groups of HIV-1-infected individuals with different HLA-B*35 genotypes. J. Virol. **76:** 12603-12610.

42. **Bailey, E.**1980. Identification and genetics of horse lymphocyte alloantigens. Immunogenetics **11:** 499-506.

43. **Bailey, E., E. Marti, D. G. Fraser, D. F. Antczak, S. Lazary.** 2000. Immunogenetics of the horse. A. T. Bowling, Jr, and A. Ruvinsky, Jr, eds. The Genetics of the Horse CABI Publishing, New York.

44. **Bernoco, D., D. F. Antczak, E. Bailey, K. Bell, R. W. Bull, G. Byrns, G. Guerin, S. Lazary, J. McClure, J. Templeton, et al.** 1987. Joint report of the Fourth International Workshop on lymphocyte alloantigens of the horse, Lexington, Kentucky, 12–22 October, 1985. Anim. Genet. **18:** 81

45. **Lazary, S., D. F. Antczak, E. Bailey, T. K. Bell, D. Bernoco, G. Byrns, J. J. McClure.** 1988. Joint report of the Fifth International Workshop on lymphocyte alloantigens

of the horse, Baton Rouge, Louisiana, 31 October–1 November 1987. Anim. Genet. **19:** 447-456.

46. **Barbis, D. P., J. K. Maher, J. Stanek, B. A. Klaunberg, D. F. Antczak.** 1994. Horse cDNA clones encoding two MHC class I genes. Immunogenetics **40:** 163

47. **Carpenter, S., J. M. Baker, S. J. Bacon, T. Hopman, J. Maher, S. A. Ellis, D. F. Antczak.** 2001. Molecular and functional characterization of genes encoding horse MHC class I antigens. Immunogenetics **53:** 802-809.

48. **Chung, C., S. R. Leib, D. G. Fraser, S. A. Ellis, T. C. McGuire.** 2003. Novel classical MHC class I alleles identified in horses by sequencing clones of reverse transcription-PCR products. Eur. J. Immunogenet. **30:** 387-396.

49. **Ellis, S. A., A. J. Martin, E. C. Holmes, W. I. Morrison.** 1995. At least four MHC class I genes are transcribed in the horse: phylogenetic analysis suggests an unusual evolutionary history for the MHC in this species. Eur. J. Immunogenet. **22:** 249-260.

50. **Holmes, E. C., S. A. Ellis.** 1999. Evolutionary history of MHC class I genes in the mammalian order Perissodactyla. J. Mol. Evol. **49:** 316-324.

51. **McGuire, T. C., S. R. Leib, R. H. Mealey, D. G. Fraser, D. J. Prieur.** 2003. Presentation and binding affinity of equine infectious anemia virus CTL envelope and matrix protein epitopes by an expressed equine classical MHC class I molecule. J. Immunol. **171:** 1984-1993.

52. **Tallmadge, R. L., T. L. Lear, D. F. Antczak.** 2005. Genomic characterization of MHC class I genes of the horse. Immunogenetics **57:** 763-774.

53. **Bailey, E.** 1983. Population studies on the ELA system in American standardbred and thoroughbred mares. Anim. Blood Groups Biochem. Genet. **14:** 201-211.

54. **Terasaki, P. I., D. Bernoco, M. S. Park, G. Ozturk, Y. Iwaki.** 1978. Microdroplet testing for HLA-A, -B, -C, and -D antigens: The Phillip Levine Award Lecture. Am. J. Clin. Pathol. **69:** 103-120.

55. **Ellis, S. A., R. E. Bontrop, D. F. Antczak, K. Ballingall, C. J. Davies, J. Kaufman, L. J. Kennedy, J. Robinson, D. M. Smith, M. J. Stear, et al.** 2006. ISAG/IUIS-VIC Comparative MHC Nomenclature Committee report, 2005. Immunogenetics : 1-6.

56. **Miller, A. D., G. J. Rosman.** 1989. Improved retroviral vectors for gene transfer and expression. BioTechniques **7:** 980-982.

57. **Lonning, S. M., W. Zhang, S. R. Leib, T. C. McGuire.** 1999. Detection and induction of equine infectious anemia virus-specific cytotoxic T-lymphocyte responses by use of recombinant retroviral vectors. J. Virol. **73:** 2762-2769.

58. **Shimizu, Y., D. E. Geraghty, B. H. Koller, H. T. Orr, R. DeMars.** 1988. Transfer and expression of three cloned human non-HLA-A,B,C class I major histocompatibility complex genes in mutant lymphoblastoid cells. Proc. Natl. Acad. Sci. USA **85:** 227-231.

59. **Ridgely, S. L., T. C. McGuire.** 2002. Lipopeptide stimulation of MHC class I-restricted memory cytotoxic T lymphocytes from equine infectious anemia virus-infected horses. Vaccine **20:** 1809-1819.

60. **Ridgely, S. L., B. Zhang, T. C. McGuire.** 2003. Response of ELA-A1 horses immunized with lipopeptide containing an equine infectious anemia virus ELA-A1-restricted CTL epitope to virus challenge. Vaccine **21:** 491-506.

61. **Rivera, J. A., T. C. McGuire.** 2005. Equine infectious anemia virus-infected dendritic cells retain antigen presentation capability. Virology **335:** 145-154.

62. **Zhang, W., D. B. Auyong, J. L. Oaks, T. C. McGuire.** 1999. Natural variation of equine infectious anemia virus Gag protein cytotoxic T lymphocyte epitopes. Virology **261:** 242-252.

63. **Allen, T. M., D. H. O'Connor, P. Jing, J. L. Dzuris, B. R. Mothe, T. U. Vogel, E. Dunphy, M. E. Liebl, C. Emerson, N. Wilson, et al.** 2000. Tat-specific cytotoxic

T lymphocytes select for SIV escape variants during resolution of primary viraemia. Nature **407:** 386-390.

64. **Siliciano, R. F., A. D. Keegan, R. Z. Dintzis, H. M. Dintzis, H. S. Shin.** 1985. The interaction of nominal antigen with T cell antigen receptors. I. Specific binding of multivalent nominal antigen to cytolytic T cell clones. J. Immunol. **135:** 906-914.

65. **Alexander-Miller, M. A., G. R. Leggatt, A. Sarin, J. A. Berzofsky.** 1996. Role of antigen, CD8, and cytotoxic T lymphocyte (CTL) avidity in high dose antigen induction of apoptosis of effector CTL. J. Exp. Med. **184:** 485-492.

66. **Derby, M., M. Alexander-Miller, R. Tse, J. Berzofsky.** 2001. High-avidity CTL exploit two complementary mechanisms to provide better protection against viral infection than low-avidity CTL. J. Immunol. **166:** 1690-1697.

67. **del Guercio, M. F., J. Sidney, G. Hermanson, C. Perez, H. M. Grey, R. T. Kubo, A. Sette.** 1995. Binding of a peptide antigen to multiple HLA alleles allows definition of an A2-like supertype. J. Immunol. **154:** 685-693.

68. **Greenwood, F. C., W. M. Hunter, J. S. Glover.** 1963. The preparation of I-131-labelled human growth hormone of high specific radioactivity. Biochem. J. **89:** 114-123.

69. **Marti-Renom, M. A., A. C. Stuart, A. Fiser, R. Sanchez, F. Melo, A. Sali.** 2000. Comparative protein structure modeling of genes and genomes. Annu. Rev. Biophys. Biomol. Struct. **29:** 291-325.

70. **Miles, J. J., D. Elhassen, N. A. Borg, S. L. Silins, F. E. Tynan, J. M. Burrows, A. W. Purcell, L. Kjer-Nielsen, J. Rossjohn, S. R. Burrows, J. McCluskey.** 2005. CTL recognition of a bulged viral peptide involves biased TCR selection. J. Immunol. **175:** 3826-3834.

71. **Luthy, R., J. U. Bowie, D. Eisenberg.** 1992. Assessment of protein models with three-dimensional profiles. Nature **356:** 83-85.

72. **Canutescu, A. A., A. A. Shelenkov, R. L. Dunbrack, Jr.** 2003. A graph-theory algorithm for rapid protein side-chain prediction. Protein Sci. **12:** 2001-2014.

73. **Tynan, F. E., N. A. Borg, J. J. Miles, T. Beddoe, D. El Hassen, S. L. Silins, W. J. van Zuylen, A. W. Purcell, L. Kjer-Nielsen, J. McCluskey, et al.** 2005. High resolution structures of highly bulged viral epitopes bound to major histocompatibility complex class I: implications for T-cell receptor engagement and T-cell immunodominance. J. Biol. Chem. **280:** 23900-23909.

74. **Gabb, H. A., R. M. Jackson, M. J. Sternberg.** 1997. Modelling protein docking using shape complementarity, electrostatics and biochemical information. J. Mol. Biol. **272:** 106-120.

75. **Katchalski-Katzir, E., I. Shariv, M. Eisenstein, A. A. Friesem, C. Aflalo, I. A. Vakser.** 1992. Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. Proc. Natl. Acad. Sci. USA **89:** 2195-2199.

76. **Moont, G., H. A. Gabb, M. J. Sternberg.** 1999. Use of pair potentials across protein interfaces in screening predicted docked complexes. Proteins **35:** 364-373.

77. **Mancini, A. L., R. H. Higa, A. Oliveira, F. Dominiquini, P. R. Kuser, M. E. Yamagishi, R. C. Togawa, G. Neshich.** 2004. STING Contacts: a web-based application for identification and analysis of amino acid contacts within protein structure and across protein interfaces. Bioinformatics **20:** 2145-2147.

78. **Neshich, G., R. C. Togawa, A. L. Mancini, P. R. Kuser, M. E. Yamagishi, G. Pappas, Jr, W. V. Torres, T. Fonseca e Campos, L. L. Ferreira, F. M. Luna, et al.** 2003. STING Millennium: a web-based suite of programs for comprehensive and simultaneous analysis of protein structure and sequence. Nucleic Acids Res. **31:** 3386-3392.

79. **Day, C. L., A. K. Shea, M. A. Altfeld, D. P. Olson, S. P. Buchbinder, F. M. Hecht, E. S. Rosenberg, B. D. Walker, S. A. Kalams.** 2001. Relative dominance of

epitope-specific cytotoxic T-lymphocyte responses in human immunodeficiency virus type 1-infected persons with shared HLA alleles. J. Virol. **75:** 6279-6291.

80. **Moudgil, K. D., J. Wang, V. P. Yeung, E. E. Sercarz.** 1998. Heterogeneity of the T cell response to immunodominant determinants within hen eggwhite lysozyme of individual syngeneic hybrid F1 mice: implications for autoimmunity and infection. J. Immunol. **161:** 6046-6053.

81. **Tynan, F. E., D. Elhassen, A. W. Purcell, J. M. Burrows, N. A. Borg, J. J. Miles, N. A. Williamson, K. J. Green, J. Tellam, L. Kjer-Nielsen, J. McCluskey, J. Rossjohn, S. R. Burrows.** 2005. The immunogenicity of a viral cytotoxic T cell epitope is controlled by its MHC-bound conformation. J. Exp. Med. **202:** 1249-1260.

82. **Shields, M. J., N. Assefi, W. Hodgson, E. J. Kim, R. K. Ribaudo.** 1998. Characterization of the interactions between MHC class I subunits: a systematic approach for the engineering of higher affinity variants of β2-microglobulin. J. Immunol. **160:** 2297-2307.

83. **Burrows, S. R., J. Rossjohn, J. McCluskey.** 2005. Have we cut ourselves too short in mapping CTL epitopes?. Trends Immunol. **27:** 11-16.

84. **Green, K. J., J. J. Miles, J. Tellam, W. J. van Zuylen, G. Connolly, S. R. Burrows.** 2004. Potent T cell response to a class I-binding 13-mer viral epitope and the influence of HLA micropolymorphism in controlling epitope length. Eur. J. Immunol. **34:** 2510-2519.

85. **Bjorkman, P. J., M. A. Saper, B. Samraoui, W. S. Bennett, J. L. Strominger, D. C. Wiley.** 1987. Structure of the human class I histocompatibility antigen, HLA-A2. Nature **329:** 506-512.

86. **Fremont, D. H., M. Matsumura, E. A. Stura, P. A. Peterson, I. A. Wilson.** 1992. Crystal structures of two viral peptides in complex with murine MHC class I H-2Kb. Science **257:** 919-927.

87. **Matsumura, M., D. H. Fremont, P. A. Peterson, I. A. Wilson.** 1992. Emerging principles for the recognition of peptide antigens by MHC class I molecules. Science **257:** 927-934.

CHAPTER 5. GENERAL CONCLUSIONS

Characterization of macromolecular interactions is a basic step for understanding complex cellular events. Most essential cellular functions involve macromolecular interactions, such as protein-protein and protein-RNA interactions. Although most viruses do not encode all of their own components, they also require various macromolecular interactions for replication, with virally-encoded proteins often interacting with host cellular components. Lentiviruses, especially, as single stranded RNA viruses, evolve faster than their eukaryotes hosts because of a high mutation rate and frequent recombination events within the viral genome. It is important to understand how macromolecular interactions in the lentiviruses have been maintained in the presence of such strong evolutionary selective pressure. We have focused our studies on Equine Infectious Anemia Virus (EIAV), one of the most divergent known lentiviruses. It has a similar viral life cycle and overall genome structure to other lentiviruses, but it has a different disease course, with alternating periods of acute, chronic, and inapparent disease. EIAV is a valuable model system for studying virus-host macromolecular interactions because it has a relatively small genome and a limited number of gene products, compared to other more complex lentiviruses like HIV-1. By studying the basic mechanisms involved in the EIAV replication, we can understand the essential conserved features of lentiviruses.

IDENTIFICATION OF A BIPARTITE RNA BINDING DOMAIN IN EIAV REV

The Rev protein of equine infectious anemia virus (EIAV) and other lentiviruses is an essential regulatory protein that facilitates export of incompletely spliced viral RNAs from

the nucleus to the cytoplasm (18). Discrete functional domains of Rev mediate protein-RNA and protein-protein interactions that are required for nuclear import, RNA binding, multimerization, and nuclear export. Rev binds a specific sequence in the lentiviral RNA, termed the Rev-responsive element (RRE). Previously it was shown that the nuclear export signal (NES) domain is conserved among the lentiviral Rev proteins, in terms of its leucine-rich sequence composition. Moreover, the NES domain is interchangeable between lentiviral Revs, even though the spacing of the leucine residues within NES is atypical within EIAV (5). The C-terminal region of EIAV is important for nuclear localization and the central region is involved in RNA binding (2, 3, 8). To investigate interactions between EIAV Rev and the RRE in viral RNA, we characterized the RNA-binding activity of EIAV Rev and directly analyzed the roles of putative RNA-binding motifs for EIAV Rev binding for the first time, as described in Chapter 2 (12). Truncated Rev proteins were expressed as MBP-fusion proteins and tested for RNA binding in vitro using a UV crosslinking assay. RNA-binding activity was abrogated by deletions in either the central region or a region near the C-terminus that overlaps the nuclear localization signal. Analysis of site-specific mutations identified two short arginine-rich motifs (ARMs) that are essential for RNA-protein interaction: an RRDRW motif in the central region (residues 76-80), and a KRRRK motif near the C-terminus (residues 159-163). Each motif was found to be necessary, but not sufficient, for RNA binding in vitro and nuclear export activity in vivo. These results strongly suggest that the two ARMs interact with the RRE in concert, and thus comprise a bipartite RNA-binding domain. Mapping the motifs on a three-dimensional structural model of EIAV Rev (9) indicated that the two motifs are located in close proximity within the predicted three-dimensional structure of the folded protein, where they could form a single RNA-binding domain (Fig 5.1).

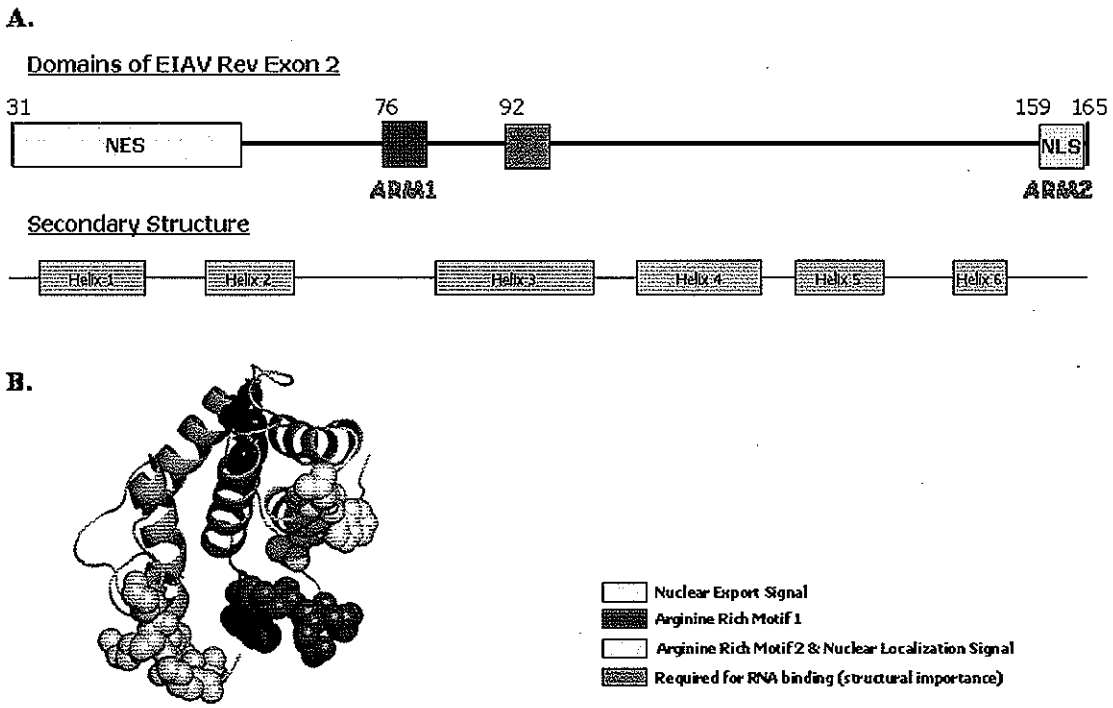


Figure 5.1. Summary of functional domains and predicted model of EIAV Rev Exon 2. **(A)** The functional domains of EIAV Rev Exon 2 are represented as boxes on a line representing the primary sequence (amino acids 31-165) **(B)** The predicted structure of EIAV Rev Exon 2 is illustrated using cartoon rendering. The NES, nuclear export signal; ARM1, arginine-rich motif 1 (RRDRW motif); NLS, nuclear localization signal (KRRRK motif); and L95, Leucine 95 and L109, Leucine 109, are shown in space-fill representation.

PROBING RNA SECONDARY STRUCTURE OF REV-RRE INTERACTIONS IN EIAV

During lentiviral replication, Rev-RRE interaction is crucial for the regulated expression of differentially spliced mRNAs. The RRE in most lentiviruses, including HIV-1, is located near the junction between SU and TM genes (13). Previously, the RRE in EIAV had been mapped in two different regions within the Env gene (15). Based on an in vivo Rev export assay, a 555 nt fragment designated ERRE-1, located in 5′ end of the Env gene has more functional activity (60%) than the rest of the Env gene (20%) (1, 2). In Chapter 3, to better understand the interaction between EIAV Rev and the RRE, we performed chemical probing experiments to determine the ERRE-1 secondary structure and footprinting experiments to analyze complexes formed between EIAV Rev and ERRE-1. Interestingly, the secondary structures for ERRE-1 based on the combination of computational and experimental approaches, are relatively unstable compared with the secondary structure predicted using computational approaches alone. There are several distinct and relatively stable stem-loop structures. In particular, the two stem-loop structures near the 3′ splice site and 5′ splice site for exon 1 of Rev not only are stable secondary structural elements based on both computational and experimental studies, but also are very well conserved in sequences of EIAV variants. Previously, several studies have implicated stem-loop structures in regulated splicing events, including alternative splicing of HIV Tat exon (10). Therefore we hypothesized that these secondary structure elements near the Rev exon 1 splice sites can act as cis-regulatory elements for splicing and alternative splicing events by interacting with trans-acting proteins such as splicing factors of the host. Previously, EIAV Rev was shown to bind to an Exonic Splicing Enhancer (ESE), containing two purine-rich tracts in the ERRE-1 (2, 3). We identified two high affinity RNA binding regions in the ERRE-1 by footprinting analysis and filter binding experiments. We call these regions RBR-1 (Rev binding region-1) and RBR-2. RBR-1 encompasses the ESE, identified as RNA binding in the previous studies mentioned above (2, 3, 7). Evidence from filter binding experiments suggested that the second region, RBR-2 actually has higher binding affinity for EIAV Rev than RBR-1. This is

particularly interesting because our comparative computational analyses identified an RNA structural motif in RBR-2, that is similar to one found in the HIV-1 high affinity Rev binding site (4, 19) and which is also found within or near RRE regions of four additional lentiviruses (HIV-2, SIV, FIV and OLV). This work thus provides the basis for future comparative analyses of lentiviral Rev-RRE interactions.

COMPUTATIONAL MODELING OF EQUINE MHC CLASS I MOLECULES AND EPITOPES OF REV, ENV AND GAG IN EIAV

The virus-specific Cytotoxic T Lymphocytes (CTL) response is the core immune response in virally-infected host. Viral infection induces the expression of MHC I complexes (MHC class I protein and processed viral peptides) on the surfaces of infected cell membranes. CTLs recognize the MHC I complexes and kill the infected cell. In EIAV infected horses, it has been shown that CTL epitopes in Rev, Env and Gag are crucial for immune control of lentiviruses (16, 17). Therefore, a detailed molecular understanding of the mechanism for recognition of CTL epitopes is important for the development of vaccines against lentiviruses. The key for differential recognition of different CTL epitopes by MHC class I proteins, is in polymorphisms that usually occur in the α-1/2 and β-1 domain of MHC class I molecules. In Chapter 4, two equine MHC I molecules that differ in only single amino acid within the α-2 domain, were identified and characterized to investigate how different MHC I molecules recognize Rev, Env and Gag epitopes during EIAV-specific CTL immune responses. Three different CTL peptides, Rev-QW11, Env-RW12 and Gag-GW12, were synthesized and tested to assess two different recognition processes: i) the recognition of MHC class I complexes by EIAV-specific CTL, and ii) the differential peptide binding affinities of MHC class I proteins. Briefly, the experiment results showed that a single

^{152}E→V substitution in the α-2 domain of the 141 molecule did not affect the recognition of Env-RW12 by CTL, but did result in more efficient recognition of Rev-QW11 by CTL. Also, CTL recognition of the Gag-GW12/141 MHC class I complex was abrogated, despite Gag-GW12 binding to 141 molecule.

Because the 3D structures of equine MHC class I molecules are not available, we used molecular modeling techniques to predict the 3D structures of equine MHC class I proteins and the three EIAV epitope peptides. Docking was performed using the modeled 7-6 and 141 MHC class I molecules with each of three epitope peptides to investigate the molecular basis for the differential experimentally determined binding affinities of MHC class I proteins for the peptides. Docking scores for each MHC class I molecule-peptide complex and detailed interaction statistics extracted from the proposed docking models supported the experimental observations. For example, the docking score difference between 7-6/Rev-QW11 and 141/Rev-QW11, and the larger number of salt bridges in the 141/Rev-QW11 complex docking model, can explain why the 141 molecule recognizes the Rev-QW11 more efficiently than the 7-6 molecule. In conclusion, molecular modeling provided a mechanistic explanation for experimental results in which a single amino acid change in an equine MHC class I molecule affected the equine CTL immune response; a similar phenomenon has been observed in the human CTL immune response to HIV-1 infection (6, 11).

FUTURE STUDIES

The combined results from the wet-lab experimental and computational approaches used in this study have brought us closer to understanding the macromolecular recognition mechanisms involved in protein-RNA and protein-peptide interactions during lentiviral

infection. In Chapter 2, we characterized the RNA binding domain of the EIAV Rev protein, and obtained important clues as to how the bipartite RNA binding domain in Rev specifically recognizes viral RNA. Our results suggest that EIAV Rev differs from HIV-1 Rev in terms of domain organization: two RNA binding domains separated in primary sequence are both required for high affinity binding to the EIAV RRE. Rev of EIAV is similar to that of HIV-1 in that short arginine rich motifs are important for the RRE binding and appear to contact the RNA directly. An intriguing discovery is that the L95 residue (located in the protein core, based on the predicted Rev structure), is required for RNA binding, but we don't yet understand why.

There are at least two hypotheses regarding the structure and function of Rev that should be tested in future analyses: First is the suggestion that the "bipartite" RNA binding domain forms a single RNA binding domain in the context of the folded structure. Our current hypothesis is that, even though they are 79 amino acids apart in primary sequence, the RRDRW and KRRRK motifs are brought into close proximity within the three-dimensional structure of folded protein (Fig 5.1). The best way to test this hypothesis would be to experimentally determine the 3D structure by NMR spectroscopy or X-ray crystallography. Another option would be to measure the distance between the two RNA binding domains using biophysical methods such as FRET. A second important hypothesis to test is related to the role of the leucine 95 in the RNA binding. We know that L95 is crucial for both ERev export function and RNA activity, based on a functional in vivo activity assay and an in vitro RNA binding assay. Our preferred hypothesis is that the L95 residue stabilizes EIAV Rev structure through hydrophobic interactions with other hydrophobic residues in neighboring alpha-helices in the protein core, based on our prediction of the secondary and tertiary structure of EIAV Rev. This is supported by our preliminary observation that the circular dichroism (CD) spectrum of L95D mutant Rev differs from that of wildtype. The best way to test this hypothesis is the same as for the first hypothesis: experimentally

determine the 3D structure of EIAV Rev. If it is not possible to obtain a complete structure, it may be possible to obtain important information regarding the distances between critical residues in Rev using NMR.

In Chapter 3, we characterized the secondary structure of the EIAV RRE and its interactions with EIAV Rev. This study complements the characterization of the RNA binding domain in EIAV Rev, described in Chapter 2, by revealing sequence and structural feature of Rev binding site in RNA. Therefore, I want to propose more detailed analyses that can be built on the results in Chapter 2. These should include: i) more detailed mapping to identify correlations between specific nucleotides in EIAV RRE and Rev amino acids that contact them, and ii) more systematic comparative analyses of lentiviral Rev-RRE interactions, using both experimental and computational approaches. A tantalizing observation is that RRE is also recognized by the host splicing factor, SF2/ASF (2, 3, 7, 14). Previously, two models have been proposed to explain interactions among EIAV RRE, Rev, and the host SF2/ASF protein. The first model is an exclusive model, in which the two proteins, EIAV Rev and SF2/ASF, directly compete with each other for binding to the EIAV RRE sequence. The other alternate, non-exclusive model, is that EIAV Rev and SF2/ASF can bind to the EIAV RRE at the same time, but that the binding of one protein affects the activity of other protein. Several indirect lines of evidence have been presented previously, but direct and detailed analysis has not been performed to test either model so far. Comparison of the footprinting results obtained on Rev/ERRE-1 versus SF2/ERRE-1 complexes, or direct binding competition assays between Rev and SF2 would give us a better understanding of the potentially competitive relationship between host proteins and EIAV proteins in regulating EIAV gene expression. In addition, this will be a good example for investigating how two different RNA-binding proteins may use the same (or different) recognition sites to interact with same RNA substrate.

Molecular modeling techniques are becoming increasing useful for analyzing and understanding experimental results regarding protein complexes for which high resolution structural data are lacking. In Chapter 4, we used homology modeling to analyze and rationalize how a single amino acid change in an equine MHC I molecule affects the recognition of three different viral epitope peptides, from Rev, Env and Gag proteins. The modeling processes used in this study provided rough models for equine MHC class I molecules and their cognate epitope peptides. I recommend more extensive searches for models be made by simulating dynamic features of MHC class I and peptide molecules using molecular dynamics approaches. Such results would be more reliable than the modeling procedures used here. Also, because only one docking method was applied in this study, it would be worthwhile to compare results obtained using different docking methods to model complexes of MHC class I molecules and epitope peptides.

REFERENCES

1. **Belshan, M., M. E. Harris, A. E. Shoemaker, T. J. Hope, and S. Carpenter.** 1998. Biological characterization of Rev variation in equine infectious anemia virus. J Virol **72:**4421-6.

2. **Belshan, M., G. S. Park, P. Bilodeau, C. M. Stoltzfus, and S. Carpenter.** 2000. Binding of equine infectious anemia virus rev to an exon splicing enhancer mediates alternative splicing and nuclear export of viral mRNAs. Mol Cell Biol **20:**3550-7.

3. **Chung, H., and D. Derse.** 2001. Binding sites for Rev and ASF/SF2 map to a 55-nucleotide purine-rich exonic element in equine infectious anemia virus RNA. J Biol Chem **276:**18960-7.

4. **Cook, K. S., G. J. Fisk, J. Hauber, N. Usman, T. J. Daly, and J. R. Rusche.** 1991. Characterization of HIV-1 REV protein: binding stoichiometry and minimal RNA substrate. Nucleic Acids Res **19:**1577-83.

5. **Fridell, R. A., K. M. Partin, S. Carpenter, and B. R. Cullen.** 1993. Identification of the activation domain of equine infectious anemia virus rev. J Virol **67:**7317-23.

6. **Gao, X., G. W. Nelson, P. Karacki, M. P. Martin, J. Phair, R. Kaslow, J. J. Goedert, S. Buchbinder, K. Hoots, D. Vlahov, S. J. O'Brien, and M. Carrington.** 2001. Effect of a single amino acid change in MHC class I molecules on the rate of progression to AIDS. N Engl J Med **344:**1668-75.

7. **Gontarek, R. R., and D. Derse.** 1996. Interactions among SR proteins, an exonic splicing enhancer, and a lentivirus Rev protein regulate alternative splicing. Mol Cell Biol **16:**2325-31.

8. **Harris, M. E., R. R. Gontarek, D. Derse, and T. J. Hope.** 1998. Differential requirements for alternative splicing and nuclear export functions of equine infectious anemia virus Rev protein. Mol Cell Biol **18:**3889-99.

9. **Ihm, Y., O. W. Sparks, J. H. Lee, W. Wannemuehler, M. Terribilini, H. Cao, C. Z. Wang, S. Carpenter, K.-H. Ho, and D. Dobbs.** 2007. Structural model of the Rev regulatory protein from Equine Infectious Anemia Virus (EIAV). In preparation.

10. **Jacquenet, S., D. Ropers, P. S. Bilodeau, L. Damier, A. Mougin, C. M. Stoltzfus, and C. Branlant.** 2001. Conserved stem-loop structures in the HIV-1 RNA region containing the A3 3' splice site and its cis-regulatory element: possible involvement in RNA splicing. Nucleic Acids Res **29:**464-78.

11. **Jin, X., X. Gao, M. Ramanathan, Jr., G. R. Deschenes, G. W. Nelson, S. J. O'Brien, J. J. Goedert, D. D. Ho, T. R. O'Brien, and M. Carrington.** 2002. Human

immunodeficiency virus type 1 (HIV-1)-specific CD8+-T-cell responses for groups of HIV-1-infected individuals with different HLA-B*35 genotypes. J Virol **76:**12603-10.

12. **Lee, J. H., S. C. Murphy, M. Belshan, W. O. Sparks, Y. Wannemuehler, S. Liu, T. J. Hope, D. Dobbs, and S. Carpenter.** 2006. Characterization of functional domains of equine infectious anemia virus Rev suggests a bipartite RNA-binding domain. J Virol **80:**3844-52.

13. **Lesnik, E. A., R. Sampath, and D. J. Ecker.** 2002. Rev response elements (RRE) in lentiviruses: an RNAMotif algorithm-based strategy for RRE prediction. Med Res Rev **22:**617-36.

14. **Liao, H. J., C. C. Baker, G. L. Princler, and D. Derse.** 2004. cis-Acting and trans-acting modulation of equine infectious anemia virus alternative RNA splicing. Virology **323:**131-40.

15. **Martarano, L., R. Stephens, N. Rice, and D. Derse.** 1994. Equine infectious anemia virus trans-regulatory protein Rev controls viral mRNA stability, accumulation, and alternative splicing. J Virol **68:**3102-11.

16. **Mealey, R. H., A. Sharif, S. A. Ellis, M. H. Littke, S. R. Leib, and T. C. McGuire.** 2005. Early detection of dominant Env-specific and subdominant Gag-specific CD8+ lymphocytes in equine infectious anemia virus-infected horses using major histocompatibility complex class I/peptide tetrameric complexes. Virology **339:**110-26.

17. **Mealey, R. H., B. Zhang, S. R. Leib, M. H. Littke, and T. C. McGuire.** 2003. Epitope specificity is critical for high and moderate avidity cytotoxic T lymphocytes associated with control of viral load and clinical disease in horses with equine infectious anemia virus. Virology **313:**537-52.

18. **Pollard, V. W., and M. H. Malim.** 1998. The HIV-1 Rev protein. Annu Rev Microbiol **52:**491-532.

19. **Tiley, L. S., M. H. Malim, H. K. Tewary, P. G. Stockley, and B. R. Cullen.** 1992. Identification of a high-affinity RNA-binding site for the human immunodeficiency virus type 1 Rev protein. Proc Natl Acad Sci U S A **89**:758-62.

APPENDIX A. STRIKING SIMILARITIES IN DIVERSE TELOMERASE PROTEINS REVEALED BY COMBINING STRUCTURE PREDICTION AND MACHINE LEARNING APPROACHES

A paper accepted in *Pacific Symposium on Biocomputing (PSB) 2008*

Jae-Hyung Lee, Michael Hamilton, Colin Gleeson, Cornelia Caragea, Peter Zaback, Jeffry D. Sander, Xue Li, Feihong Wu, Michael Terribilini, Vasant Honavar, Drena Dobbs

ABSTRACT

Telomerase is a ribonucleoprotein enzyme that adds telomeric DNA repeat sequences to the ends of linear chromosomes. The enzyme plays pivotal roles in cellular senescence and aging, and because it provides a telomere maintenance mechanism for ~90% of human cancers, it is a promising target for cancer therapy. Despite its importance, a high-resolution structure of the telomerase enzyme has been elusive, although a crystal structure of an N-terminal domain (TEN) of the telomerase reverse transcriptase subunit (TERT) from *Tetrahymena* has been reported. In this study, we used a comparative strategy, in which sequence-based machine learning approaches were integrated with computational structural modeling, to explore the potential conservation of structural and functional features of TERT in phylogenetically diverse species. We generated structural models of the N-terminal domains from human and yeast TERT using a combination of threading and homology modeling with the *Tetrahymena* TEN structure as a template. Comparative analysis of predicted and experimentally verified DNA and RNA binding residues, in the context of these structures, revealed significant similarities in nucleic acid binding surfaces of

Tetrahymena and human TEN domains. In addition, the combined evidence from machine learning and structural modeling identified several specific amino acids that are likely to play a role in binding DNA or RNA, but for which no experimental evidence is currently available.

INTRODUCTION

In most eukaryotes, a remarkable ribonucleoprotein enzyme, telomerase, is responsible for the synthesis and maintenance of telomeres, the ends of linear chromosomes (2, 5, 9). Many exciting discoveries have been made in telomerase biology since 1984, when the enzyme was first identified in the ciliate, *Tetrahymena thermophila*, by Greider and Blackburn(13). Recently, pivotal roles for telomerase in signaling pathways that regulate cancer, stress response, apoptosis and aging have been demonstrated (4, 6, 10, 27).

Two essential roles of telomeres are protecting or "capping" chromosome ends and facilitating their complete replication (reviewed in (2, 5, 9)). Typically, telomeres consist of arrays of simple DNA sequence repeats, ranging from ~50 copies of 5'-TTGGGG-3' in *Tetrahymena*, to ~1000 copies of 5'-TTAGGG-3' in humans and other vertebrates. The sequence of telomeric repeats is specified by an RNA template (TER), which varies in length from ~160 nts in ciliates to ~1500 nts in vertebrates, and is an essential component of the catalytically active form of telomerase (4, 9). Human telomerase is composed of hTER and two bound proteins, the telomerase reverse transcriptase component (hTERT) and dyskerin (8). The regulation of telomerase activity involves interactions with a variety of other cellular proteins, many of which are essential for telomere homeostasis (10, 16).

Telomerase is a promising target for cancer therapy because it is generally present in very low levels in normal somatic cells, but it is highly active in many human malignancies (12). Telomerase targeting strategies have included short interfering RNA (siRNA) knockdown of endogenous hTER and a combination of siRNA and expression of mutant

forms of the hTER RNA, which become incorporated into the enzyme and inhibit proliferation in variety of different human cancer cell lines (12).

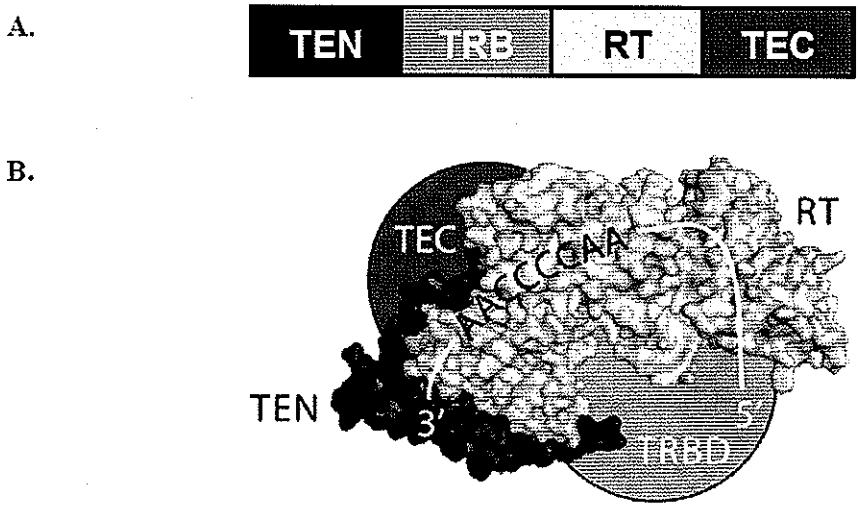


Figure A.1. TERT domain architecture. A) The telomerase reverse transcriptase (TERT) comprises 4 functional domains: essential N-terminal (TEN) domain, RNA-binding domain (TRBD), reverse transcriptase (RT), and C-terminal extension (TEC). B) Cartoon illustrating TERT domain organization, and the RNA template (TER). The TEN domain is *Tetrahymena* structure (PDB ID: 2B2A), and RT domain is from HIV-RT (PDB ID: 3HVT). Figure modeled after Collins, 2006 (9).

Despite its obvious clinical importance, currently there are no experimentally determined structures for the telomerase ribonucleoprotein complex or for telomerase complexes bound to telomeric DNA substrates, presumably because these are multisubunit structures. The telomerase reverse transcriptase component, TERT, is generally thought to consist of four functional domains (see Figure A.1): the essential N-terminal (TEN) domain, an RNA-binding domain (TRBD), reverse transcriptase (RT), and a C-terminal extension

(TEC). Recently, a crystal structure of the essential N-terminal domain of TERT from *Tetrahymena* has been reported (18) and appears to represent a novel protein fold. Several conserved sequence motifs have been identified within the TEN domain on the basis of multiple sequence alignments and mutagenesis experiments (11, 35). In addition, experiments directed at mapping DNA and RNA binding sites within TERTs from several organisms have identified specific amino acids that appear to contact either the DNA template or the RNA component (reviewed in(2)). In human telomerase, the TEN domain binds both DNA, specifically interacting with telomeric DNA substrates, and RNA, apparently binding in a non-sequence specific manner (18).

Although vertebrate TEN domain sequences share a high degree of sequence similarity, the TEN domains from more diverse species share very little sequence similarity (<30% identity), suggesting that a homology modeling approach to predicting the structure of the human TEN domain would be difficult. However, an alignment of the N-terminal sequences of TERTs from organisms ranging from human to *T. thermophila* to *S. cerevisiae*, revealed several highly conserved residues distributed throughout the N-terminal domain, suggesting that TEN domains from diverse organisms may share similar architectures (18). Based on this suggestion, we set out to test the hypothesis that the N-terminal domains of TERTs in diverse organisms not only share a similar overall three-dimensional fold, but may also have phylogenetically conserved DNA and RNA binding surfaces. We used a strategy in which comparative protein structural modeling approaches were integrated with sequence-based machine learning approaches for predicting DNA or RNA binding residues.

DATASETS, MATERIALS AND METHODS

Datasets

RNA-protein interface dataset

A dataset of protein–RNA interfaces was extracted from structures of known protein–RNA complexes in the Protein Data Bank (PDB) (3)solved by X-ray crystallography. Proteins with >30% sequence identity or structures with resolution worse than 3.5 Å were removed using PISCES(32). The resulting dataset, RB147 (31), contains 147 non-redundant polypeptide chains. RNA-binding residues were identified according to a distance-based cutoff definition: an RNA-binding residue is an amino acid containing at least one atom within 5 Å of any atom in the bound RNA. RB147 contains a total of 6157 RNA-binding residues and 26,167 non-binding residues. The RB147 dataset (31) is larger than the RB109 dataset used in our previous studies (29, 30).

DNA-protein interface dataset

A dataset of protein-DNA interfaces was extracted from structures of known protein-DNA complexes in the PDB(3). Proteins with >30% sequence identity or structures with resolution worse than 3.0 Å and R factor > 0.3 were removed using PISCES(32). The resulting dataset, DB208, contains 208 polypeptide chains, each at least 40 amino acids in length. DNA-binding residues were identified according to a definition based on reduction in solvent accessible surface area (ASA): an amino acid is a DNA-binding residue if its ASA computed in the protein-DNA complex using NACCESS (15) is less than its ASA in the unbound protein by at least 1 $Å^2$ (19). DB208 contains a total of 5,721 interface residues and 39,815 non-interface residues. The DB208 dataset is larger than the DB171 dataset used in our previous studies (36).

Algorithms for predicting interfacial residues

We used sequenced-based Naïve Bayes classifiers (22, 33) for predicting protein-RNA interfaces (29, 30) and protein-DNA interfaces (36). Briefly, the input to the classifier is a contiguous window of 2n+1 amino acid residues consisting of the target residue and n sequence neighbors to the left and right of the target residue, obtained from the protein sequence using the "sliding window" approach. The output of the classifier is a probability that the target residue is an interface residue given the identity of the 2n+1 amino acids in the input to the classifier. With Naïve Bayes classifiers, it is possible to tradeoff the rate of true positive predictions against the rate of false positive predictions, by using a classification threshold, θ, on the output probability of the classifier. The target residue is predicted to be an interface residue if its probability returned by the classifier is greater than θ, and a non-interface residue otherwise. The length of the window was set to 21 in the experiments described here.

We used the implementation of the Naive Bayes classifier available in WEKA, an open source machine learning package (33) for training classifiers used to predict interface residues in this study. The performance of the protein-RNA interface predictor trained on RB147 dataset (RNABindR, http://bindr.gdcb.iastate.edu/RNABindR/), and estimated using leave-one-out sequence-based cross-validation, is documented in (31). The performance of protein-DNA interface predictor trained on the DB208 dataset (DNABindR, http://cild.iastate.edu/DNABindR) and estimated using 10-fold sequence-based cross-validation, is comparable to that of the previously published protein-DNA interface predictor, which was trained on the DB171 dataset (36). The RNA interface predictions on TEN domains were obtained by using Naïve Bayes classifiers trained on the RB147 dataset (high specificity setting of RNAbindR). The DNA interface predictions were obtained by DNABindR (θ=0.168) trained on the DB208 dataset.

Structural modeling of telomerase TEN domains in human and yeast

The N-terminal domains from human telomerase (GENBANK NP_937986) and yeast telomerase (GENBANK NP_013422) sequences, were threaded onto the *T. thermophila* telomerase N-terminal domain (TEN) structure (PDB: 2b2a chain A) using FUGUE(28). The output alignments were used for generating 3D coordinates for the N-terminal domains of human and yeast telomerase by MODELLER (25). Among 15 generated models, the highest ranking model was chosen and refined using SCWRL (7) to reposition side-chains. Energy minimization was performed by 400 steps of steepest descent using the GROMOS96 force field (26) with a 9Å non-bonded cutoff in the Deep View/Swiss PDB-viewer (14). One human TEN model was based on the *Tetrahymena* TEN structure in the PDB: 2b2aA, N-terminal domain of tTERT. For a second model, several templates were selected using PSI-BLAST (1) and the Swiss-Model HMM template library (20) to detect remote homologs of hTERT. The chosen templates were portions of the following PDB structures: 1imhC, Tonicity-responsive enhancer binding protein (TONEBP)-DNA complex; 1jfiB, Negative Cofactor 2-TATA box binding protein-DNA complex (NC2-TBP-DNA); 2dyrM, bovine heart cytochrome C oxidase; 1b1uA, bifunctional inhibitor of Trypsin and Alpha-amylase from Ragi seeds; 2b2aA, N-terminal domain of tTERT. The templates were aligned and models were generated using the procedure described above. All generated structures were evaluated using the ANOLEA server (21).

Experimental identification of RNA and DNA binding residues

Experimentally determined DNA and RNA binding sites in hTERT and tTERT were collected by mining relevant literature. Point mutations that affect RNA binding have not been reported, but Moriarty et al. showed that deletions at positions 30-39 and 110-119 in hTERT result in reduced RNA and DNA association, respectively (23, 24). Conserved

primer grip regions have been mapped in the TEN and RT domains of hTERT, between amino acids 137-141 and 930-934 (34). Alanine substitutions in the C-terminal region of TEN at positions Q168, F178, and W187 have been shown to substantially decrease tTERT association with DNA (18).

RESULTS

Rationale

Computational and bioinformatic analyses can provide valuable insight into protein sequence-structure-function relationships, especially when the structure of a protein or complex is difficult to solve using experimental approaches. Surprisingly, despite the fascinating structural and regulatory complexity of telomerase, its pivotal role in cellular signal pathways, and its critical interactions with DNA, RNA and protein partners, very few studies have exploited bioinformatic or computational structural biology approaches to investigate the structure and function of telomerase. In this work, we use a combination of comparative structural modeling and sequence-based machine learning methods to test the hypothesis that the N-terminal domains of TERTs in diverse organisms share a similar overall architecture and conserved DNA and RNA binding surfaces.

Sequence-based prediction of RNA and DNA binding sites in human and *Tetrahymena* TERT

Conserved domains within the telomerase reverse transcriptase protein of human (hTERT) and *Tetrahymena* (tTERT) are illustrated in Figure A.2. In previous work, we used a sequence-based machine learning approach to predict RNA binding residues in TERT sequences and showed that our predictions compared favorably with available experimental

data(30). Results of these previously published predictions are included in Figure A.2 for comparison with DNA binding residues predicted in the current study (see Materials and Methods). The predicted DNA and RNA binding regions in hTERT and tTERT are indicated by boxes under the middle sections of Figures A.2A and A.2B, respectively. The lower portion of each figure shows specific examples, with boxed amino acids representing short deletions (in hTERT) or alanine-substitution mutations (in tTERT), that have been shown to compromise or abolish DNA binding. Note that for hTERT, the predictions either overlap or surround the amino acids implicated by deletion (Figure A.2A). For tTERT, two of three experimentally-identified DNA binding residues lie within the predicted DNA binding region (Figure A.2B).

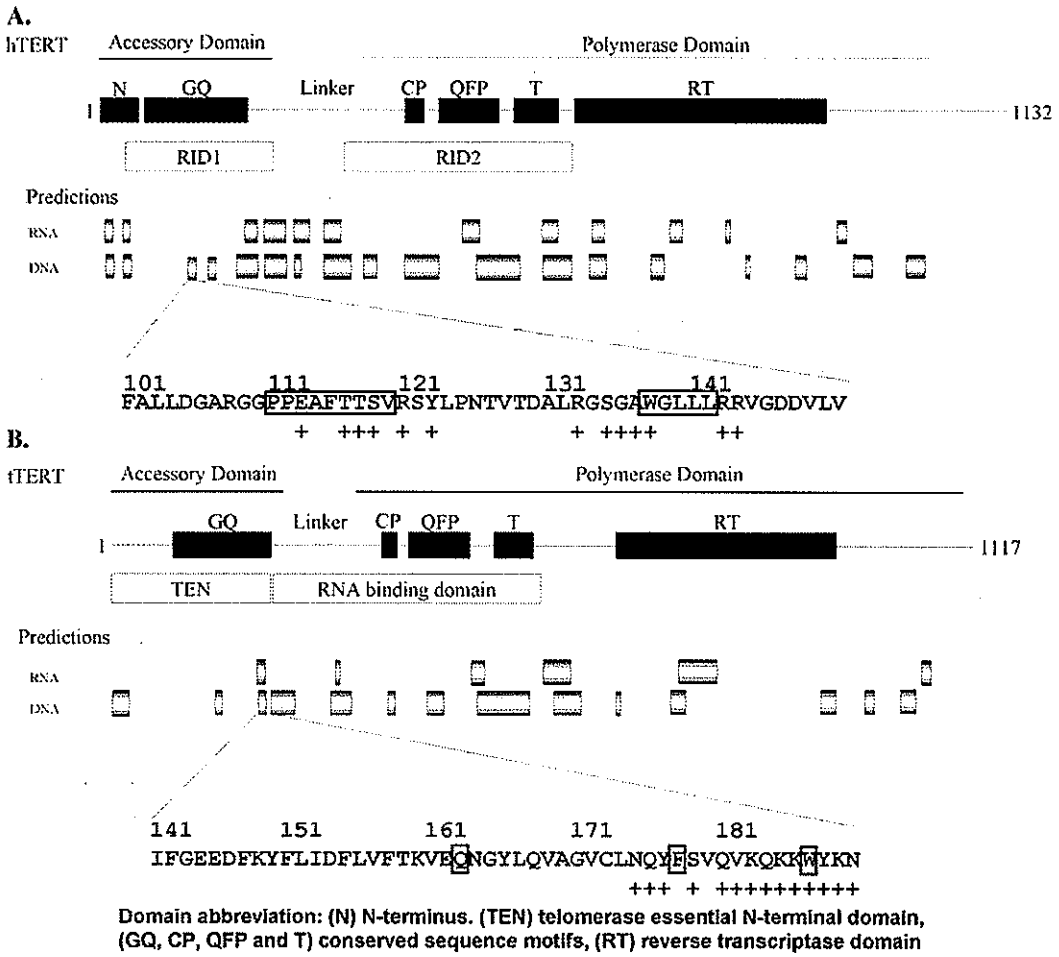


Figure A.2. Predicted interface residues and conserved domains for telomerase reverse transcriptase (TERT). Mapped functional domains and conserved motifs of TERT are shown above shaded boxes representing clusters of predicted RNA and DNA interface residues. Predicted interface residues are indicated by a + below the amino acid sequence. A) Human telomerase reverse transcriptase (hTERT). In the sequence shown, boxed amino acids 110-119 and 137-141, correspond to the template anchor site and a putative primer grip, implicated in forming the hTERT-DNA active complex (23, 24, 34). B) *Tetrahymena* telomerase reverse transcriptase (tTERT). The amino acid sequence shown represents the C-terminal end of the TEN domain. Alanine mutations at positions Q168, F178 and W187 have been shown to significantly reduce hTERT-DNA

association. Predicted interactions spanning amino acids 181-190 are located in a highly flexible, disordered region(18).

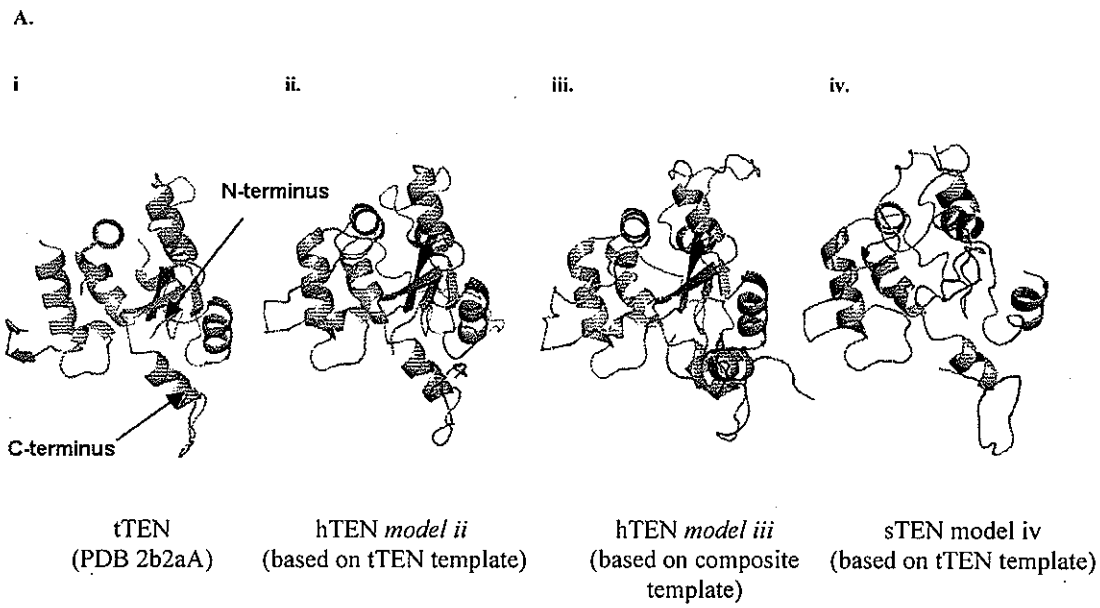


Figure A.3. Comparison of TEN domain structures and sequences and in *Tetrahymena*, human and yeast, *S. cerevisiae*. A) Comparison of *Tetrahymena* TEN domain structure determined by X-ray crystallography with modeled structures of TEN domains from other species. i) *T. thermophila*, experimentally-determined structure, PDB ID: 2b2aA (18); ii) human structural model, based on threading using the *T. thermophila* 2b2aA structure as template; iii) human structural model, based on threading using a composite of

several different structures as template; iv) yeast, *S. cerevisiae*, structural model, based on threading using the *T. thermophila* 2b2aA structure as template. B) Multiple sequence alignment of telomerase TEN domains from *T. thermophila*, *H. sapiens*, and *S. cerevisiae*(18). Amino acids conserved in all 3 species in the multiple sequence alignment are highlighted.

Structural modeling of N-terminal domain of TERT from human and yeast

Our initial attempts to generate structural models of the human and yeast TEN domains by submitting their sequences to several web-based homology modeling servers were unsuccessful, due to failure of the servers to identify appropriate homology modeling templates (the pairwise sequence identity between TEN domains of hTERT and tTERT is < 20%). However, the results of multiple sequence alignment (Figure A.3B) and predicted secondary structure similarities (data not shown), led us to try threading, using the FUGUE server (see Materials and Methods). The *Tetrahymena* TEN domain structure (PDB ID 2b2aA) was identified as the highest scoring structural template for both the human and yeast TEN domain sequences (hTERT: certain, with 99% confidence; sTERT: likely, with 95% confidence). Based on the alignments generated by FUGUE, we generated all-atom models and performed energy minimization to generate the final models illustrated in Figure A.3A (see Materials and Methods for details). Two different models for the human TEN domain, *model ii*, based on the *Tetrahymena* TEN template, and *model iii*, based on a composite template from several different structures, were very similar to one another as well as to model iv, for the yeast TEN domain, despite their highly divergent amino acid sequences. Table A.1 shows the root mean square deviation (RMSD) values calculated for comparison of the *Tetrahymena* TEN domain structure (determined by X-ray crystallography(18)) with the hTEN and sTEN modeled structures, using TOPOFIT (17) for structural alignment.

Aligned Structures	RMSD (Å)
tTEN *vs* hTEN	**1.11**
tTEN *vs* sTEN	**1.41**
sTEN *vs* hTEN	**1.39**

Table A.1. RMSD computed from structural alignments of TEN domain structures: tTEN, *Tetrahymena*, PDB structure, 2b2aA (Fig.3A, structure i); hTEN, human, modeled structure (Fig. 3A, *model ii*); sTEN, yeast, modeled structure (Fig. 3A, *model iv*). Alignments were performed using TOPOFIT (17)

Analysis of RNA and DNA binding surfaces in human and *Tetrahymena* TEN domains

To compare RNA and DNA binding surfaces in human and *Tetrahymena* TEN domains, we examined both our predicted nucleic acid binding sites and available experimental data in the context of the experimentally determined structure of *Tetrahymena* TEN domain (18) and modeled structure of the human TEN domain (*model ii*, Figure A.3A). Examples of these analyses are illustrated in Figures A.4 and A.5. The predicted RNA binding residues in hTEN overlap with several RNA binding sites implicated by deletion experiments (Figure A.4A, compare left and right models). Furthermore, additional putative RNA binding residues on the "back" side of the hTEN model (Figure A.4B, left, in oval) co-localize with an experimentally defined RNA binding site mapped onto the tTEN crystal structure (Figure A.4B, right, in oval).

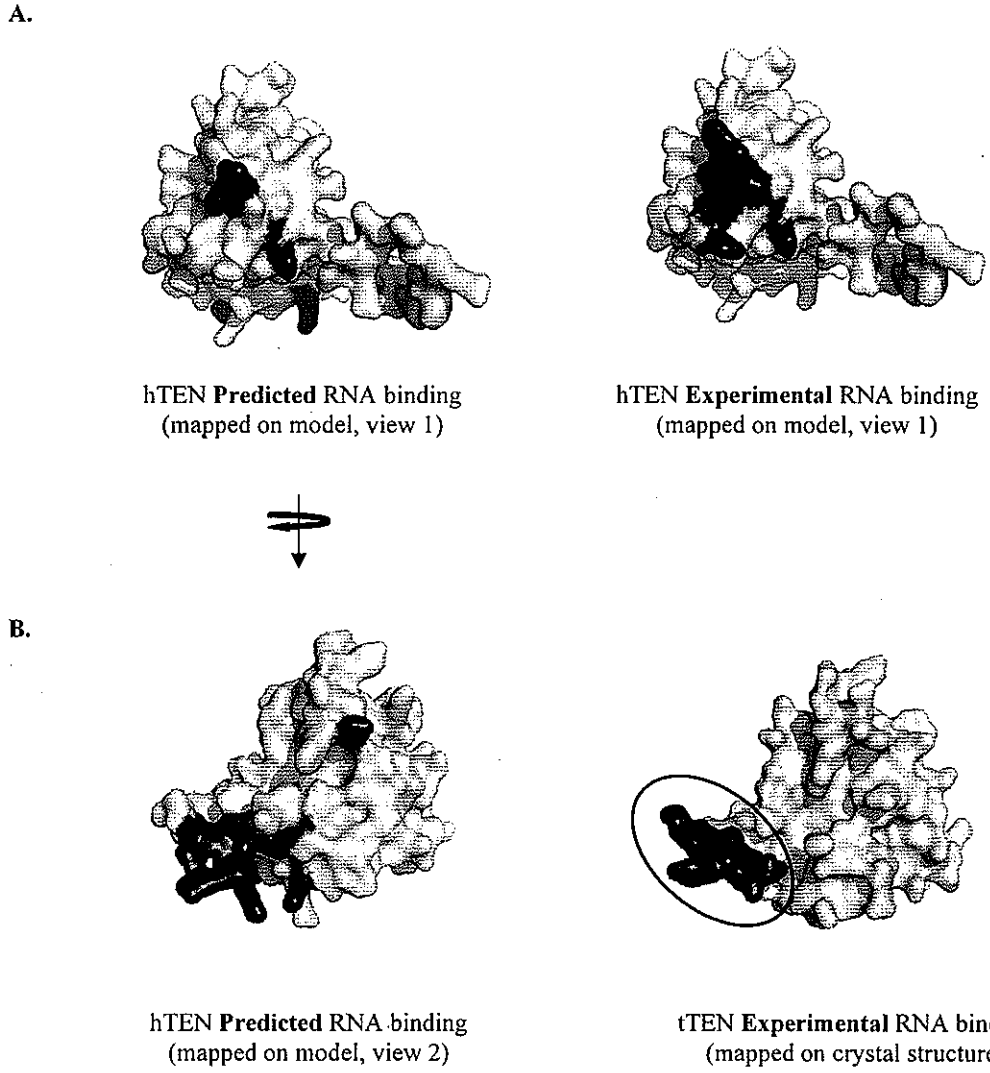
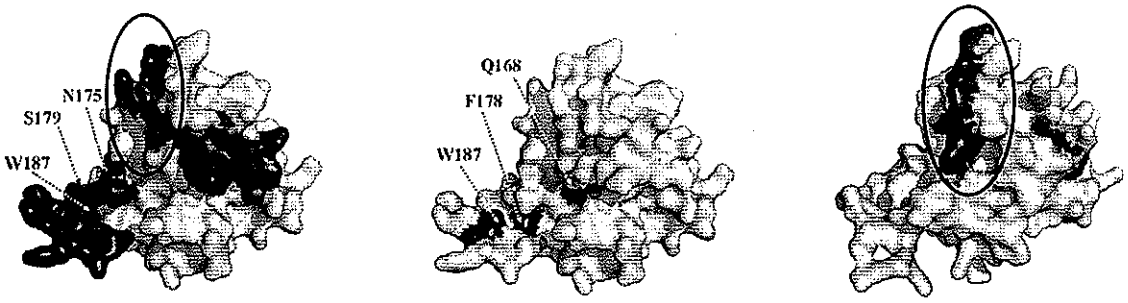


Figure A.4. Comparison of predicted and experimentally determined RNA binding surfaces in TEN domains. A) Sequence-based RNA binding site predictions mapped onto the hTERT TEN domain *model ii* (left) overlap with experimentally determined RNA binding residues (right); Black residues are predicted (left) or actual (right) RNA binding residues. B) Another patch of predicted RNA binding residues in the hTEN model (left, in oval) co-localizes with an experimentally verified RNA binding region in tTEN (right). Figures A.4 and A.5 were generated using PyMol (http://pymol.sourceforge.net/).



A. tTEN **Predicted** DNA binding (mapped on crystal structure)

B. tTEN **Experimental** DNA binding (mapped on crystal structure)

C. hTEN **Experimental** DNA binding (mapped on model, view 2)

Figure A.5. Comparison of predicted and experimentally determined DNA binding surfaces in TEN domains. A) Residues predicted to interact with DNA (black), mapped onto tTEN, PDB 2b2aA. Predicted binding sites encompass residues shown in B) which illustrates the only 3 experimentally defined DNA binding residues in tTEN (see Fig. 2B). Note that additional predicted DNA binding residues in A (in oval) are consistent with C), which shows experimentally validated DNA binding residues in the human protein mapped onto our modeled structure of hTEN.

Only three DNA binding residues in the TEN domain of tTERT have been experimentally identified: Q168, F178, and W187 (Figure A.5B). Several additional putative DNA binding residues are predicted by our machine learning classifiers (Figure A.5A). Some of these predicted residues in tTEN (in oval) co-localize with experimentally defined DNA binding residues in the human protein, when viewed in the context of our modeled structure of the hTEN domain (Figure A.5C).

Taken together, these results support our hypothesis that TEN domains in diverse organisms have similar three dimensional structures and conserved nucleic acid binding surfaces. Further, they identify additional putative interface residues that could be targeted in experiment studies.

SUMMARY AND DISCUSSION

Telomerase is one of several clinically important regulatory proteins for which it has been difficult to obtain high resolution structural information. The recent experimental determination of the structure of the N-terminal domain of tTERT, the telomerase reverse transcriptase component from *Tetrahymena*, suggests that at least partial structural information for human telomerase may soon become available. It seems unlikely, however, that experimental elucidation of the structure of the multisubunit RNP complex corresponding to the catalytically active form of telomerase will occur in the near future. Thus, the integrative strategy proposed here, in which structural information gleaned from comparative modeling is combined with machine learning predictions of functional residues, can be expected to provide valuable insights into the sequence and structural correlates of function for telomerase and other "recalcitrant" proteins. We are currently pursuing several avenues for improving the reliability of machine learning predictions, including the use of different sequence representations and additional sources of input information (e.g., structure and phylogenetic information, when available) and more sophisticated machine learning algorithms. We are also pursuing additional approaches for protein structure prediction, including ab initio and fold recognition methods capable of incorporating predicted protein-protein contacts as constraints. Given the large number of proteins with which telomerase interacts and the essential roles of telomerase in cellular signaling, aging, cancer, and other human diseases, this should continue to be rich and challenging area of research.

ACKNOWLEDGEMENT

This research was supported in part by NIH GM 066387, NIH-NSF BSSI 0608769, NSF IGERT 0504304 and by the ISU Center for Integrated Animal Genomics. We thank

Fadi Towfic for critical comments on the manuscript and members of our groups for helpful discussions.

REFERENCES

1. **Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman.** 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res **25:**3389-402.

2. **Autexier, C., and N. F. Lue.** 2006. The structure and function of telomerase reverse transcriptase. Annu Rev Biochem **75:**493-517.

3. **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.

4. **Blackburn, E. H.** 2005. Telomerase and Cancer: Kirk A. Landon - AACR Prize for Basic Cancer Research Lecture. Mol Cancer Res **3:**477-482.

5. **Blackburn, E. H.** 2005. Telomeres and telomerase: their mechanisms of action and the effects of altering their functions. FEBS Letters **579:**859.

6. **Blasco, M. A.** 2007. The epigenetic regulation of mammalian telomeres. Nat Rev Genet **8:**299-309.

7. **Canutescu, A. A., A. A. Shelenkov, and R. L. Dunbrack, Jr.** 2003. A graph-theory algorithm for rapid protein side-chain prediction. Protein Sci **12:**2001-14.

8. **Cohen, S. B., M. E. Graham, G. O. Lovrecz, N. Bache, P. J. Robinson, and R. R. Reddel.** 2007. Protein composition of catalytically active human telomerase from immortal cells. Science **315:**1850-3.

9. **Collins, K.** 2006. The biogenesis and regulation of telomerase holoenzymes. Nat Rev Mol Cell Biol **7:**484-94.

10. **de Lange, T.** 2005. Shelterin: the protein complex that shapes and safeguards human telomeres. Genes Dev **19:**2100-10.

11. **Friedman, K. L., and T. R. Cech.** 1999. Essential functions of amino-terminal domains in the yeast telomerase catalytic subunit revealed by selection for viable mutants. Genes Dev **13:**2863-74.

12. **Goldkorn, A., and E. H. Blackburn.** 2006. Assembly of Mutant-Template Telomerase RNA into Catalytically Active Telomerase Ribonucleoprotein That Can Act on Telomeres Is Required for Apoptosis and Cell Cycle Arrest in Human Cancer Cells. Cancer Res **66:**5763-5771.

13. **Greider, C. W., and E. H. Blackburn.** 1985. Identification of a specific telomere terminal transferase activity in *Tetrahymena* extracts. Cell **43:**405-13.

14. **Guex, N., and M. C. Peitsch.** 1997. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis **18:**2714-23.

15. **Hubbard, S. J., S. F. Campbell, and J. M. Thornton.** 1991. Molecular recognition. Conformational analysis of limited proteolytic sites and serine proteinase protein inhibitors. J Mol Biol **220:**507-30.

16. **Hug, N., and J. Lingner.** 2006. Telomere length homeostasis. Chromosoma **115:**413-425.

17. **Ilyin, V. A., A. Abyzov, and C. M. Leslin.** 2004. Structural alignment of proteins by a novel TOPOFIT method, as a superimposition of common volumes at a topomax point. Protein Sci **13:**1865-74.

18. **Jacobs, S. A., E. R. Podell, and T. R. Cech.** 2006. Crystal structure of the essential N-terminal domain of telomerase reverse transcriptase. Nat Struct Mol Biol **13:**218-25.

19. **Jones, S., and J. M. Thornton.** 1996. Principles of protein-protein interactions. Proc Natl Acad Sci U S A **93:**13-20.

20. **Kopp, J., and T. Schwede.** 2004. The SWISS-MODEL Repository of annotated three-dimensional protein structure homology models. Nucleic Acids Res **32:**D230-4.

21. **Melo, F., and E. Feytmans.** 1998. Assessing protein structures with a non-local atomic interaction energy. J Mol Biol **277:**1141-52.

22. **Mitchell, T.** 1997. Machine Learning. McGraw-Hill.

23. **Moriarty, T. J., S. Huard, S. Dupuis, and C. Autexier.** 2002. Functional multimerization of human telomerase requires an RNA interaction domain in the N terminus of the catalytic subunit. Mol Cell Biol **22:**1253-65.

24. **Moriarty, T. J., R. J. Ward, M. A. Taboski, and C. Autexier.** 2005. An anchor site-type defect in human telomerase that disrupts telomere length maintenance and cellular immortalization. Mol Biol Cell **16:**3152-61.

25. **Sanchez, R., and A. Sali.** 1997. Evaluation of comparative protein structure modeling by MODELLER-3. Proteins **Suppl 1:**50-8.

26. **Scott, W. R. P., P. H. Hunenberger, I. G. Tironi, A. E. Mark, S. R. Billeter, J. Fennen, A. E. Torda, T. Huber, P. Kruger, and W. F. van Gunsteren.** 1999. The GROMOS Biomolecular Simulation Program Package. J. Phys. Chem. A **103:**3596-3607.

27. **Shay, J. W., and W. E. Wright.** 2007. Hallmarks of telomeres in ageing research. J Pathol **211:**114-23.

28. **Shi, J., T. L. Blundell, and K. Mizuguchi.** 2001. FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. J Mol Biol **310:**243-57.

29. **Terribilini, M., J. H. Lee, C. Yan, R. L. Jernigan, S. Carpenter, V. Honavar, and D. Dobbs.** 2006. Identifying interaction sites in "recalcitrant" proteins:

predicted protein and RNA binding sites in rev proteins of HIV-1 and EIAV agree with experimental data. Pac Symp Biocomput:415-26.

30. **Terribilini, M., J. H. Lee, C. Yan, R. L. Jernigan, V. Honavar, and D. Dobbs.** 2006. Prediction of RNA binding sites in proteins from amino acid sequence. Rna **12:**1450-62.

31. **Terribilini, M., J. D. Sander, J. H. Lee, P. Zaback, R. L. Jernigan, V. Honavar, and D. Dobbs.** 2007. RNABindR: a server for analyzing and predicting RNA-binding sites in proteins. Nucleic Acids Res **35:**W578-W584.

32. **Wang, G., and R. L. Dunbrack, Jr.** 2003. PISCES: a protein sequence culling server. Bioinformatics **19:**1589-91.

33. **Witten, I. H. a. F., E.** 2005. Data Mining: Practical Machine Learning Tools and Techniques, 2nd ed. Morgan Kaufmann.

34. **Wyatt, H. D., D. A. Lobb, and T. L. Beattie.** 2007. Characterization of physical and functional anchor site interactions in human telomerase. Mol Cell Biol **27:**3226-40.

35. **Xia, J., Y. Peng, I. S. Mian, and N. F. Lue.** 2000. Identification of functionally important domains in the N-terminal region of telomerase reverse transcriptase. Mol Cell Biol **20:**5196-207.

36. **Yan, C., M. Terribilini, F. Wu, R. L. Jernigan, D. Dobbs, and V. Honavar.** 2006. Predicting DNA-binding sites of proteins from amino acid sequence. BMC Bioinformatics 7:262.

APPENDIX B. CHARACTERIZATION OF ISOMERIZING PROLINES USING SEQUENCE AND STRUCTURE INFORMATION

Preliminary data and results described

ABSTRACT

In polypeptide chains, there are two potential configurations (*cis* and *trans*), for the peptide bond that connects neighboring amino acids, *cis* and *trans*. For thermodynamic reasons, the *cis* configuration is almost never observed for amino acids in stably folded proteins, with the notable exception of peptidyl-prolyl bonds, which are observed in the *cis* configuration for ~5% of prolines that have been characterized. The structure of proline, which lacks an amide hydrogen atom, lowers the thermodynamic barrier between *cis* and *trans* peptide bond configurations. Peptidyl-prolyl *cis/trans* isomerization has considerable biological significance in the regulation of protein folding and in protein structure stabilization. In some cases, however, proline residues have been shown to undergo *cis/trans* isomerization within stably folded proteins, a change that has been shown to alter the recognition properties of several proteins. Consequently, it has been suggested that proline *cis/trans* isomerization could represent a novel post-translational regulatory mechanism. Here we present a systematic characterization of the amino acid sequence correlates of peptidyl-prolyl *cis/trans* isomerization. In addition, we use machine learning approaches to generate classifiers for predicting whether a specific prolyl peptide bond is likely to be in the *trans*, *cis* or *isomerizing (isom)* configuration, using only the amino acid sequence of a query protein as input. The configuration of every peptidyl-prolyl bond for all proline residues (>1 million) in the PDB database of protein structures was determined and comprehensive

datasets for training and testing algorithms to perform *cis* versus *trans* and *cis* or *trans* versus *isom* classification were generated. Two machine leaning algorithms, Naïve Bayes and Support Vector Machine (SVM), were used for constructing classifiers and 4 different features were used as input for classifiers. Analysis of the secondary structure environment and the propensity for certain amino acid residues to be located at specific positions relative to the target proline residue showed that proline residues that undergo *cis/trans* isomerization have a distinguishable secondary structure environment and a different neighboring amino acid propensity distribution from *cis* and *trans* cases. In classification tasks, the *cis/trans* classifier outperformed either *isom/trans* or *isom/cis* classifiers. Generally, SVM classifiers and using a combination of different physicochemical features as input gave the best classification results.

INTRODUCTION

Post-translational modifications of proteins, such as phosphorylation and glycosylation, represent an important mechanism for regulating the cellular functions of proteins. Recently, a new potential regulatory "switch" has been proposed: a conformational change mediated by *cis/trans* isomerization of proline residues (1, 2). Proline is unique among amino acids in its capacity to accommodate the thermodynamic change between *cis* and *trans* peptide bond configurations because it lacks the amide hydrogen atom (8). Although most proline residues are observed in either the *trans* (95%) or the *cis* (5%) conformation, conversion of the planar peptide bond of a proline residue between the *trans* and *cis* form is thermodynamically feasible, and this phenomenon is referred to as *cis/trans* isomerization (Fig. B.1). The ability to monitor protein dynamics using NMR spectroscopy led to the discovery of a "proline switch" in the Itk tyrosine kinase, in which proline *cis/trans* isomerization determines the specificity of Itk protein-protein interactions by regulating the transition between two alternative binding surfaces (14). Generally, proline isomerization

plays a important role in many biological functions such as protein folding (19, 30) and cell signaling (16, 20, 32).

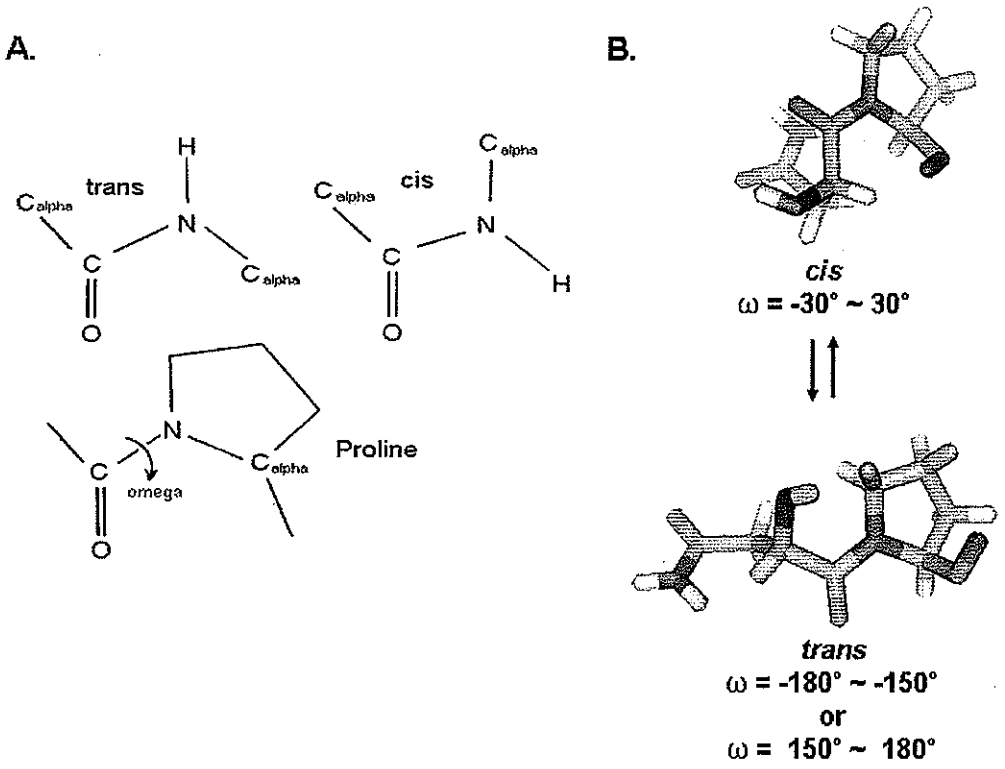


Figure B.1. *Cis* and *trans* configurations of a proline residue. **(A)** There are two configurations of the C-N bond in polypeptides. **(B)** Based on the omega (ω) angle, proline residues can have either configuration: cis or trans.

Several computational approaches have been used to analyze *cis/trans* conformations of prolyl peptide bonds in protein structures. Peptide bond conformation at proline residues can be investigated using 3D structures of proteins that have been solved using high-resolution experimental structure determination methods (e.g., from the PDB database (4). In addition, several studies have reported the analysis of the prolyl peptide bonds using other

biochemical or biophysical methods. The first data-driven computational analysis of sequence determinants of proline isomerization was reported by Frommel, et al. (9). They analyzed the distribution of amino acids in the neighborhood of prolyl residues, using a dataset extracted from the PDB, and calculated the propensity for certain amino acid residues to be located in the neighborhood of *cis*-prolyl residues. Based on these amino acid propensities, they tried to predict the conformation of prolyl peptide bonds on the basis of sequence, but their dataset was too small to generate reliable predictions. Recently, the same group updated the study and showed that *cis*-prolyl bonds in proteins are well-conserved during evolution, compared with *trans*-prolyl residues (13). Pahlke, et al. (17) analyzed *cis* and *trans* conformations of proline residues using Chou-Fasman parameterization, an occurrence matrix for amino acids, and secondary structures. Recently, Wang et al. (29), used a Support Vector Machine (SVM) algorithm to construct classifiers for predicting the *cis/trans* conformation of proline residues. Enhanced predictions on *cis/trans* conformation have done by Song, et al. (23)..In their study, different machine leaning algorithms and various data features for *cis* and *trans* prolyl residues were systematically tested. No studies to date have attempted to systematically identify proline residues that can undergo *cis/trans* isomerization within stably folded proteins, or to distinguish such proline residues from those that are thought to exist in only the *cis* or only the *trans* conformation.

At the time we initiated this study, *cis/trans* isomerization had been observed in only a handful of proteins. This is at least partly due to the technical difficulty of detecting the phenomenon: protein crystallization would be expected to "freeze" potentially isomerizing prolines in a single configuration. Proline isomerization would be detected only if two different structures produced by X-ray crystallography happen to "capture" a proline in two different configurations. Only with NMR spectroscopy is it possible to directly measure cis/trans isomerization and evaluate it dynamically. Therefore, we set out to explore whether proline-mediated isomerization might actually occur at a significant frequency in folded

proteins by performing a systematic examination of the structural environment of every proline residue in *all* known protein structures. Machine learning approaches were used to differentiate three (rather than just two) different configuration of prolyl peptide bonds; *cis*, *trans* and isomerization (*isom*) cases, using either sequence or a combination of sequence and structural information extracted from the PDB. The ultimate goal of this study is to develop a computational model that can be used to identify additional "proline-switch" candidates, which can then be experimentally evaluated, using NMR spectroscopy.

MATERIALS AND METHODS

Exploring and analyzing proline configurations in PDB

To generate data for our analyses of the sequence correlates of each of the three different proline conformations, *cis*, *trans* and *isom*, we extracted datasets from the PDB (April 20th, 2007). At that time, there were 42,861 PDB entries, including 92,210 protein chains that are greater than 20 amino acids in length. Every proline residue in all chains was analyzed to calculate the omega angle of the peptide bond. Each proline entity in our database of extracted peptidyl proline bonds contains the PDB chain ID, the absolute omega angle value, the amino acid identities of residues in a 21 amino acid window containing the target proline residue (including 10 residues before and after the proline), the corresponding secondary structure information for residues in the 21 amino acid window, and the calculated configuration of the prolyl peptide bond (if the absolute omega angle value of the prolyl peptide bond is $< 30°$, the bond is classified as in *cis* configuration; if it is $>150°$, the bond is classified as in *trans* configuration). A total 1,019,085 proline residues were extracted and analyzed. Based on the omega angle, 948,081 (93%) of these proline residues are in the *trans* configuration and 45,353 (4.5%) are in *cis* configuration. The omega angle for 25,651 prolyl peptide bonds could not be classified as *cis* or *trans*. To generate a dataset for the

isomerization cases (i.e., proteins in which a specific proline has been observed in both the *cis* and *trans* configurations) it was necessary to collect "redundant" structures. Therefore, to compare different protein structures that have the same or almost the same sequences, datasets of clustered protein sequences, available from the PDB were used. These are generated by an algorithm that clusters all protein sequences in PDB, based on sequence similarities (to reduce redundancy of the PDB database for certain applications). Clusters of proteins with ≤95%, ≤90%, ≤70%, and ≤50% sequence homology are generated. For generating our primary dataset, the 90% sequence homology cluster dataset and a 2.5 A° resolution cutoff for structures determined by X-ray crystallography were used. In the PDB cluster data, there are 15,271 different clusters and within each cluster, protein chains that have more than 90% sequence homology are listed. Initially, all 21 amino acid windows from all protein chains in one cluster were extracted and the 21 amino acid windows were compared using pair-wise sequence alignment. If two *identical* 21 amino acid windows from the same cluster have *different* configurations of target proline (i.e., one *trans* and one *cis*), we assigned the proline centered in this 21 amino acid window the *isom* case label. Using this procedure, 460 unique isomerization cases were collected and used for further structural and sequence analysis.

Generation of datasets for classification tasks

To avoid redundancy problems in the classification tasks, the entire dataset of 21 amino acid windows generated to evaluate the relative frequency of different proline configurations was not used. Instead, protein chains from PISCES server (28) (less than 30% sequence identity and structures with resolution greater than 2.5A°) were collected and used for identifying 21 amino acid windows (hereafter referred to as "windows") for further analysis. This resulted in total 46,472 windows containing 44,148 *trans* cases, 2,140 *cis* cases

and 184 *isom* cases. Four different features associated with each instance were used for construction of dataset. The sequence information, i.e., amino acid identity, was the primary data feature and always used for the construction of dataset and derivation of other features that were explored for providing better performance. Secondary structure information was obtained using DSSP software (12) to analyze each PDB structure and the output was pre-processed and presented simply as an assignment of one of three basic secondary structure classes; helix (H), sheet (E) and coil (C). The other two data features, extracted from the HSSP database (21), were sequence entropy information and profiles for each amino acid residue. The HSSP database is a database that combines three dimensional structure and primary sequence information for proteins. Each entity in HSSP database has a multiple sequence alignment from sequence homologs and a value representing the sequence variability at each position, calculated based on the alignment. Because the HSSP database entity originally came from PDB structures, the information from HSSP database contains evolutionary information based not only on sequence information but also on the structural conservation of the protein. The entropy for each amino acid position represents how well-conserved an amino acid is in the multiple sequence alignment. The HSSP profile has more information about the conservation of the amino acid position than the entropy, because the profile for each amino acid position includes 20 different vectors that encode the frequency of occurrence of each of the 20 amino acid in a multiple sequence alignment.

Five different combinations of data features were evaluated in the classification tasks: i) sequence alone (seq); ii) sequence + secondary structure (seq+ss); 3) sequence + HSSP entropy (seq+hssp_ent); iv) sequence + HSSP profile (seq+hssp_pro); v) sequence + secondary structure + HSSP profile (seq+ss+hssp_pro). To examine the effect of window length (for local amino acid features near the target proline) for constructing classifiers, nine different window lengths (3, 5, 7, 9, 11, 13, 15, 17, 19, 21 aa), with the target proline always in the middle of the window, were generated and evaluated.

Naïve Bayes method

Naïve Bayes learning is based on Bayes theorem (15). Bayesian learning infers the optimal hypothesis by weighing the evidence supporting alternative hypothesis. The probability of a hypothesis given a set of data D, $P(h|D)$, can be calculated based on its prior probability $P(h)$, the probability of observing the data given the hypothesis $P(D|h)$ and the probability of the data $P(D)$ using Bayes theorem.

$$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)}$$

In a given data set $D = \{X_1, X_2, \ldots, X_N\}$, naïve Bayes method estimates the probability of each hypothesis $P(h_i)$ for all $h_i \in$ hypothesis space H, Given an instance X with attribute value, $\{a_1 = x_1, a_2 = x_2, \ldots, a_n = x_n\}$ the Bayesian approach to classify X is to assign to it the most probable hypothesis h_{MAP},

$$\begin{aligned} h_{MAP} &= \arg\max P(h_i \mid a_1 = x_1, a_2 = x_2, \ldots, a_n = x_n) \\ &= \arg\max \frac{P(a_1 = x_1, a_2 = x_2, \ldots, a_n = x_n \mid h_i)P(h_i)}{P(aa_1 = x_1, a_2 = x_2, \ldots, a_n = x_n)} \\ &= \arg\max P(a_1 = x_1, a_2 = x_2, \ldots, a_n = x_n \mid h_i)P(h_i) \end{aligned}$$

for all $h_i \in$H.

If we assume that class attributes are independent of each other, and then,

$$\begin{aligned} h_{MAP} &= \arg\max P(a_1 = x_1, a_2 = x_2, \ldots, a_n = x_n \mid h_i)P(h_i) \\ &= \arg\max_{\text{for all } i} P(h_i)\prod_{i=1}^{n} P(a_i = x_i \mid h_i) \end{aligned}$$

for all $h_i \in$H.

Sometimes this assumption may not be true in real situations, but various Naïve Bayes applications have shown good performance in many practical examples and have good

computational efficiencies. Recently, bioinformatic applications based on Naïve Bayes method showed good performance for prediction of interface residues in protein-protein, protein-DNA and protein-RNA interactions (e.g., 24-26, 33, 34).

Support Vector Machine (SVM) method

Support Vector Machine, which is based on the perceptron algorithm, tries to find the best hyperplane boundary between classes in a high dimensional space (27). If the class patterns are not be separable in the original n-dimensional pattern space, a non-linear kernel function can be used to implicitly map the patterns in the n-dimensional input space into a higher dimensional feature space in which the patterns become separable. (Fig B.2A). Also in the SVM, the overfitting problem for the training dataset is resolved by selecting the hyperplane that maximizes the margin of separation between two classes from among all separating hyperplanes (Fig B.2B).

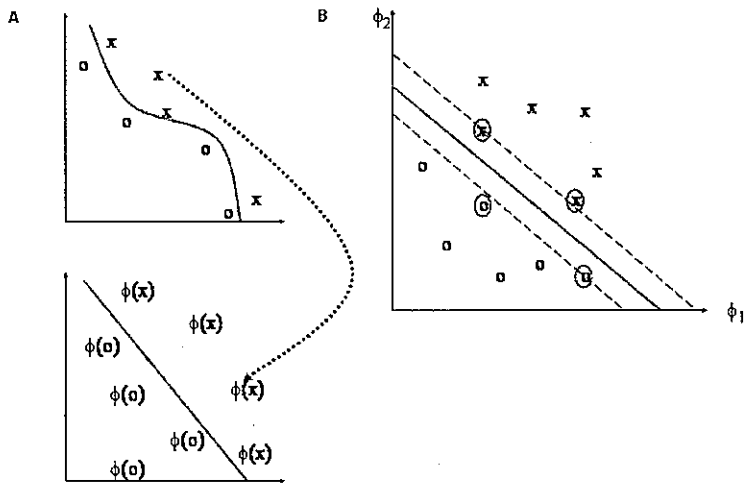


Figure B.2. The role of kernel function in SVM and maximization of the margin. **(A)** The kernel function can map linearly non-separable class patterns onto a high dimensional space in which the classes can

be separated. **(B)** A maximal margin hyperplane with its support vector highlighted in the 2-dimensional feature space ($\phi 1$, $\phi 2$).

Recently, several biological classification problems such as protein function classification (7), protein secondary structure prediction (10), sub-cellular localization prediction (11), and protein-protein interaction sites prediction (33) were analyzed using different machine learning classification algorithms.

Evaluating machine learning classification performance

Previous studies (3, 33) mentioned various measures for evaluating the performance of classifiers. Generally, the measurements include: 1) sensitivity for a class: the probability of correctly predicting an example of that class, 2) specificity for a class: the probability that a positive prediction for that class is correct, 3) accuracy: the overall probability that a prediction is correct, and 4) Matthew's correlation coefficient: a measure of how predictions correlate with the actual data.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$Correlation\ Coefficient = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}}$$

$$Specificity^{POS} = \frac{TP}{TP + FP}$$

$$Sensitivity^{POS} = \frac{TP}{TP + FN}$$

In our performance measures, we define,

- TP (true positives): the number of proline residues predicted to be members of the positive classes (*cis* or *isom* cases) that actually are members of positive classes (*cis* or *isom* cases).

- TN (true negatives): the number of proline residues predicted to be members of negative class (*trans* cases) that actually are members of the negative class (*trans* cases)

- FP (false positive): the number of proline residues predicted to be positive (*cis* or *isom*) cases that actually are negative (*trans)* cases.

- FN (false negative): the number of proline residues predicted to be negative (*trans*) cases that actually are positive classes (*cis* or *isom*) cases.

The correlation coefficient has values ranging from -1 to 1, and the more accurate the predictions are, the closer correlation coefficient is to 1. On the other hand, a correlation coefficient of -1 means the opposite prediction is always obtained, and a correlation coefficient of 0 means the predictions are random. Because any single measure is not sufficient to adequately evaluate a classifier, all of the above measurements are used.

Analysis of the distribution of amino acid types and secondary structure near proline residues

The datasets for classification were first used to analyze the distribution of amino acid types and secondary structural features in the neighborhood of proline residues. There are three different classes (*trans*, *cis* and *isom)* of 21 amino acid windows (containing the target proline residue in the middle). The frequency of occurrence of each of the 20 amino acids or of each of the 3 secondary structural assignments (helix, sheet and coil) at each residue

position in the windows for each class (*trans*, *cis* and *isom*), was determined and normalized, based on the natural distribution of amino acids. The natural distribution of 20 amino acids was calculated based on the entire PISCES non-redundant dataset. The calculated distributions are presented as a 20 X 20 matrix for each class (Table B.1) and if the value in a cell is > 1, it means the amino acid at that position is over-represented relative to other amino acids.

RESULTS

Distribution of amino acids and secondary structures near proline residues

Table B.1 shows the normalized propensities calculated for 10 residues preceding and 10 residues following the central proline residue in datasets of 21 aa windows. The columns for amino acids are arranged on the table column on the basis their hydrophobicity; Phenylalanine (F) is the most hydrophobic amino acid and the arginine (R) is the most hydrophilic residue (5). The over-represented amino acids (values >1) are colored red and under-represented amino acids (values <1) are colored blue. The amino acid propensities for each of the 3 proline configurations has a different distribution trend. In *trans* cases, hydrophobic residues (FLIYWV) are overrepresented within window and hydrophilic residues (except arginine) are underrepresented. The overall trend of propensities for *cis* cases is similar to that of *trans*, but two hydrophobic residues, leucine (L) and isoleucine (I) and two hydrophilic residues, asparagine (N) and histidine (H), have the opposite propensities, relative to the *trans* cases. In the *isom* cases, there is no general tendency based on the hydrophobicity of amino acids, but a definite "signal" exists in the propensity distribution, with specific favorable and unfavorable amino acids readily detectable.

	F	L	I	Y	W	V	M	P	C	A	G	T	S	K	Q	N	H	E	D	R
-10	1.0267	1.0319	1.0577	1.0466	1.1458	1.0587	0.9506	0.9935	1.0827	1.0362	1.0377	1.0001	0.9149	0.9896	1.0015	0.96	0.9726	0.9522	0.9776	1.0422
-9	1.081	1.0778	1.0198	1.0888	1.2084	1.0866	0.9525	0.9797	1.0699	1.0006	1.0284	1.0039	0.907	0.9872	0.9671	0.9804	0.9824	0.9291	0.93	1.0216
-8	1.0804	1.0896	1.0924	1.1495	1.2002	1.0868	0.7902	0.9787	1.0699	1.0009	1.0167	0.9361	0.9384	0.9935	0.9599	0.9822	0.9716	0.9868	0.9154	1.0229
-7	1.0945	1.0862	1.0533	1.0677	1.1491	1.1047	0.8056	1.017	1.0093	0.9846	1.0037	1.0204	0.9613	0.942	0.9254	0.9552	0.9402	0.9577	0.971	1.0119
-6	1.1064	1.0638	1.0113	1.1152	1.182	1.077	0.8175	1.0111	1.2057	1.0123	1.0015	1.009	0.9246	0.9115	0.9532	0.9685	0.9128	0.9171	0.9644	1.0474
-5	1.1652	1.0676	1.1049	1.1218	1.2414	1.0695	0.7344	1.07	1.1121	0.9743	0.9944	0.9532	0.9317	0.9597	0.9115	0.986	0.9222	0.9903	0.9765	0.9712
-4	1.0866	1.0862	1.019	1.0842	1.1177	1.0361	0.7615	1.1102	1.1378	0.9515	1.0108	0.998	0.9705	0.9397	0.8903	0.9617	0.9427	0.9653	0.9619	1.0163
-3	1.1194	0.9911	0.9852	1.1079	1.1369	1.0146	0.6997	1.1298	1.0057	0.987	1.1274	1.061	1.0057	0.9675	0.9701	0.9772	0.9885	0.9526	0.9272	0.9213
-2	1.1612	1.08	1.073	1.1323	1.1392	1.0276	0.7676	1.0621	1.0314	0.9393	1.1926	1.0475	0.9795	0.9775	0.9103	1.1325	0.9957	0.9124	0.9877	0.9007
-1	1.0188	1.1547	1.1884	1.0248	0.7765	1.0822	0.8506	0.8438	1.2185	0.8268	0.7622	1.2215	0.9915	0.9115	0.9302	1.3194	1.2316	0.7095	1.2005	0.9156
1	0.9764	0.9825	0.9432	1.0274	1.0716	1.0988	0.9919	0.8443	0.8405	0.9476	1.1134	0.9846	1.0053	0.9587	1.0682	1.0347	0.9837	1.2852	1.1773	0.8534
2	1.1211	1.0444	1.0009	1.1099	1.1128	1.0538	0.9457	1.0253	0.9745	0.9951	1.106	1.0602	1.0109	0.9063	0.949	0.9514	0.976	0.9512	0.9966	0.8503
3	1.0855	0.9981	1.0714	1.1337	1.1721	1.1982	0.8087	1.1092	0.8772	0.9937	0.6856	1.0788	0.9915	0.9483	1.0155	0.9691	0.9213	0.9387	0.9213	0.9804
4	1.0663	0.9796	1.0202	1.0373	1.2084	1.0881	0.7377	1.1083	0.9414	1.047	0.9416	0.9886	0.9358	0.9839	1.0094	0.9063	0.8786	1.0157	0.9714	1.0465
5	1.0634	1.0707	1.0682	1.0756	1.1787	1.1243	0.8321	1.0744	1.0167	1.032	0.927	1.0263	0.9368	0.9834	0.9562	0.9859	0.876	0.962	0.9408	1.0255
6	1.129	1.107	1.0594	1.1198	1.215	1.108	0.8934	0.9651	1.0112	1.0367	0.9233	1.0378	0.9358	0.8712	0.9465	0.9883	0.9171	0.9034	0.9462	0.9957
7	1.0589	1.0506	1.0779	1.0492	1.1491	1.093	0.8486	1.0062	1.0809	1.0406	0.9592	0.9768	0.9309	0.9912	0.981	0.9356	0.894	0.9571	0.985	1.0413
8	1.0493	0.9684	1.0057	1.0908	1.1309	1.0374	0.8078	0.9576	1.0883	1.0531	0.9774	1.0098	0.9156	0.9225	1.0052	0.969	0.8672	1.0091	1.0427	1.0606
9	1.0193	1.0395	1.0363	1.1026	1.0765	1.0035	0.625	0.965	1.1341	1.0451	1.0417	1.0077	0.9541	0.9045	0.9574	0.9828	0.882	0.9781	0.9924	1.0395
10	1.0578	1.0705	0.9926	1.0677	1.1342	1.0521	0.8243	0.9984	1.1011	1.0689	1.0031	0.9306	0.9287	0.9479	0.9544	0.9525	0.9716	0.9833	0.9199	1.0343

Table B.1A. Amino acid propensity near proline residue in *trans* case. Propensities >1 are shown in red, otherwise values are shown in blue (see text for additional details).

	F	L	I	Y	W	V	M	P	C	A	G	T	S	K	Q	N	H	E	D	R
-10	1.1647	0.9826	0.9668	1.1671	1.1563	0.9569	0.6023	1.0425	1.0979	1.0195	1.0644	1.0918	0.9098	1.0573	0.9477	0.9312	0.9816	1.02	1.0796	0.956
-9	0.9564	1.1244	1.1896	1.1299	1.0203	1.0715	0.9035	1.0728	0.9843	0.9737	0.9441	0.9521	0.9635	0.9685	0.9355	0.9866	0.9816	1.0267	0.9676	1.0393
-8	1.0497	0.9978	0.7797	1.2388	1.2584	1.1052	0.6031	1.0627	1.4387	1.0425	1.0525	1.1704	0.9689	0.9555	0.7482	1.0085	1.005	0.9254	0.9916	0.9128
-7	1.0497	1.0029	1.1646	1.1979	1.2584	1.186	0.7629	1.0728	1.5144	0.9795	0.9441	1.0045	0.9637	1.138	1.035	0.7426	0.9816	0.7836	0.9516	0.9314
-6	1.2596	1.0788	0.9567	1.4702	1.5305	1.3141	0.7629	0.9514	1.3251	0.8764	1.0663	0.9346	0.8327	0.9685	0.8729	1.1194	0.8616	0.743	0.9636	0.723
-5	1.0147	1.0079	0.9816	1.0482	1.7005	1.2399	0.9436	1.3562	1.3629	0.9394	1.0716	1.1879	0.9483	0.7586	0.8729	1.0418	0.8816	0.7363	0.7997	0.9037
-4	1.2363	0.9523	1.0399	1.266	1.5845	1.186	0.7629	1.3967	1.1736	0.9534	1.0334	1.0569	1.1025	0.7909	0.7482	1.0861	1.0756	0.7565	0.6478	1.0122
-3	1.2013	1.0333	1.173	1.0346	1.1904	1.1119	0.8634	1.3866	1.2115	0.9222	1.0525	1.1093	1.0562	0.7506	0.8853	1.0529	1.1108	0.6173	0.6233	0.8314
-2	1.2363	0.861	0.782	1.1027	1.0883	1.0445	0.9035	1.4777	1.4387	0.7332	1.1673	1.1268	1.0871	0.8797	1.0225	1.1637	1.1108	0.7633	0.9037	0.967
-1	1.4463	0.7344	0.5574	2.4095	2.1766	0.5324	0.9622	1.1943	0.9465	0.9222	1.5501	0.8622	1.0562	0.8716	1.035	1.2866	0.8616	0.9187	0.7117	0.8043
1	1.0078	0.9573	0.9905	2.0147	1.1223	1.0984	0.5019	1.174	1.0601	1.094	0.9623	0.9084	1.0331	0.8053	0.9352	1.2302	1.4811	0.5404	0.7357	0.8314
2	1.1663	0.8205	1.0315	1.2796	1.1563	0.9906	0.7428	2.1254	1.0979	0.6472	1.1099	1.2753	0.902	0.7264	0.5361	1.2967	1.0579	0.699	1.1116	0.8957
3	0.9331	1.0282	1.0648	1.1979	1.1223	1.3006	0.8232	1.6789	1.0601	0.7446	0.6675	1.153	0.9175	0.8151	0.7482	0.942	1.1108	0.8649	0.9516	1.0303
4	1.283	0.6256	1.2728	1.3613	1.3604	1.1793	0.6023	1.255	0.795	0.8993	0.9823	1.1268	0.8712	0.7829	0.8853	1.1526	1.1813	0.6376	0.9586	0.7953
5	1.0964	0.8472	0.9151	1.2116	1.3944	1.1591	0.8232	1.255	1.0979	0.9305	1.0079	1.0307	1.1265	0.7829	0.9851	0.9642	1.0579	0.7701	1.3275	0.7501
6	1.1663	0.9421	1.2146	0.9933	1.6325	1.0243	0.9634	1.174	1.2115	0.9451	1.0972	1.0132	0.9175	0.7748	0.7856	0.942	0.9345	0.9187	1.1596	0.6405
7	1.1663	0.9472	0.9698	1.0482	1.4824	1.1456	0.8623	0.9608	1.0601	0.8878	1.0716	1.118	0.9252	0.8878	1.0973	0.8645	1.0227	1.1888	0.8397	0.976
8	1.108	0.856	0.9151	0.9121	1.1223	1.1523	0.783	1.1133	1.0222	1.0539	1.078	1.2229	0.9461	0.9636	0.9976	0.9642	1.199	1.0536	0.9516	0.9947
9	1.1313	0.932	0.9234	1.1571	1.2244	1.1119	0.8432	0.6704	1.7415	1.1169	1.1992	0.7599	0.8013	0.8878	1.0474	0.931	0.6111	0.9727	1.0716	1.0122
10	1.0497	0.932	1.1397	1.1299	1.4964	1.0176	0.9235	0.8293	0.9843	1.0711	0.976	1.153	0.7364	0.8958	0.8729	1.1415	0.9169	0.9795	0.9676	1.0935

Table B.1B. Amino acid propensity near proline residue in *cis* case. The color scheme is same with the Table B.1A.

	F	L	I	Y	W	V	M	P	C	A	G	T	S	K	Q	N	H	E	D	R
-10	1.2343	0.9529	1.2716	0.8003	0.3999	1.347	0.7082	0.595	0.8903	0.7408	1.1251	1.2325	1.2692	0.7592	0.7331	1.1728	0.922	0.7148	1.2224	1.2752
-9	1.0971	1.3698	1.2716	0.9604	1.1997	1.1865	0.2361	0.952	0.4452	0.7406	0.525	1.4379	0.272	1.0439	0.8797	0.9122	1.8659	1.0325	1.5045	0.7438
-8	0.6228	1.072	1.2716	1.2805	2.3994	1.347	1.1803	0.595	1.7806	1.0776	1.0501	1.1297	0.7252	1.1388	0.6797	1.0425	0.2073	0.7942	0.7522	0.8376
-7	0.6357	1.1315	1.1738	0.9604	0.7998	0.9508	0.4721	1.3091	0.4452	1.0102	0.6	1.746	1.2692	1.2337	1.3196	1.0425	0.4146	0.7942	1.1204	0.4251
-6	1.3714	1.0124	1.076	1.2805	0.7998	1.347	0.7082	1.071	1.3355	0.7408	1.2001	0.7189	0.6346	1.5184	0.5665	0.6516	0.8293	1.3502	0.5642	0.9564
-5	1.7828	0.7742	1.1738	0.9604	1.5996	1.347	0.4721	1.071	3.5613	0.6735	0.9	0.9243	1.3598	1.0439	0.8797	1.0425	0.622	0.7942	0.7522	0.8376
-4	1.6457	0.7147	0.5369	0.8003	0.7996	0.7924	0.7082	1.5471	2.6709	0.3755	1.0501	1.3352	1.4505	1.1388	0.5365	0.7319	0.3293	0.7942	0.9463	1.2752
-3	1.0971	1.072	0.7625	1.4406	0.7993	0.8716	0.4721	1.6661	2.6709	0.7408	1.2001	0.7189	1.3598	1.0439	1.0263	0.9122	0.8293	0.9531	0.2621	1.1689
-2	0.6557	0.8238	1.1738	1.1205	1.1997	0.8716	0.9443	1.7851	0.8905	0.202	1.2751	1.027	1.2692	1.4235	0.8797	0.6516	0.4146	1.0325	1.4105	0.9564
-1	0.6557	0.2976	0.5669	1.1205	1.5996	0.2377	0.2361	1.071	0	1.1449	2.2501	0.9243	1.1785	1.3286	0.5797	1.8244	1.0366	1.3502	0.7522	0.9564
1	2.3314	1.1315	0.5669	2.0809	0.3999	1.1885	0.2361	0.952	0.4452	0.8082	1.0501	1.1297	0.9972	0.5694	0.8797	0.7619	1.8659	0.9531	0.9403	0.4251
2	1.0971	0.7742	0.6847	1.6007	0.7996	1.0301	0.9443	1.6661	0.8903	0.5388	1.1251	1.027	0.7252	1.3286	1.1729	0.9122	0.3293	1.2708	0.9403	0.9564
3	0.6557	1.1315	1.1738	0.8003	0.3999	0.6339	0.9443	1.071	2.2258	0.4041	1.0501	1.3352	0.9972	1.1388	0.7331	0.3909	1.8659	1.2708	1.5045	0.9564
4	1.5085	0.9529	1.1738	0.4802	0.7998	0.8716	0.9443	1.19	0.4452	1.1449	0.9	1.3352	0.9065	0.6541	1.0263	1.1728	1.0366	0.6737	0.7522	1.1689
5	0.6357	0.3573	1.5651	0.8003	0.3999	1.1885	0.9443	1.3091	1.3355	0.6061	0.75	1.2325	1.9944	1.0439	1.3196	0.6516	0.4146	1.0325	1.3164	0.9564
6	1.6457	1.2507	1.6629	0.9604	0.7998	1.1093	1.4164	0.238	0.6903	0.6082	0.825	0.7189	0.7252	1.2337	1.1729	0.9122	0.622	0.7942	1.3164	0.7438
7	1.3714	1.1911	0.9782	1.6007	1.9995	1.347	0.9443	0.714	1.3355	0.6082	0.525	1.8487	1.0878	0.1898	0.6797	1.0425	0.6293	0.7942	0.9842	1.2752
8	1.3714	1.1315	1.076	0.9604	0	1.0301	0.9443	0.952	0.8903	1.2123	0.825	1.2325	0.9972	1.1388	1.0263	0.9122	0.622	0.3737	0.3465	0.8501
9	1.0971	1.1315	0.8003	1.7607	1.1997	0.9547	0.9443	0.952	1.3355	1.2796	1.0501	1.027	1.1785	1.0439	0.5665	0.9122	1.0366	0.556	1.2224	0.7438
10	1.2343	1.1315	0.7625	1.4406	0.3999	0.8716	1.4164	0.833	0.8903	1.0102	1.0501	1.027	1.1785	0.2847	0.5665	1.6941	0.4146	1.2708	0.9403	1.0626

Table B.1C. Amino acid propensity near proline residue in *isom* case. The color scheme is same with the Table B.1A.

To facilitate comparison the propensities in *cis*, *trans* and *isom* cases, propensity *differences* were calculated directly by subtracting the propensities in the *isom* cases from those of either the *trans* or *cis* cases (Table B.2). Values in blue indicate that the propensity difference between two cases is <1 at a position, which means the residue at the position is disfavored in the *isom* cases. Conversely, values in red are used to indicate residues favored in the *isom* cases. The favored and disfavored residues common to both comparisons (*trans* vs *isom* and *cis* vs *isom*) were obtained by comparing values in Table B.2A and B.2B, using boxes to highlight common values. Favored residues (for *isom*) identified in both comparisons are: i) cysteine residues at positions -5, -4, -3 and 3; and ii) tryptophan residue at position -8. There is only one common disfavored residue (for *isom)*, tryptophan residue at position 8. Interestingly, a previous study on proline isomerization using short synthetic peptides found that the rate of the isomerization was enhanced in the disulfide bridge containing peptides, Ac-Cys-Pro-Xaa-Cys-NH2 or Ac-Cys-Gly-Pro-Cys-NH2 (22). Also, in studies on oxidative folding pathways of cysteine-rich peptides derived from minicollagen-1, two different proline isomers were detected, depending on the sequence environment (6). Our analysis reported here, however, provides the first evidence based on a comprehensive analysis of prolyl-peptide bonds, that cysteine is favored near the isomerizing proline residues in native protein structures.

	F	L	I	Y	W	V	M	P	C	A	G	T	S	K	Q	N	H	E	D	R
-10	0.208	-0.079	0.214	-0.246	-0.746	0.288	-0.142	-0.398	-0.192	-0.295	0.087	0.232	0.354	-0.15	-0.268	0.213	-0.251	-0.237	0.245	0.233
-9	0.016	0.292	0.252	-0.128	-0.009	0.102	-0.616	-0.028	-0.625	-0.26	-0.503	0.434	-0.635	0.137	-0.087	-0.068	1.004	0.112	0.525	-0.278
-8	-0.258	-0.018	0.179	0.131	1.199	0.26	0.39	-0.384	0.711	0.077	0.033	0.144	-0.213	0.245	-0.08	0.08	-0.664	-0.103	-0.183	-0.385
-7	-0.409	0.045	0.121	-0.107	-0.349	-0.154	-0.334	0.292	-0.564	0.025	-0.404	0.726	0.378	0.292	0.394	0.087	-0.426	-0.163	0.157	-0.587
-6	0.265	-0.051	0.065	0.165	-0.382	0.27	-0.109	0.06	0.13	-0.271	0.199	-0.29	-0.29	0.607	-0.387	-0.257	-0.084	0.433	-0.4	-0.091
-5	0.618	-0.313	0.069	-0.161	0.358	0.278	-0.312	0.001	2.449	-0.301	-0.094	-0.039	0.428	0.174	-0.032	0.055	-0.3	-0.096	-0.224	-0.334
-4	0.559	-0.371	-0.432	-0.284	-0.318	-0.244	-0.073	0.437	1.533	-0.076	0.039	0.337	0.477	0.199	-0.304	-0.21	-0.113	-0.111	-0.136	0.259
-3	-0.022	0.081	-0.184	0.333	-0.336	-0.143	-0.228	0.536	1.665	-0.226	0.073	-0.342	0.354	0.076	0.056	-0.065	-0.079	3E-04	-0.645	0.248
-2	-0.476	-0.246	0.101	-0.012	0.061	-0.156	0.176	0.723	-0.141	-0.637	0.082	-0.02	0.29	0.546	-0.031	-0.481	-0.481	0.22	0.423	0.056
-1	-0.333	-0.857	-0.601	0.096	0.823	-0.644	-0.615	0.227	-1.219	0.318	1.488	-0.297	0.187	0.417	-0.051	0.505	-0.195	0.641	-0.448	0.041
1	1.355	0.148	-0.256	1.053	-0.672	0.09	-0.456	0.108	-0.395	-0.139	-0.083	0.145	-0.008	-0.289	-0.19	-0.253	0.982	-0.332	-0.237	-0.428
2	-0.024	-0.27	-0.324	0.491	-0.313	-0.024	0.099	0.641	-0.084	-0.456	0.019	-0.033	-0.286	0.522	0.224	-0.039	-0.147	0.32	-0.058	0.108
3	-0.4	0.135	0.102	-0.333	-0.772	-0.564	0.136	-0.038	1.349	-0.59	0.164	0.256	0.006	0.19	-0.282	-0.498	1.045	0.332	0.583	-0.024
4	0.442	-0.027	0.154	-0.557	-0.409	-0.216	0.207	0.082	-0.496	0.098	-0.041	0.336	-0.029	-0.13	0.017	0.267	0.158	-0.142	-0.219	0.122
5	-0.378	-0.713	0.497	-0.275	-0.779	0.064	0.112	0.235	0.319	-0.426	-0.177	0.206	1.056	0.161	0.383	-0.234	-0.461	0.071	0.376	-0.069
6	0.517	0.144	0.593	-0.159	-0.415	0.001	0.523	-0.747	-0.121	-0.229	-0.098	-0.319	-0.211	0.362	0.226	0.013	-0.295	-0.109	0.37	-0.252
7	0.312	0.141	-0.1	0.551	0.85	0.254	0.096	-0.292	0.255	-0.232	-0.434	0.872	0.157	-0.701	-0.101	0.107	-0.065	-0.163	-0.421	0.234
8	0.322	0.143	0.07	-0.13	-1.131	-0.007	0.137	-0.006	-0.198	0.159	-0.152	0.223	0.082	0.216	0.021	-0.077	-0.285	-0.135	-0.196	-0.21
9	0.078	0.092	-0.156	0.658	0.123	-0.449	0.109	-0.013	0.201	0.235	0.008	0.019	0.224	0.139	-0.371	-0.052	0.155	-0.42	0.23	-0.296
10	0.176	0.061	-0.21	0.373	-0.734	-0.181	0.592	-0.165	-0.211	-0.059	0.047	0.038	0.25	-0.663	-0.368	0.742	-0.457	0.288	0.02	0.028

Table B.2A. Amino acid propensity difference between *isom* and *trans* cases

	F	L	I	Y	W	V	M	P	C	A	G	T	S	K	Q	N	H	E	D	R
-10	0.08	-0.03	0.365	-0.357	-0.756	0.39	0.106	-0.447	-0.208	-0.279	0.041	0.141	0.359	-0.298	-0.215	0.342	-0.26	-0.305	0.143	0.317
-9	0.141	0.245	0.082	-0.169	0.179	0.117	-0.687	-0.121	-0.539	-0.233	-0.419	0.486	-0.592	0.075	0.044	0.026	0.984	0.006	0.537	-0.295
-8	-0.227	0.074	0.498	0.042	1.141	0.242	0.377	-0.468	0.342	0.035	-0.002	-0.041	-0.262	0.283	0.132	0.034	-0.798	-0.131	-0.239	-0.275
-7	-0.384	0.129	0.009	-0.238	-0.459	-0.235	-0.291	0.236	-1.069	0.031	-0.344	0.741	0.305	0.098	0.285	0.3	-0.467	0.011	0.177	-0.406
-6	0.112	-0.066	0.119	-0.19	-0.731	0.033	-0.055	0.12	0.01	-0.136	0.135	-0.216	-0.198	0.55	-0.286	-0.468	-0.052	0.607	-0.419	0.233
-5	0.768	-0.234	0.192	-0.088	-0.101	0.107	-0.471	-0.285	2.198	-0.286	-0.172	-0.264	0.411	0.285	0.007	7E-04	-0.26	0.058	-0.047	-0.266
-4	0.409	-0.248	-0.453	-0.466	-0.765	-0.394	-0.055	0.15	1.497	0.022	0.017	0.278	0.348	0.348	-0.162	-0.304	-0.246	0.038	0.199	0.263
-3	-0.104	0.039	-0.39	0.406	-0.391	-0.24	-0.411	0.279	1.459	-0.181	0.148	-0.39	0.304	0.293	0.141	-0.141	-0.282	0.136	-0.342	0.337
-2	-0.551	-0.027	0.392	0.018	0.111	-0.173	0.041	0.307	-0.548	-0.531	0.108	-0.1	0.182	0.544	-0.143	-0.512	-0.696	0.269	0.507	-0.011
-1	-0.761	-0.437	0.03	-1.289	-0.577	-0.295	-0.326	-0.123	-0.946	0.223	0.7	0.042	0.122	0.457	-0.155	0.539	0.155	0.432	0.041	0.152
1	0.524	0.134	-0.104	0.066	-0.722	0.09	-0.266	-0.222	-0.615	-0.286	0.068	0.221	-0.036	-0.036	-0.055	-0.448	0.385	0.413	0.205	-0.406
2	-0.069	-0.046	-0.347	0.321	-0.357	0.039	0.201	-0.459	-0.208	-0.108	0.015	-0.248	-0.177	0.602	0.587	-0.384	-0.229	0.582	-0.171	0.071
3	-0.247	0.103	0.109	-0.398	-0.722	-0.667	0.121	-0.508	1.166	-0.341	0.183	0.182	0.08	0.324	-0.015	-0.551	0.755	0.386	0.553	-0.074
4	0.226	0.127	-0.099	-0.881	-0.561	-0.308	0.342	-0.065	-0.35	0.246	-0.082	0.208	0.035	0.071	0.141	0.02	-0.145	0.036	-0.207	0.374
5	-0.411	-0.59	0.65	-0.411	-0.995	0.029	0.121	0.054	0.238	-0.224	-0.258	0.202	0.869	0.261	0.334	-0.313	-0.643	0.262	-0.011	0.206
6	0.479	0.309	0.448	-0.033	-0.833	0.085	0.533	-0.936	-0.321	-0.137	-0.272	-0.294	-0.192	0.459	0.387	-0.03	-0.313	-0.124	0.157	-0.097
7	0.205	0.244	0.071	0.552	0.537	0.201	0.081	-0.187	0.275	-0.08	-0.547	0.731	0.163	-0.698	-0.218	0.178	-0.193	-0.395	-0.275	0.299
8	0.263	0.276	0.161	0.048	-1.122	-0.122	0.181	-0.161	-0.132	0.158	-0.253	0.01	0.149	0.275	0.029	-0.052	-0.577	-0.18	-0.105	-0.045
9	-0.034	0.2	-0.043	0.604	-0.025	-0.557	0.101	0.082	-0.406	0.163	-0.149	0.267	0.377	0.156	-0.461	-0.019	0.226	-0.417	0.151	-0.268
10	0.185	0.2	-0.357	0.311	-1.097	-0.146	0.493	0.003	-0.094	-0.061	0.074	-0.126	0.392	-0.611	-0.286	0.553	-0.502	0.291	-0.027	-0.031

Table B.2B. Amino acid propensity difference between *isom* and *cis* cases

Table B.3 summarizes secondary structure tendencies, calculated for each case and represented as described above for the amino acid propensity comparisons.

	Helix	Sheet	Coil
-10	0.95486	1.08426	0.99436
-9	0.93032	1.13045	0.99106
-8	0.90113	1.16609	0.99847
-7	0.87212	1.18642	1.01469
-6	0.83003	1.21377	1.03948
-5	0.79123	1.18998	1.09108
-4	0.70395	1.16964	1.18768
-3	0.51399	1.0958	1.41529
-2	0.382	0.98989	1.60549
-1	0.26297	0.62455	1.93532
0	0.60056	0.50096	1.68032
1	0.77368	0.65752	1.4205
2	0.872	0.85258	1.21067
3	0.89845	1.00765	1.09403
4	0.92155	1.14121	0.99326
5	0.95856	1.16224	0.94501
6	0.98966	1.13174	0.93274
7	1.01832	1.0886	0.93025
8	1.0342	1.04763	0.93888
9	1.03485	1.00469	0.96343
10	1.03999	0.95909	0.98521

trans

	Helix	Sheet	Coil
-10	0.76972	1.26258	1.06935
-9	0.73647	1.30331	1.07771
-8	0.71799	1.41328	1.03111
-7	0.63179	1.54565	1.03709
-6	0.56775	1.61081	1.06098
-5	0.50986	1.60674	1.11953
-4	0.4409	1.45808	1.27366
-3	0.30542	1.28091	1.50903
-2	0.16133	1.00396	1.81132
-1	0.03818	0.63536	2.14706
0	0.03695	0.17717	2.41708
1	0.21798	0.88788	1.82446
2	0.33252	1.24629	1.50306
3	0.56159	1.40513	1.18763
4	0.63918	1.49474	1.05979
5	0.73277	1.41531	1.01558
6	0.82637	1.28091	1.00363
7	0.86701	1.29517	0.95584
8	0.91504	1.25647	0.93194
9	0.93475	1.25036	0.91641
10	0.96184	1.13429	0.95823

cis

	Helix	Sheet	Coil
-10	0.70956	1.43669	1.02556
-9	0.72404	1.58035	0.92722
-8	0.75301	1.67613	0.84293
-7	0.66612	1.50852	1.02556
-6	0.57924	1.62824	1.03961
-5	0.53579	1.36485	1.23629
-4	0.46339	1.05357	1.48917
-3	0.36202	0.83807	1.71395
-2	0.13033	0.55073	2.10731
-1	0.08689	0.43101	2.2197
0	0.14481	0.19156	2.304
1	0.33306	0.6944	1.82634
2	0.49235	1.17329	1.39083
3	0.63716	1.24513	1.20819
4	0.72404	1.29302	1.0958
5	0.88333	1.41274	0.87102
6	1.01366	1.34091	0.78673
7	0.9847	1.36485	0.80078
8	1.04262	1.17329	0.85597
9	0.99918	1.02962	0.98341
10	1.0571	0.9099	0.99746

isom

Table B.3. Secondary structure propensities near proline residues.

The distribution of secondary structure tendencies for residues near proline in the three cases is distinguishable, suggesting that secondary structure information could be useful in developing classifiers to accurately identify isomerizing cases.

Classifying Prolyl Peptide Bonds as Cis, Trans or Isomerizing

As described in the Materials and Methods section, we compared 5 different combinations of 4 different input features (seq, seq+ss, seq+hssp_ent, seq+hssp_pro and seq+ss+hssp_pro) and 2 different machine learning algorithms (Naïve Bayes and SVM, from the WEKA software package (31)). For building classifiers to distinguish the three different prolyl peptide ond classes (*trans*, *cis* and *isom*), three binary classifiers; *cis/trans*, *isom/cis*

and *isom/trans* were constructed and evaluated for classification performance. Also 5-fold cross-validation was used for all evaluations.

In Figure B.3, a comparison of several different classifiers, using different feature combinations for the *cis/trans* classification task, is presented using Receiver Operating Characteristics (ROC) curves. Except for the classifier that used only sequence information, the classifiers evaluated in the ROC curve were constructed using the SVM method, which gave us better performance than Naïve Bayes on this task. Based on the area under curve (AUC) in the ROC, the classifier that uses sequence, secondary structure and hssp profile information as input, has the best performance. Interestingly, the classifier that uses secondary structure information outperformed classifiers using either HSSP entropy or HSSP profile information. This suggests that secondary structure is an important data feature for the *cis/trans* classification problem.

For overall comparisons among the three classification tasks (*cis/trans*, *isom/cis* and *isom/trans)*, the classifier that showed the best performance for each classification task was chosen and presented in the ROC curve in shown in Figure B.4. As expected, we obtained the best performance in the *cis/trans* classification task. Unfortunately, we did not obtain comparable performance in the *isom/trans* and *isom/cis* classification tasks, although the *isom/trans* classification performance was better than that of the *isom/cis* classification. In Table B.4, the classifier performance is summarized using measures described in Materials Methods. It is important to note that in Table B.4, the classifier for each classification task was optimized based on the *correlation coefficient* (not, for example, based on specificity for the positive class). Therefore, a meaningful comparison of the overall performance of classifiers on each classification task has to be carefully considered, using both ROC curves in Figure B.2 and the measures in Table B.4.

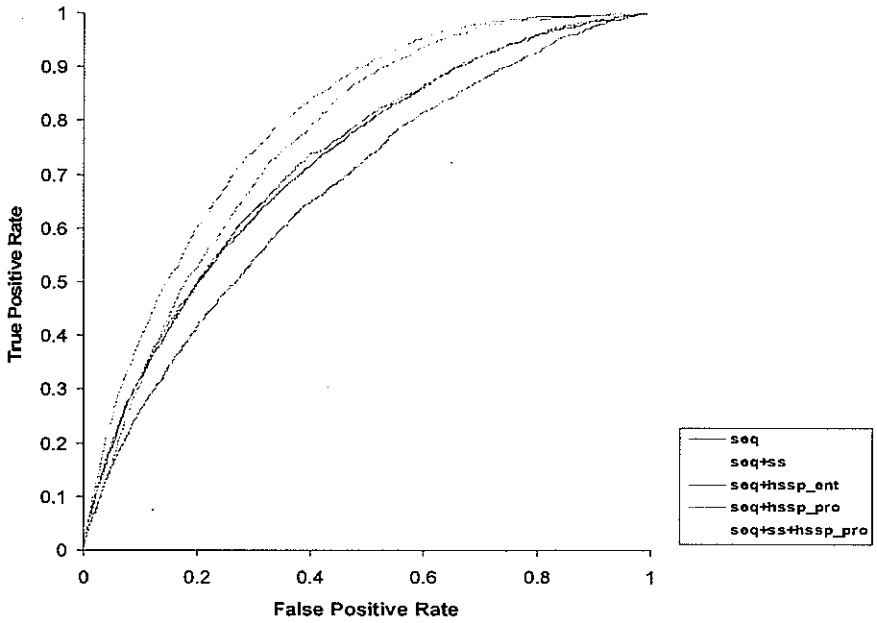


Figure B.3. Receiver Operating Characteristics (ROC) curves for classifiers using 5 different combinations of 4 different data features. Combinations of data features used are shown in the box.

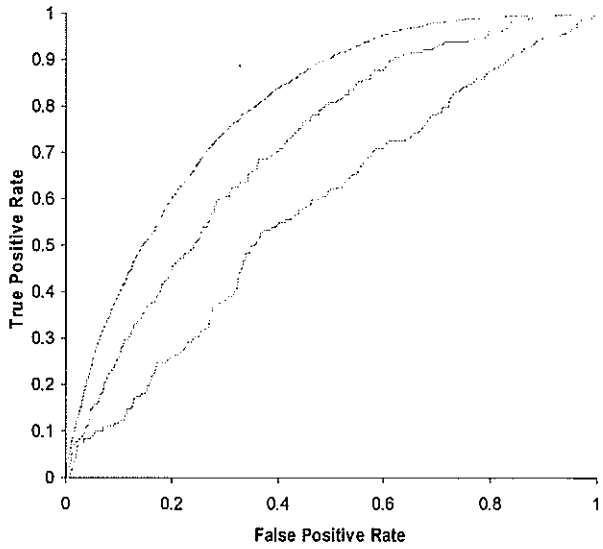


Figure B.4. ROC curves for the best classifiers for each of three different classification tasks. Blue line, *cis/trans* classification; green line, *isom/trans* classification; pink line, *isom/cis* classification task.

classification task	data features	methods	window size	ACC	CC	AUC	SP	SN
cis/trans	seq+ss+hssp_pro	SVM	**11**	**0.80**	**0.21**	**0.79**	**0.13**	**0.59**
isom/cis	seq+hssp_ent	Naïve Bayes	**11**	**0.92**	**0.10**	**0.54**	**0.57**	**0.02**
isom/trans	seq+hssp_ent	Naïve Bayes	**21**	**1.00**	**0.05**	**0.55**	**0.50**	**0.01**

Table B.4. Overall performance measures (based on optimization of correlation coefficient). ACC: accuracy, CC: correlation coefficient, AUC: area under ROC curve, SP: specificity for positive class (*cis* or *isom)*, SN: sensitivity for positive class (*cis* or *isom)*. Definitions of performance measures are provided in the Materials and Methods.

FUTURE STUDIES

Two important preliminary studies on the characterization of the proline *cis/trans* isomerization are presented in this report: i) Analysis of the amino acid and secondary structure propensities of residues near proline residues in each of three different configuration classes (*cis*, *trans* and isomerizing), and ii) Application of machine learning algorithms to attempt to classify of prolyl peptide bonds into one of three classes (*cis*, *trans* and *isom*) was attempted for the first time. The results have not been fully discussed and presented in this Appendix (which is a very preliminary draft of a future manuscript), but I would like to propose several suggestions for the future study based on the current stage of this analysis. Further discussion of the current results and additional computational experiments suggested below would solidify and enhance the quality of this study.

Systematic analysis of data on amino acid and secondary structure propensities near proline residues

We have already analyzed the amino acid sequence and secondary structure propensities for the different three classes of prolyl peptide bond configurations. The results in this study show significant differences among *cis*, *trans* and *isom* cases near the proline residues, in terms of amino acid distribution and secondary structure tendencies. Previous studies have shown that there is a statistically significant difference between the *cis* and *trans* cases, based on neighboring sequence information (9, 13, 17) and one study showed a structural differences in a protein with has two prolyl bond configurations (18). I propose a systematic characterization and comparison of three classes presented in study. Analysis can be based on both sequence information and structural information near the proline residues. Detailed analysis of the available data for the three different configurations of prolyl peptide bonds is important not only for understanding the nature of the proline isomerization but also for use of the distinguishing data features in classification tasks. The results from this analysis would give us opportunities to construct sequence and/or structural motifs defining each proline configuration and to identify additional "proline switch" candidates.

More sophisticated machine learning methods for classification tasks

The ultimate goal in this study is to find more candidates for the "proline switch" in order to investigate its possible role as a cellular mechanism for post-translation regulation of protein function. To achieve this goal, we tried to construct and evaluate classifiers for predicting the *cis*, *trans* and *isom* classes. Although the method used in this study did not achieve good classification performance in isom/*cis* and *isom*/*trans* classification tasks, we can differentiate and classify the *cis* and *trans* classes as has been shown in previous studies

described above. Difficulties in classifying *isom/trans* and *isom/cis* may be inherent to this biological phenomenon and reflect the low thermodynamic barrier that separates the *isomerizing* configuration from more "stably" *cis* and *trans* configurations in a living biological system. Still, the preliminary analyses suggest there is room to achieve better performance in identifying *isomerizing* prolines. I propose two potential strategies. First, data features correlated with *isomerizing* proline residues should be explored extensively. The above section already mentioned the data analysis, but feature extraction is a very important step in building machine learning classifiers. Therefore, if more distinctive features in the isomerization cases, such as structural features of the *isomerizing* proline or correlated amino acids near the *isomerizing* proline residues, could be used in the classification task, better performance might be achieved. A second opportunity for improving performance would be to use machine learning algorithms designed especially for multi-class classification problems and unbalanced datasets. In this study, only binary classifiers were used as a first attempt on these classification problems. We know that the isomerizing cases could have both *trans* and *cis* proline configurations. Therefore, algorithms that explicitly take into account the intrinsic features of isomerizing cases would be the best way to achieve better performance. Algorithms that are able to handle extremely unbalanced datasets would also be helpful for achieving our goal of finding more candidates for the "proline switch".

Structural analysis for isomerizing cases

Preliminarily, we found total 460 cases in which the same prolyl peptide bond can have two different configurations in a folded protein. Of these 460 cases, we used only a non-redundant subset of 184 cases for the classification datasets. If the "proline switch" is the general mechanism for regulating cellular function, it may be possible to find common structural features or motifs in different proteins. The 184 cases would be a good starting

point to investigate characteristic structural features of the isomerizing cases. The analyses could be performed using structural alignment methods or more extensive computational approaches such as molecular dynamics.

REFERENCES

1. **Andreotti, A. H.** 2003. Native state proline isomerization: an intrinsic molecular switch. Biochemistry **42:**9515-24.

2. **Andreotti, A. H.** 2006. Opening the pore hinges on proline. Nat Chem Biol **2:**13-4.

3. **Baldi, P., S. Brunak, Y. Chauvin, C. A. Andersen, and H. Nielsen.** 2000. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics **16:**412-24.

4. **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.

5. **Black, S. D., and D. R. Mould.** 1991. Development of hydrophobicity parameters to analyze proteins which bear post- or cotranslational modifications. Anal Biochem **193:**72-82.

6. **Boulegue, C., A. G. Milbradt, C. Renner, and L. Moroder.** 2006. Single proline residues can dictate the oxidative folding pathways of cysteine-rich peptides. J Mol Biol **358:**846-56.

7. **Cai, C. Z., W. L. Wang, L. Z. Sun, and Y. Z. Chen.** 2003. Protein function classification via support vector machine approach. Math Biosci **185:**111-22.

8. **Fisher, G.** 2000. Chemical aspects of peptide bond isomerisation. Chemical Society Reveiws **29:**119-27.

9. **Frommel, C., and R. Preissner.** 1990. Prediction of prolyl residues in cis-conformation in protein structures on the basis of the amino acid sequence. FEBS Lett **277:**159-63.

10. **Guo, J., H. Chen, Z. Sun, and Y. Lin.** 2004. A novel method for protein secondary structure prediction using dual-layer SVM and profiles. Proteins **54:**738-43.

11. **Hua, S., and Z. Sun.** 2001. Support vector machine approach for protein subcellular localization prediction. Bioinformatics **17:**721-8.

12. **Kabsch, W., and C. Sander.** 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers **22:**2577-637.

13. **Lorenzen, S., B. Peters, A. Goede, R. Preissner, and C. Frommel.** 2005. Conservation of cis prolyl bonds in proteins during evolution. Proteins **58:**589-95.

14. **Mallis, R. J., K. N. Brazin, D. B. Fulton, and A. H. Andreotti.** 2002. Structural characterization of a proline-driven conformational switch within the Itk SH2 domain. Nat Struct Biol **9:**900-5.

15. **Mitchell, T.** 1997. Machine Learning. McGRAW-HILL.

16. **Nelson, C. J., H. Santos-Rosa, and T. Kouzarides.** 2006. Proline isomerization of histone H3 regulates lysine methylation and gene expression. Cell **126:**905-16.

17. **Pahlke, D., C. Freund, D. Leitner, and D. Labudde.** 2005. Statistically significant dependence of the Xaa-Pro peptide bond conformation on secondary structure and amino acid sequence. BMC Struct Biol **5:**8.

18. **Reimer, U., and G. Fischer.** 2002. Local structural changes caused by peptidyl-prolyl cis/trans isomerization in the native state of proteins. Biophys Chem **96:**203-12.

19. **Reimer, U., G. Scherer, M. Drewello, S. Kruber, M. Schutkowski, and G. Fischer.** 1998. Side-chain effects on peptidyl-prolyl cis/trans isomerisation. J Mol Biol **279:**449-60.

20. **Sarkar, P., C. Reichman, T. Saleh, R. B. Birge, and C. G. Kalodimos.** 2007. Proline cis-trans isomerization controls autoinhibition of a signaling protein. Mol Cell **25:**413-26.

21. **Schneider, R., A. de Daruvar, and C. Sander.** 1997. The HSSP database of protein structure-sequence alignments. Nucleic Acids Res **25:**226-30.

22. **Shi, T., S. M. Spain, and D. L. Rabenstein.** 2004. Unexpectedly fast cis/trans isomerization of Xaa-Pro peptide bonds in disulfide-constrained cyclic peptides. J Am Chem Soc **126:**790-6.

23. **Song, J., K. Burrage, Z. Yuan, and T. Huber.** 2006. Prediction of cis/trans isomerization in proteins using PSI-BLAST profiles and secondary structure information. BMC Bioinformatics **7:**124.

24. **Terribilini, M., J. H. Lee, C. Yan, R. L. Jernigan, S. Carpenter, V. Honavar, and D. Dobbs.** 2006. Identifying interaction sites in "recalcitrant" proteins: predicted protein and RNA binding sites in rev proteins of HIV-1 and EIAV agree with experimental data. Pac Symp Biocomput:415-26.

25. **Terribilini, M., J. H. Lee, C. Yan, R. L. Jernigan, V. Honavar, and D. Dobbs.** 2006. Prediction of RNA binding sites in proteins from amino acid sequence. Rna **12:**1450-62.

26. **Terribilini, M., J. D. Sander, J. H. Lee, P. Zaback, R. L. Jernigan, V. Honavar, and D. Dobbs.** 2007. RNABindR: a server for analyzing and predicting RNA-binding sites in proteins. Nucleic Acids Res **35:**W578-84.

27. **Vapnik, V.** 1998. Statistical learning theory. Springer-Verlag., New York.

28. **Wang, G., and R. L. Dunbrack, Jr.** 2003. PISCES: a protein sequence culling server. Bioinformatics **19:**1589-91.

29. **Wang, M. L., W. J. Li, and W. B. Xu.** 2004. Support vector machines for prediction of peptidyl prolyl cis/trans isomerization. J Pept Res **63:**23-8.

30. **Wedemeyer, W. J., E. Welker, and H. A. Scheraga.** 2002. Proline cis-trans isomerization and protein folding. Biochemistry **41:**14637-44.

31. **Witten, I. H., and E. Frank.** 2000. Data Mining: Practical machine learning tools with Java implementations. Morgan Kaufmann.

32. **Wulf, G., G. Finn, F. Suizu, and K. P. Lu.** 2005. Phosphorylation-specific prolyl isomerization: is there an underlying theme? Nat Cell Biol **7:**435-41.

33. **Yan, C., D. Dobbs, and V. Honavar.** 2004. A two-stage classifier for identification of protein-protein interface residues. Bioinformatics **20 Suppl 1:**I371-I378.

34. **Yan, C., M. Terribilini, F. Wu, R. L. Jernigan, D. Dobbs, and V. Honavar.** 2006. Predicting DNA-binding sites of proteins from amino acid sequence. BMC Bioinformatics **7:**262.

ACKNOWLEDGEMENTS

Above all, I would like to express my earnest gratitude to my major professor Professor Drena Dobbs. Her advice was always helpful when I needed to talk and discuss about my graduate study. She taught me the way of studying and doing do researches as an independent scientist. Endless passion for the scientific research is the best thing I learned from her. She always brought new research projects and directions which is the best part of my graduate study and I really enjoyed it.

I also would like to thank to my co-major professor, Professor Kai-Ming Ho in physics department. He supported and guided me for providing computational advices and suggestions for my study. Also I would like to thank to Professor Susan Carpenter and Professor Gloria Culver. They provided me opportunities to perform experiments and always gave me a good advice for my graduate study. The discussion with them always corrected my research directions.

I would like to express my thanks to Professor Amy Andreotti, Professor Vasant Honavar and Professor Robert Jernigan for serving as my program study of committee. Their advices and support guided my graduate study to the right direction.

I'm also grateful to all members in Drena Dobbs's lab, Michael Terribilini, Peter Zaback and Jeffry Sander. We learned from each other in discussing lots of things from scientific research to social problems. I really enjoyed working and talking with them.

Last but certainly not least, I really want to express my sincere gratitude to my wife, Suh-Yeon. Without her, I cannot imagine my graduate life and study in Iowa State University. We always had a lot of discussion for real life problems as a friend and wife as well as scientific problems as a scientific collaborator and adviser. Most of all, I want to dedicate my work to all of my family members who supported for my graduate study.