STRUCTURES OF THE *KLEBSIELLA OXYTOCA* PHAGE PHI KO2 AND VIBRIO HARVEYI MYOVIRUS-LIKE PROTELOMERASE FAR C-TERMINAL DOMAINS

by

Diana K. Smith

A dissertation submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Medicinal Chemistry

University of Utah

August 2010

The University of Utah Graduate School

STATEMENT OF DISSERTATION APPROVAL

The dissertation of **Diana K. Smith**

has been approved by the following supervisory committee members:

Darrell Davis	, Chair	**April 28, 2010** Date Approved
Chris Ireland	, Member	**April 28, 2010** Date Approved
Michael Kay	, Member	**April 28, 2010** Date Approved
Eric Schmidt	, Member	**April 28, 2010** Date Approved
Wes Sundquist	, Member	**April 28, 2010** Date Approved

and by **Chris Ireland**, Chair of the Department of **Medicinal Chemistry**

and by Charles A. Wight, Dean of The Graduate School.

ABSTRACT

A small but growing number of bacteria and phages are known to contain linear, hairpin-ended genomes. The hairpin "protelomeres" are created by the action of a dedicated enzyme known as protelomerase that acts on a palindromic DNA target sequence. Phage protelomerases are typically longer than their bacterial counterparts and contain an additional far C-terminal region of limited sequence conservation. Studies of the protelomerase of the *Klebsiella oxytoca* phage ΦKO2 have shown that although the far C-terminal region is not required to produce hairpin ends, truncation of the region has a drastic effect on enzyme kinetics. To date, no other studies have been reported on the far C-terminal region of this or any other protelomerase. We present the solution structures of the far C-terminal regions of two phage protelomerases. The regions form homologous, compact structures that adopt a fold similar to the canonical double-stranded RNA-binding domain and have been called the far C-terminal domains. Sequence alignment and secondary structure predictions show that all known and putative phage protelomerases contain C-terminal regions which will almost certainly form homologous domains. A sequence comparison of these proteins with all known protelomerases is presented, along with an analysis of the sequence and structure of proteins which adopt a similar fold. Based on structure homology and comparative sequence conservation of key binding regions, we propose that the domain belongs to the growing family of three stranded β-sheet DNA-binding proteins that is a subclass of the double-stranded RNA-binding domain superfamily.

To my Matt, my Mimsy, and my munchkins

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Background

The "chromosome end problem"

Genomic stability rests upon the ability to replicate and repair a complete genome with fidelity. Fidelity is ensured in both repair and replication by the action of a polymerase, which interprets the blueprint of a complementary strand to synthesize a new polynucleotide sequence. Typically, the synthesis of a new strand by a polymerase also requires a primer, an existing oligonucleotide with a free 3'-OH, on to which it adds additional nucleotides. The necessity of an existing 5' primer results in a requirement for specialized care for terminal genomic DNA. All known DNA polymerases operate in a 5'-3' direction. Successive rounds of genomic replication lead to ongoing shortening of the lagging strand of linear DNA with subsequent eventual loss of genetic information (*1*). Additionally, free or exposed genome ends are recognized by the cell as damaged or foreign DNA, and are subject to normal means of cellular protection and repair, including degradation by exonucleases and recombination via nonhomologous end joining (*2*, *3*). The complications of replication and repair at regions of the terminal regions of genomes constitutes what has been called the "chromosome end problem."

Organisms across all kingdoms have developed different strategies for maintenance of genomic termini. Typically, eukaryotes protect their linear genomes with

the addition of noncoding sequences to DNA ends. Drosophila chromosomes, for instance, are maintained by the addition of a transposon sequence to the termini ((*4*, *5*), and (*6*) for a recent review.) More common is the insertion of repeated, guanine rich sequence regions which fold into specialized structures. The folded structures interact with dedicated proteins, forming the protective nucleoprotein complexes called telomeres ((*7-9*) for reviews). Using an entirely different strategy, bacterial and Archaeal genomes are usually circular, bypassing the issue of exposed DNA ends entirely (summarized in (*10*).

Although relatively rare, notable exceptions to this broad categorization of genome architecture are known. There are several known examples, for instance, of bacteria with linear genomes. Among these, there are two general methods used to protect exposed genomic termini. See Figure 1.1 for an overview of the classes of linear bacterial and phage genomes. In the first class, as with linear eukaryotic chromosomes, terminal DNA ends are protected by associated proteins. Representatives of this class include *Streptomyces lividans* (*11*) and *Saccharopolyspora erythraea (12)*, in which the free ends of linear genomic DNA are protected by 5' covalent association with "terminal proteins" (TPs.) Some viral pathogens also employ this method, including adenovirus (*13*) and *Bacillus subtilis* phage Φ29 (*14*). Under this means of genome protection, the covalently attached proteins also serve to prime DNA synthesis during replication. In adenovirus and phage Φ29 replication, the attached protein is the origin of replication, while in *Streptomyces*, replication initiates from an interior location, and the attached terminal protein serves to help initiate "end patching" of the single-stranded region that results from replication ((*15*), and (*16*) for a recent review of the *Streptomyces*

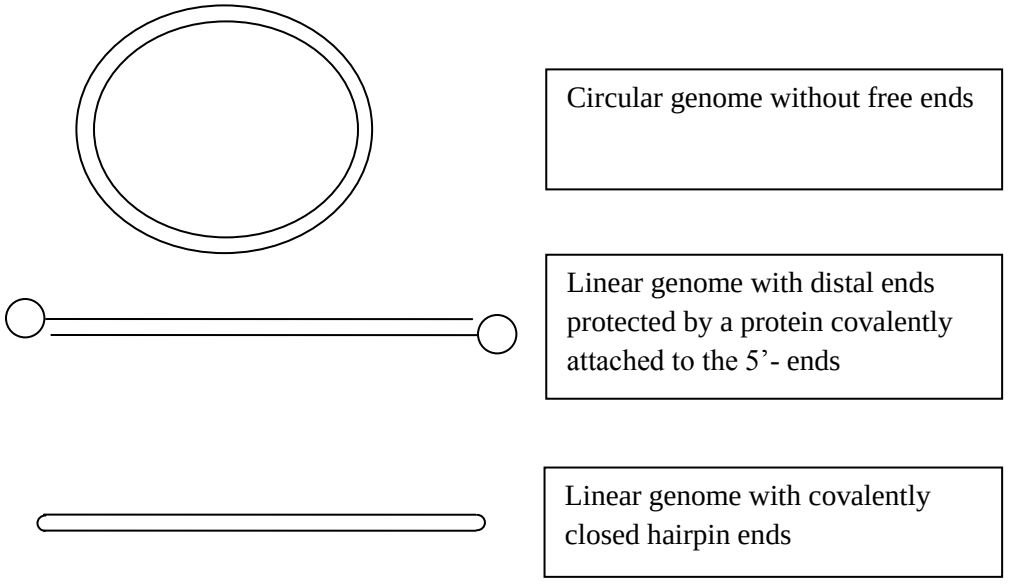


Figure 1.1 Approaches to the "chromosome end problem" by bacteria and viruses

chromosome.) Another solution to the chromosome end replication problem has also been discovered in both bacteria and viruses with linear genomes. In these systems, the termini of linear chromosomes or plasmids are covalently closed and protected by terminal "hairpin caps," which prevent both exonucleolytic degradation and terminus shortening during replication (*17, 18*). The hairpin caps have been shown to be covalently closed by a combination of denaturation/renaturation studies, direct visual examination of denatured, single stranded circles through microscopy, and Maxam-Gilbert sequence analysis (*19-21*). Because of the protection afforded the genome termini by the hairpin caps, the term "protelomere" was coined to represent "prokaryotic telomeres" (*22*).

Hairpin-capped genomes and their protelomeres

In all systems so far studied, replication of a linear, hairpin-ended genome is relatively simple. Common replication enzymes initiate bi-directional replication at an interior location of origin resulting in a circular, concatenated genomic dimer (*23*, *24*). In vitro assays have shown that the interlinked genome is resolved into monomers by the action of a single dedicated enzyme *(17, 25-29).* The process of resolution into monomers produces the terminal hairpin ends, so the enzyme has been called both a "protelomerase" and a "telomere resolvase" (*25, 26*). The presence of a protelomerase-type protein in a bacterium is a good indication that the system carries a linear replicon. Table 1.1 lists the known protelomerases and the organisms in which they were found, and gives a brief description of each system.

.

Table 1.1

Organisms that contain known or putative protelomerases

Organism	Enzyme length	Description
Borrelia burgdorferi	449	A gram-negative spirochetal bacteria that is a significant, tick-borne, human pathogen and the causative agent of Lyme disease (*30*). Symptoms of Lyme disease include fever, chills, headache, and musculoskeletal pain. Progressive illness results in facial palsy, arthritis, heart palpitations, and chronic neurological complaints, including tremor and short term memory loss. Infections have been associated with non-Hodgkin lymphoma.(*31*) The fragmented genome consists of both circular and linear chromosomes plus 12 linear and nine circular plasmids (*19, 20, 32-34*). The protelomerase gene (called ResT) is essential for *B. burgdorferi* survival (*35, 36*).
Borrelia hermsii	449	A gram-negative, spirochetal bacteria and a significant, tick-borne human pathogen, *B. hermsii* is the causative agent of relapsing fever. Manifestations include headache, myalgia, arthralgia, chills, vomiting, and abdominal pain (*37, 38*).
Agrobacterium tumifaciens	442	A rod-shaped, gram-negative bacterium of the family Rhizobiaceae found ubiquitously in soil, *A. tumefaciens* causes crown gall disease in plants by inserting a fraction of its DNA (called T-DNA) in the host genome (*39, 40*). The wide variety of plants affected by the bacterium (including grape vines, stone fruits, nut trees, sugar beets, horseradish, and rhubarb) makes it of great concern to the agriculture industry (*41*). The segmented genome consists of one linear and one circular chromosome, and two circular plasmids.

Table 1.1 continued

Organism	Enzyme length	Description
Escherichia coli phage N15	630	A non-integrative, temperate, lambdoid-like phage of the Siphoviridae family (*22*, *42*). *E. coli* is a gram negative rod-shaped bacterium commonly found in the lower intestine of warm-blooded organisms (endotherms).
Halomonas aquamarina phage ΦHAP-1	520	A myovirus-like temperate phage induced from *H. aquamarina*, the gram-negative halophilic gammaproteobacterium that has been isolated from a variety of marine and hypersaline environments, including the pelagic ocean, deep-sea hydrothermal vents, the brine-seawater interface of deep-sea brine pools, and coastal surface waters. The PhiHAP-1 genome shares synteny and gene similarity with coliphage N15 and vibriophages VP882 and VHML (*43*).
Klebsiella oxytoca phage ΦKO2	640	A non-integrated prophage of *K. oxytoca* with a genome size of approximately 51.6kb. Genome organization is similar to the coliphage N15 (*44*). *K. oxytoca* is a gram-negative, rod-shaped bacterium closely related to *K. pneumoniae*
Vibrio harveyi phage VHML	509	VHML (Vibrio harveyi myovirus like) is a temperate phage classified in the Myoviridae family. It infects the free living gram-negative marine bacterium *V. harveyi*, which causes the normally non-pathogenic host to become virulent to a variety of aquatic organisms (*45*). Economic loss in recent years due to infection of aquacultured species has been particularly devastating (*46*, *47*). The phage was sequenced, but the host-phage system was reported as lost (*43*). VHML has also been reported to lyse the strains the strains *Vibrio alginolyticus* ACMM102 and *Vibrio cholerae* ATCC 14035.
Vibrio parahaemolyticus O3:K6 *phage VP58.5*		A 42.612 kb myovirus closely related to VHML that was isolated from a *V. parahaemolyticus* strain belonging to the serovar O3:K6 pandemic clonal complex. The clone has been associated with many seafood-borne diarrhea outbreaks in Southeast Asia and South America, particularly Chile (*48*).

Table 1.1 continued

Organism	Enzyme length	Description
Vibrio parahaemolyticus O3:K6 phage VP882	538	A Myoviridae bacteriophage isolated from a pandemic strain of *V. parahaemolyticus* that also infects and lyses high proportions of *Vibrio vulnificus* and *Vibrio cholerae* (*49*). *V. parahaemolyticus* is a gram negative, rod-shaped, motile, facultatively aerobic bacterium found in brackish saltwater. When ingested, it causes gastrointestinal illness in humans. Wound, eye, and ear infections can also develop from swimming or working in affected waters (*50*).
Yersinia enterococolitica phage PY54	617	*Yersinia enterocolitica* is a member of the family Enterobacteriaceae, some strains of which are enteropathogenic to humans, and are predominantly transmitted by ingestion of undercooked meat (*51, 52*). Non-pathogenic strains are found readily in the environment and may have potential to become pathogenic. PY54 is a linear plasmid prophage with a genome size of approximately 46kb which infects *Y. enterococolitica*. (*53*)
Vibrio campbelli	691	A gram-negative, facultative anaerobe which contains a putative protelomerase (DBSOURCE accession code ABGR01000073.1.) It is likely that the protelomerase is encoded by a phage which infects the bacteria.

Protelomerase overview

Creation and maintenance of hairpin protelomeres

Creation and maintenance of a hairpin protelomeres requires only a protelomerase and the DNA recognition sequence upon which it acts. Indeed, normally circular plasmids which have been engineered to encode the *Escherichia coli* phage N15 protelomerase and its DNA recognition sequence have been shown to linearize *in vivo* once transformed into *E. coli*, and to replicate stably as hairpin-ended linear plasmids (*18*). The DNA binding site of all known protelomerases is a palindromic inverted repeat at the junction of the genomic dimers (see Figure 1.2). Studies on the protelomerases from the ΦKO2 and N15 phages and from the bacterial *Borrelia burgdorferi* enzyme (usually called ResT) have shown that the protelomerase enzymes, which are monomeric in solution, bind target DNA as dimers, creating transient staggered cuts six base-pairs apart on opposite strands, each three base-pairs away from the axis of dyad symmetry of the palindrome (*29, 54, 55*). The 6 bp overhangs created by the staggered cuts fold over and are joined to the opposite strand to form the hairpin caps.

Protelomerase nearest neighbor and reaction mechanism

Protelomere resolvases share sequence homology and an expected two-step reaction mechanism with the site-specific integrases called tyrosine recombinases or lambda integrases, as well as with type IB topoisomerases (*28, 29, 56-58*) (see also (*59*) for a minireview.) Site-specific tyrosine recombinases/lambda integrases are tetrameric and bind to two duplex DNAs, catalyzing the exchange of four DNA strands to integrate or excise a DNA segment from the host genome (*60, 61*). IB topoisomerases are found across eukarya, in many bacterial genera, and two known families of eukaryotic viruses

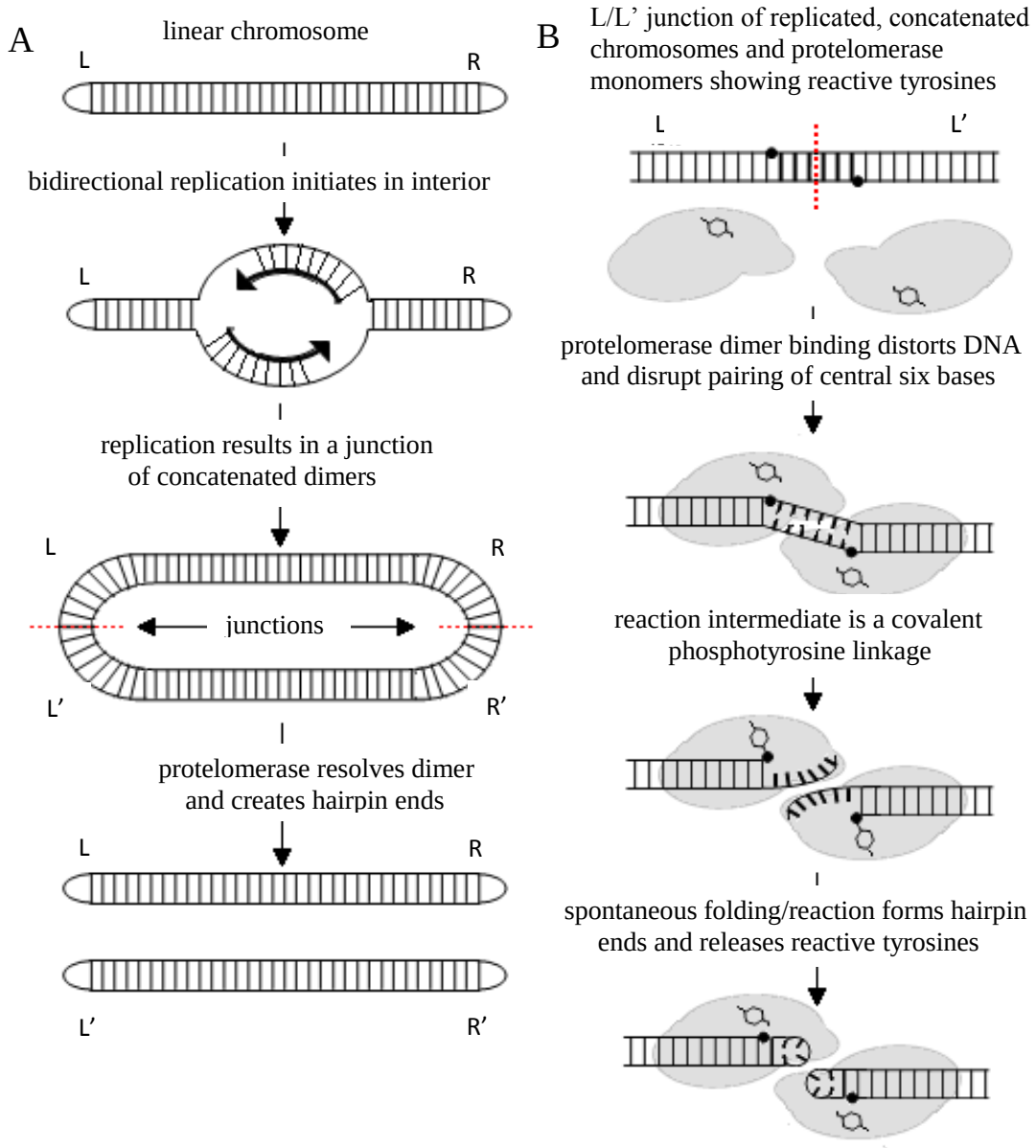


Figure 1.2 Schematic: Protelomerases resolve replicated DNA to form hairpin telomeres

Based on the figure from Aihara et al, Molecular Cell, Volume 27, Issue 6, 21 September 2007, Pages 901-913

A) Cartoon showing the current understanding of replication of hairpin telomere-containing linear genomes. Replication of a linear chromosome with hairpin telomeres produces a dimeric circular intermediate that is resolved into unit-length chromosomes by the activity of protelomerase. L and R refer to left and right hairpin ends, respectively.

B) A model for the hairpin formation reaction by the protelomerase Tel-KO2, proposed based on the crystal structure presented in the study. The dots represent the phosphates at the sites of cleavage.

(poxvirus and mimivirus, (*59, 62*).) Type IB topoisomerases are monomeric, and act to relax supercoiling by cleaving and rejoining a single strand of duplex DNA, which is first allowed to rotate relative to the opposite strand (*63*). The reaction mechanism for all three enzyme families is similar. A conserved catalytic pentad (R-K-R-H-Y) acts to execute the cleavage and religation of DNA by way of consecutive transesterification reactions without the use of a cofactor. In the reaction, a nucleophilic attack by a catalytic tyrosine on the scissile phosphodiester bond within the DNA target site results in the formation of a covalent 3' DNA-phosphotyrosyl linkage and a free 5' hydroxyl. In the second half of the reaction, the free 5' hydroxyl from the same strand (IB topoisomerases), complementary strand (protelomerases), or a strand from a neighboring duplex (lambda integrases) attacks the phosphotyrosyl intermediate in a second nucleophilic reaction, regenerating a joined DNA strand and releasing the catalytic tyrosine.

Protelomerase domain organization

All known protelomerases contain a highly conserved central region surrounded by N-terminal and C-terminal regions of variable length, sequence, and secondary structure prediction. Protelomerases can be roughly broken into two clades based primarily on size. The bacterial protelomerases are shorter, approximately 450 residues in length, while phage protelomerases are longer; most consist of more than 600 residues. The phage protelomerases contain an additional C-terminal region of varying length which terminates in a short region of moderate conservation. The conserved central core is highly homologous to tyrosine recombinases and Type IB topoisomerases, and contains all five of the catalytic pentad residues.

Protelomerase structure

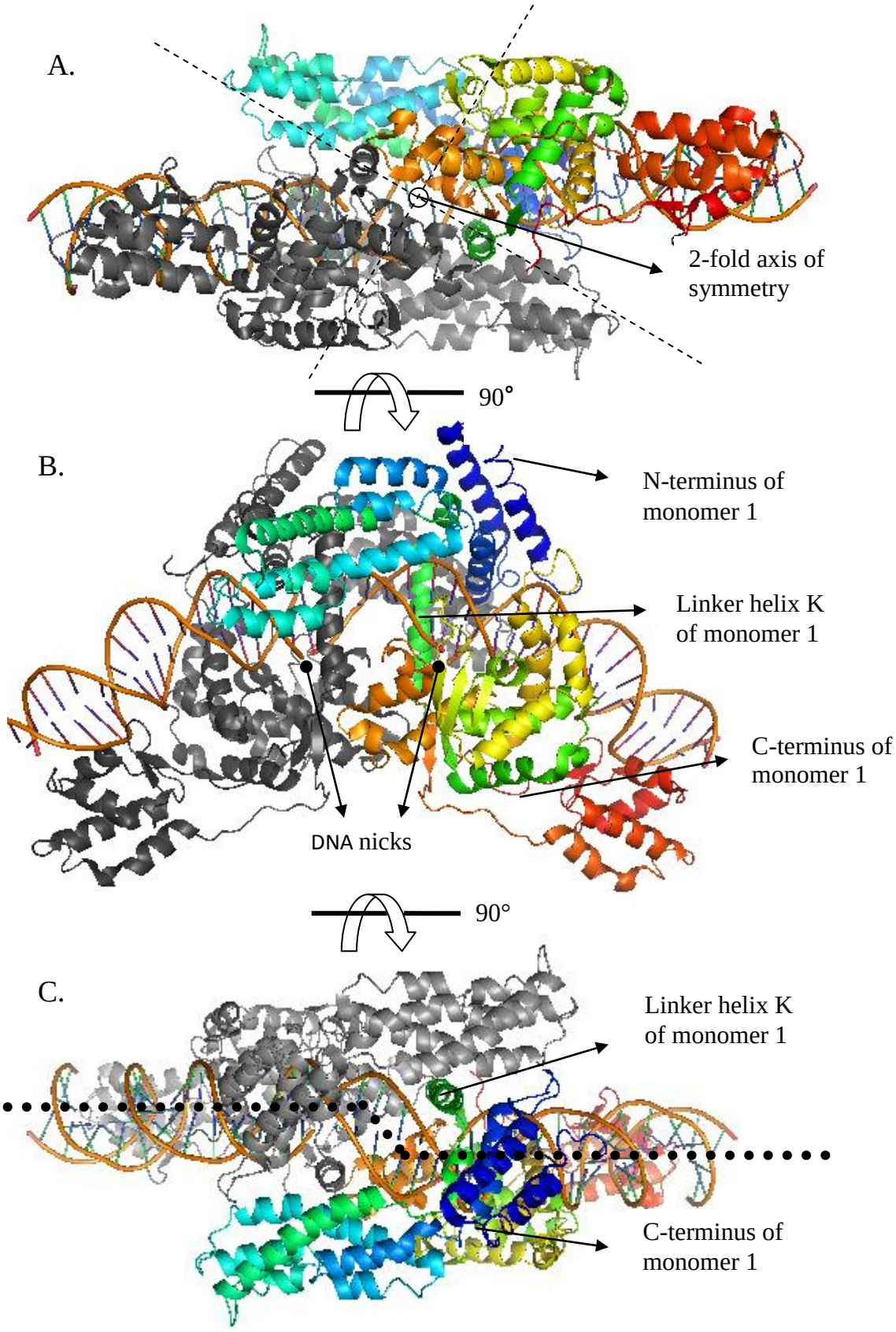
The crystal structure of the minimally catalytic, N-terminal portion (531 of 640 residues) of the *Klebsiella oxytoca* phage ΦKO2 protelomerase (Tel-KO2) complexed to the central 44 base-pairs of its 50 base-pair DNA target site was recently solved (64). To "trap" the enzyme in an intermediate state, the DNA target was nicked, and contained orthovanadates (VO_4^{3-}) to mimic the pentavalent transition state of the scissile DNA phosphate cleavage. The crystallography was simplified by slightly altering the DNA sequence to make it symmetric (two bases on each side normally interrupt the otherwise perfect symmetry) with an additional nucleotide substitution at the distal end. The structure shows two interlocked Tel-KO2 monomers each bound to a half-site of the DNA target (see Figure 1.3). The dimerically bound proteins show extensive interprotein subunit and protein-DNA contacts. The Tel-KO2 monomer subunits bind to their DNA target with their C-terminal domains on the distal ends of the DNA target site and the N-terminal domains interlinking near the central six base-pairs of the target site. Each monomer contains an extended linker helix (K) in between its first and second domains. When the monomer binds to its target site, helix K rests in the major groove, allowing the monomer to wrap completely around the DNA while still remaining predominantly on one side of the DNA. The authors propose that the enzyme dimer functions by bending and "springloading" the DNA, and that following nicking, the 6 base-pair, 5'-overhanging ends reorganize spontaneously within the enzyme active site, allowing the protelomerase to act as a ligase to produce two hairpins ends from the original duplex DNA. The authors further surmise that the reaction is made essentially irreversible through the surprising level of distention of the DNA visible in the crystal structure,

Figure 1.3 Structure of the Tel-KO2 Dimer Bound to DNA.

A) "Bottom view" of the structure oriented parallel to the two-fold non-crystallographic axis with the DNA helix extending horizontally from left to right, showing the extensive coverage of the DNA by the protelomerase dimer.

B) "Side view" of the structure (rotated 90° along the z-axis relative to the view in (A)) showing the protein-induced curvature of the DNA. DNA nicking sites are indicated.

C) "Top view" of the structure (rotated 90° along the z axis relative to the view in (B)) showing the offset in the DNA helical axis, highlighted by dotted lines indicating the axes for each half-site. In all views, one monomer is colored grey while the second monomer is colored with a spectrum in which the N-terminus begins in blue and the C terminus finishes in red. N and C termini and linker helix K of monomer 1 are indicated.


A.
2-fold axis of symmetry
90°
B.
N-terminus of monomer 1
Linker helix K of monomer 1
C-terminus of monomer 1
DNA nicks
90°
C.
Linker helix K of monomer 1
C-terminus of monomer 1

which they presume is induced by protein-binding. The DNA is not only bent in a curved arc (73° within a plane parallel to the two-fold axis of the complex), but the ends are bent outward from their otherwise linear path. The structure indicates that the disruption of the dimer interface may be required for hairpin formation following initial strand cleavage. The Tel-KO2 monomers are described as having three domains: an amino-terminal domain, the central portion of which the authors refer to as the "muzzle," the highly conserved catalytic domain containing all five residues which constitute the catalytic pentad, and a carboxy-terminal domain, which the authors call the "stirrup" (see Figure 1.4). The linker helix K connects the N-terminal and catalytic domains. The C-terminal/stirrup domain is connected to the catalytic domain by an extended segment. The "stirrup" contacts the DNA, but does not interact extensively with other portions of the protein. The carboxy end of the stirrup domain seems to loop back toward the catalytic domain, but there is no known or presumed function for this configuration, which may or may not be an artifact of the shortened construct. The protelomerase dimer extensively covers the DNA, with the exception of a small, 8-9 basepair stretch of the target site eleven basepairs from either side of the palindrome center.

The domain organization of the Tel-KO2 protein monomers is similar to, but distinct from, the lambda integrase/tyrosine recombinases with which protelomerase resolvases share sequence homology. The core portion of lambda integrase consists of a two-lobed architecture consisting of an amino-terminal domain (sometimes called a core-binding domain) and a catalytic, or core domain, connected by an extended linker (*60*). The catalytic domains of the lambda integrase and Tel-KO2 proteins are especially well-conserved by both sequence similarity and secondary structural organization, as

Figure 1.4 Surface of the Tel-KO2 dimer.

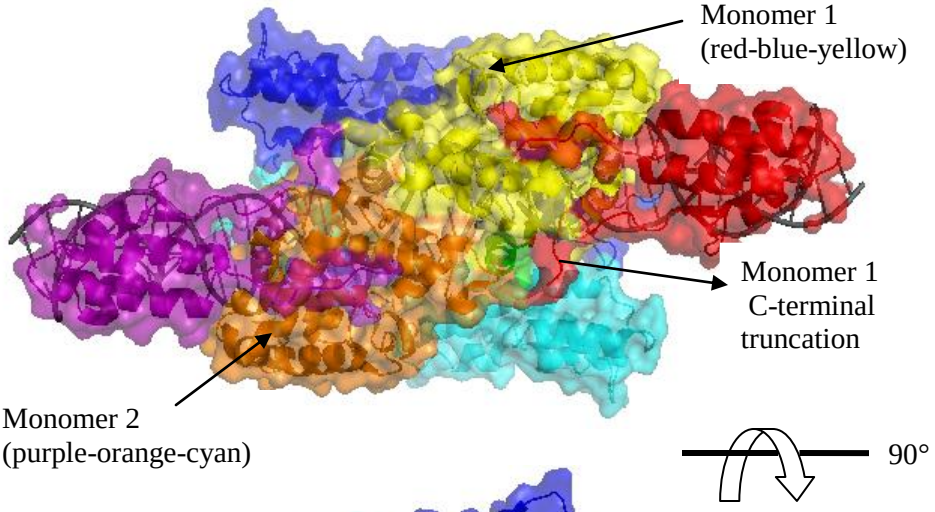
The figures are colored as follows: monomer 1: N-terminal muzzle domain in blue, linker helix in green, catalytic domain in yellow, and C-terminal stirrup domain in red; monomer 2: N-terminal muzzle domain in cyan, linker helix in gray-green, catalytic domain in orange, and C-terminal stirrup domain in purple. The views are oriented identically to those in figure X.X as follows:

A) Bottom view of the dimer, showing the extensive coverage of the DNA binding site by Tel-KO2, and the interactions of the catalytic domains (colored orange and yellow) of the monomer subunits.

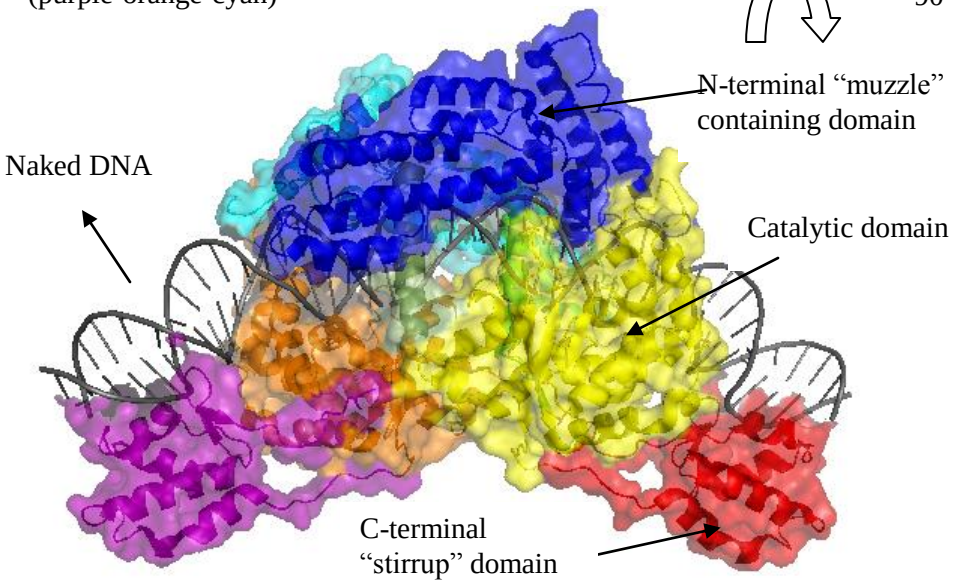
B) Side view of the dimer, showing that while each monomer wraps around the DNA, each monomer lies predominantly on one side of the DNA arc.

C) Top view of the dimer, showing the extensive interaction of the N-terminal muzzle domains (colored blue and cyan) with each other, which also interact with the DNA on the neighboring half-site. Each N-terminal muzzle domain cradles the linker helix K of its neighbor helix.

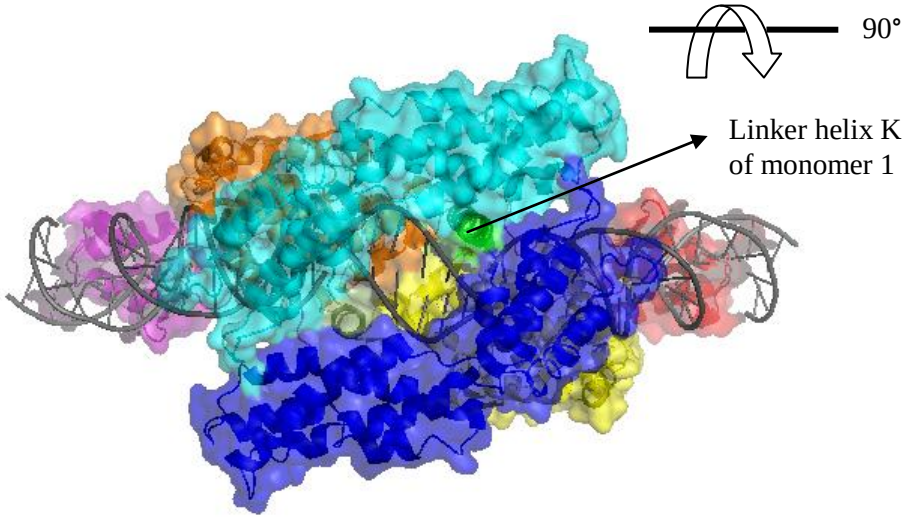
A.


Monomer 1
(red-blue-yellow)
Monomer 1
C-terminal
truncation
Monomer 2
(purple-orange-cyan)
90°

B.


N-terminal "muzzle"
containing domain
Naked DNA
Catalytic domain
C-terminal
"stirrup" domain

C.


90°
Linker helix K
of monomer 1

observable in a calculated DaliLite Pairwise comparison (see Figure 1.5) (*65*). However, the structures differ in several instances. Lambda integrase does not contain a homologous carboxy terminal "stirrup" domain, and the amino-terminal/core-binding domain is much smaller, seemingly lacking the majority of the "muzzle" present in protelomerase KO2 (helices D-I, residues 79-200). Additionally, lambda integrase contains an amino terminal arm-binding domain not present in Tel-KO2, which is connected by an extended linker. This domain is crucial to the functioning of lambda integrase. It interacts with the arm-binding domains of three other integrase subunits to form a functioning tetramer that binds the accessory (arm) DNAs involved in recombination. The arm-binding tetramer/accessory DNA interaction appears to shape the recombination complex in a way that suggests arm binding shifts the reaction equilibrium in favor of recombinant products (*60*).

Preliminary work

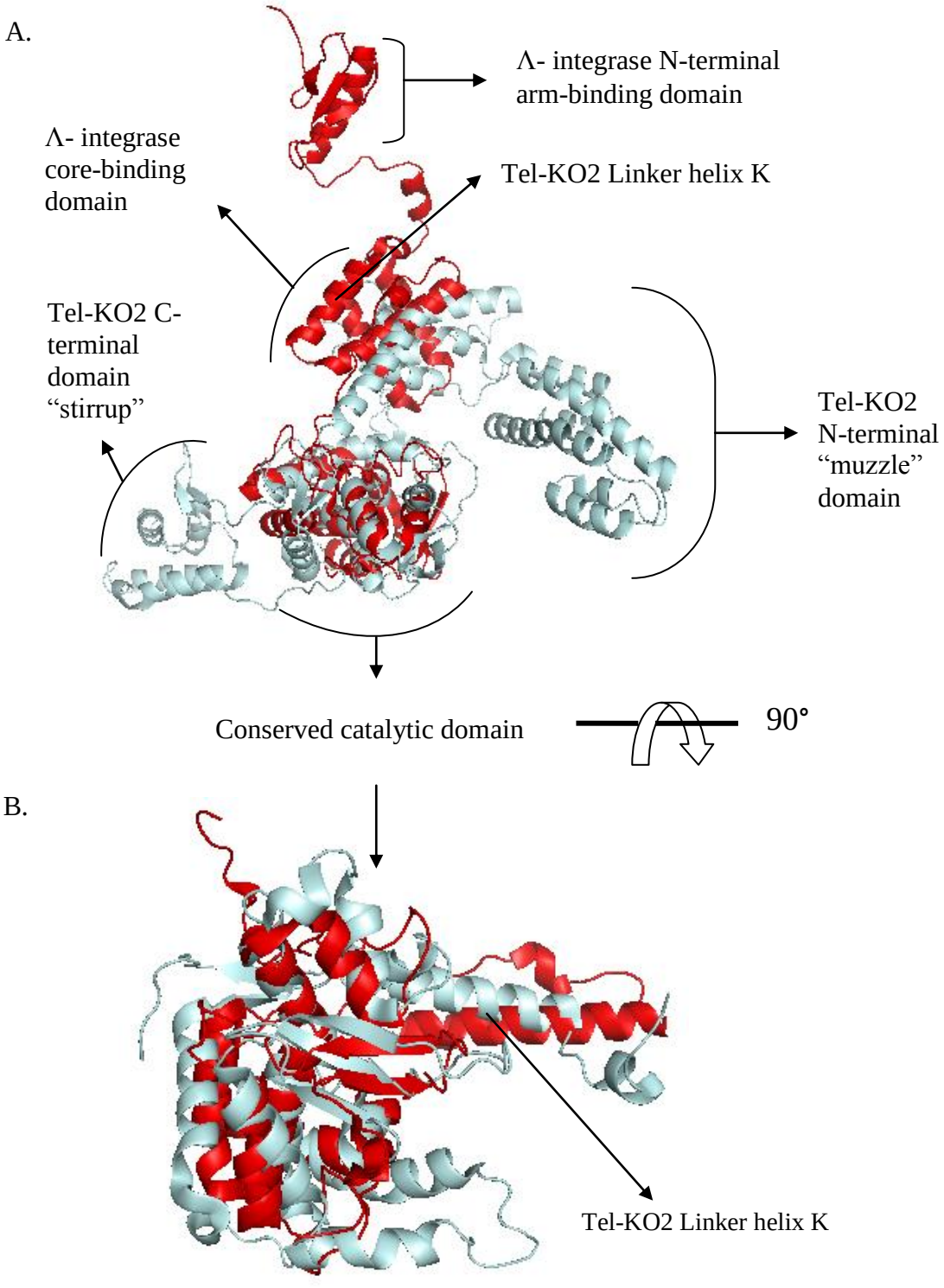
Far C-terminal region

The length of the Tel-KO2 construct used for crystallographic analysis represented the shortest construct shown to be catalytically active during *in vitro* deletion studies (Huang, W.M., personal communication, unpublished data.) The studies show that C-terminally truncated enzyme can resolve oligomer substrates containing the full-length Tel-KO2 target site into hairpin products *in vitro.* All deletion mutants studied (truncated to 605, 545, 538, and 531 from the full length 640-residue protein) are able to produce hairpins in the *in vitro* assay. However, a resolution kinetics comparison study of the shortest truncation (TelK531) with full length (TelK640) enzyme show that while the truncated enzyme is still functional, the efficiency of hairpin turnover is greatly

Figure 1.5 Domain architecture comparison of the published structures of the C-terminally truncated *Klebsiella oxytoca* phage ΦKO2 protelomere resolvase (Tel-KO2) and lambda integrase as calculated in a DaliLite pairwise comparison.

(A) Superposition of monomers. The Tel-KO2 monomer is shown in light cyan and the integrase monomer is shown in red.

(B) Close-up view of conserved catalytic regions, which have been rotated 90° along the z axis relative to the full-protein overlay. Coloring is the same as in (A).


A.
Λ- integrase N-terminal arm-binding domain
Λ- integrase core-binding domain
Tel-KO2 Linker helix K
Tel-KO2 C-terminal domain "stirrup"
Tel-KO2 N-terminal "muzzle" domain
Conserved catalytic domain
90°
B.
Tel-KO2 Linker helix K

reduced, and functions only at approximately 1/50th the rate of full-length enzyme (see Figure 1.6.)

The Huang group also studied the effect of truncating the DNA target site on protelomerase K function. An *in vitro* analysis revealed that distally truncating the DNA target site by as little as three base pairs on both sides of the palindromic 56 base-pair target sequence reduced the efficiency of full-length Tel-KO2 (1-640) by 75%. Further truncating the DNA target site to contain only the central 42 base pairs (seven base-pairs removed from each end) decreased efficiency of Tel-K640 to less than 1%. Surprisingly, the most highly truncated enzyme (TelK531) was better able to process the shortened target site substrates, with an efficiency that was restored to nearly 15% of full-length Tel-KO2 acting on full-length oligomer substrates (see Figure 1.7). Since this combination permitted catalytic activity (although kinetic efficiency was reduced compared to full length protein and substrate), a C-terminally truncated Tel-KO2 and a minimal DNA target were selected for structural studies by crystallographic analysis.

Although a truncated Tel-KO2 was found to be sufficient for processing truncated DNA target sites *in vitro*, the unique effects of the C-terminal region on the function of Tel-KO2 in substrate processing indicated that this region warranted further study. Removal of the C-terminal domain negatively affects the ability of Tel-KO2 to produce hairpins from linear, double-stranded target DNA, and the distal ends of the target site also have an effect on catalysis that cannot be explained by interaction with the catalytically active portion of the protelomerase. Given the *in vitro* study results, it seems probable that the far C-terminal region of the protelomerase interacts with the distal DNA ends, the core portion of the protelomerase, or both. The C-terminal region of Tel-KO2

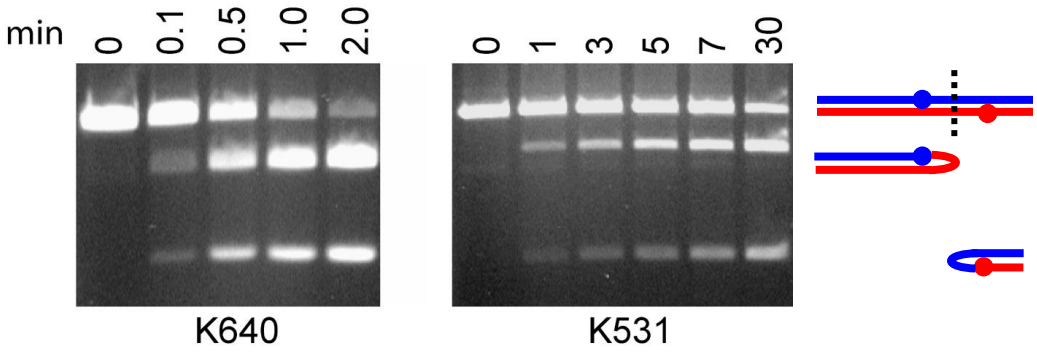
Figure 1.6 Comparison of the kinetics of hairpin resolution of replicated oligonucleotide substrate by full length (1-640) and truncated (1-531) Tel-KO2.

The oligonucleotide target sequence contains the full length 56 base pair binding site plus an additional ten base-pairs for better visualization of the resulting hairpins. Hairpin turnover by truncated Tel-KO2 occurs at $1/50^{th}$ the rate of full length (1-640) Tel-KO2.

A) Agarose gels comparing resolution progression.

B) Reaction profile showing the rate of truncated Tel-KO2(1-531) activity (squares) compared to the rate of full-length Tel-KO2 (1-640) activity (circles). Activity is scaled to percent hairpin formation of wild type (full-length) Tel-KO2. Wai Mun Huang lab, Pathology Department, University of Utah, unpublished data

A.


min
0
0.1
0.5
1.0
2.0
0
1
3
5
7
30
K640
K531

B.

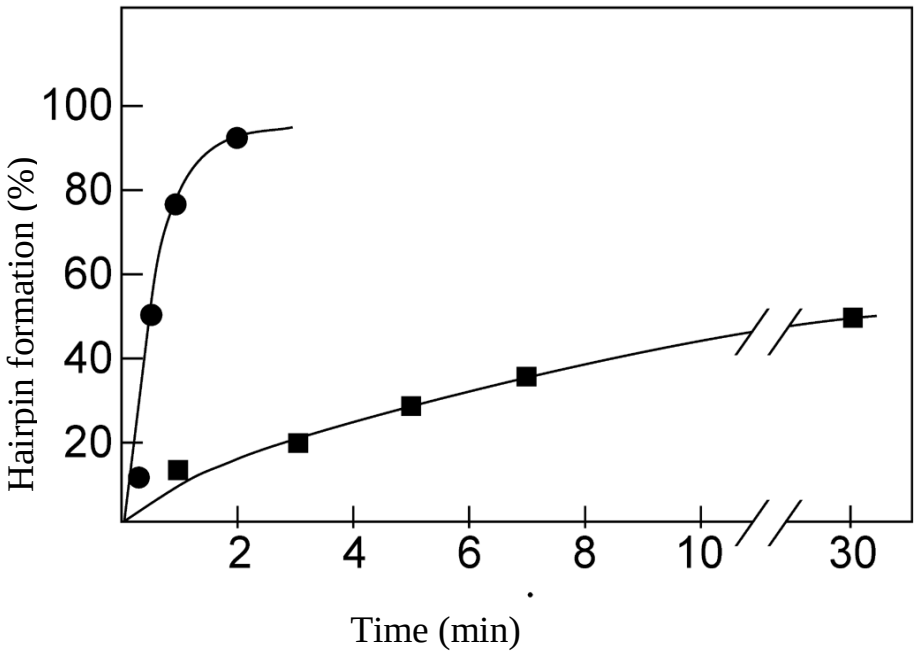
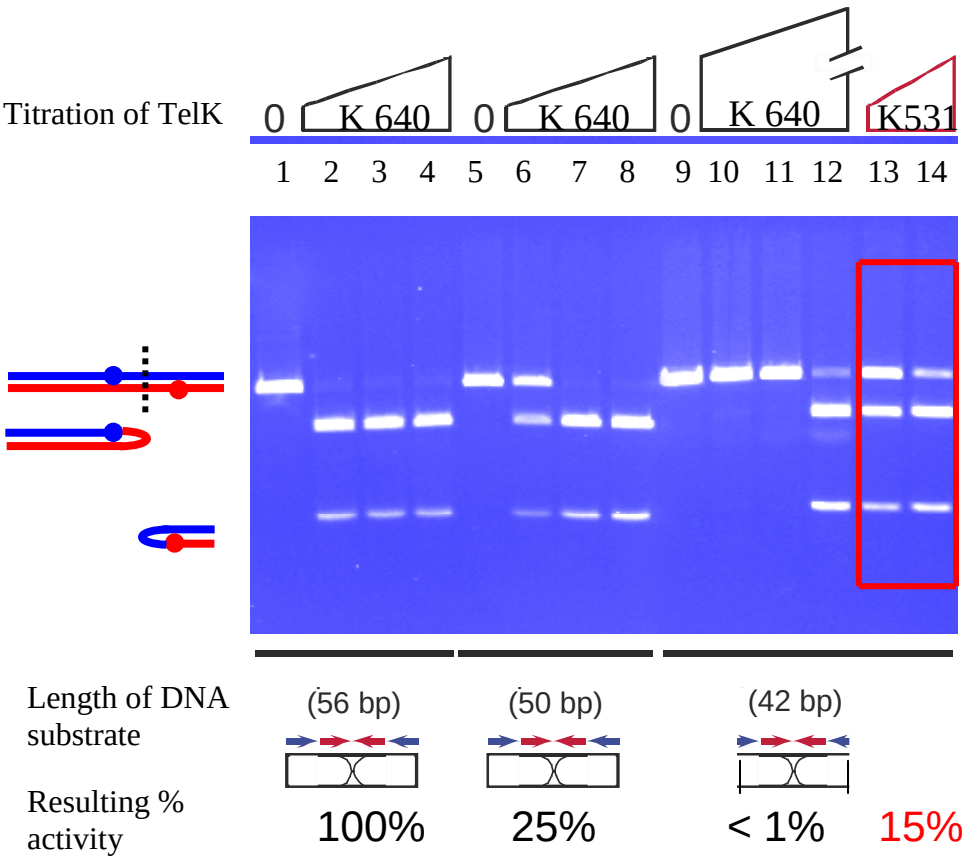

Hairpin formation (%)
100
80
60
40
20
2
4
6
8
10
30
Time (min)

Figure 1.7 Effect of DNA target site truncation on efficiency of Tel-KO2 hairpin resolution.

Reducing the length of the oligomer substrate decreases the efficiency of full-length Tel-KO2 in hairpin resolution. A C-terminally truncated Tel-KO2 (1-531) can partially restore efficiency for the shortest truncated substrate.

Lanes 1-4: oligonucleotide substrate contains full-length (50 base pair) binding site. Lanes 5-8: Tel-KO2 binding site contained in the oligonucleotide substrate has been truncated by three base-pairs on either end. Resulting oligonucleotide contains the central 50 base-pairs of the Tel-KO2 binding site. Lanes 9-14: Tel-KO2 binding site contained in the oligonucleotide substrate has been truncated by seven base-pairs on either end. Resulting oligonucleotide contains the central 42 base-pairs of the Tel-KO2 binding site. Lanes 1, 5, and 9: no protelomerase added. Lanes 2-4 and 6-8: increasing amounts of full length Tel-KO2 were titrated with the corresponding oligomer substrate. Lanes 10-12: A much greater higher concentration of Tel-KO2 used in lanes 2-4 and 6-8 was incubated with the 42 base-pair oligomer. Lanes 13 and 14: Truncated Tel-KO2 (1-531) was incubated with the 42 base-pair oligomer. Efficiency of activity of the full-length and truncated Tel-KO2 enzymes is shown as a percentage of efficiency of full-length enzyme acting on full-length oligomer substrate.

Wai Mun Huang lab, Pathology Department, University of Utah, unpublished data


Titration of TelK
0
K 640
0
K 640
0
K 640
K531
1 2 3 4 5 6 7 8 9 10 11 12 13 14
Length of DNA substrate
(56 bp)
(50 bp)
(42 bp)
Resulting % activity
100%
25%
< 1%
15%

contains an unusually high number of acidic residues and was deemed unsuitable for crystallographic analysis, so structural studies by NMR were proposed in order to elucidate the function of this interesting domain.

Research overview

Although they are relatively rare, the number of phage and bacterial organisms known to use covalently closed hairpin ends as a way of protecting their linear genomes is growing, and this means of genomic conservation is just beginning to be understood. The creation and maintenance of hairpin ends is accomplished by the action of a single, dedicated enzyme known as a protelomere resolvase, or protelomerase. A sequence alignment of protelomerases from phage and bacterial origins reveals that phage protelomerases are generally longer, and contain an additional C-terminal region consisting of a linker of variable length terminating in a short region of limited conservation. We have called this region the far C-terminal domain (farCTD.) The role of this extended region is not understood, but studies of the protelomerase from *Klebsiella oxytoca* phage ΦKO2 have shown that truncating the far CTD has drastic effects on enzyme functioning in *in vitro* studies. Protelomerase constructs lacking the far C-terminal domain can function *in vitro* to produce hairpins from synthetic DNA oligomers containing the native protelomerase binding site, but the truncated protelomerase functions at only a fraction of the normalized activity of the full length enzyme. Additionally, truncating the DNA binding site in synthetic DNA oligomers also dramatically decreases the efficiency of the protelomerase in a manner that cannot be explained by interaction with the catalytic portion of the molecule described in the solved structure. The function of the far C-terminal region is unknown. However, the

comparative in vitro studies of truncated and full-length protelomerase indicate that the far C-terminal region of Tel-KO2 plays an important role in enzyme function, either through interaction with the catalytic portion of the protelomerase, or with the DNA target itself.

To understand the role of the far C-terminal domain, we have asked a number of questions. How much of the C-terminal region of phage protelomerases is structured, and what is the structure of this domain? How much sequential and structural conservation is there between the far C-terminal domain of the *Klebsiella oxytoca* phage ΦKO2 protelomerase and those of other phage protelomerases? What is the likely binding partner of the protelomerase far CTD? And lastly, what relation does the phage protelomerase far CTD have to the bacterial protelomerase?

Chapter two reports my work on this project in attempting to answer these questions. We report the solution structure of the far CTD of two phage protelomerases. A comparison of the sequence of these protelomerase domains to the homologous regions of all known phage protelomerases is presented. We also include information about likely structural analogues in other proteins, a report of initial studies seeking to identify the farCTD binding partner, and a comparison to bacterial protelomerases.

CHAPTER 2

STRUCTURES OF THE *KLEBSIELLA OXYTOCA* PHAGE ΦKO2 AND *VIBRIO HARVEYI* PHAGE VHML PROTELOMERASE FAR C TERMINAL DOMAINS

Abstract

We present the solution structures of the far C-terminal domains of two phage protelomerases. A sequence comparison of these structures with all known protelomerases is presented, along with an analysis of the sequence and structure of proteins which adopt a similar fold. The solution structures show that the far C-terminal domain of phage protelomerases adopts a fold which is similar to the canonical dsRBD. Based on structure homology and limited sequence conservation, we propose that the domain belongs to the growing family of three stranded β-sheet DNA-binding proteins.

Introduction

A small but growing number of bacteria and viruses have been discovered to have linear genomes protected by covalently closed hairpin ends. The hairpin ends function like eukaryotic telomeres in protecting the genome from aberrant recombination and repair as well as from exonucleolytic degradation and have therefore been called "protelomeres" (*17-22, 43*). In organisms which employ this method of genomic

protection, bidirectional replication results in a concatenated genome dimer that is resolved into unit lengths by the action of a single, dedicated enzyme called a protelomere resolvase (protelomerase) (*23, 24, 26-29, 54*). The protelomerase binds as a dimer on a palindromic DNA sequence at the end-to-end junctions of the genome dimers and acts to produce and maintain the hairpin caps following replication.

Known protelomerases have a similar architectural domain organization. Protelomerases have a highly conserved central catalytic region, with amino and carboxyl regions of varying length and conservation. Phage protelomerases tend to be longer than their bacterial homologues, and have an additional region of varying length with limited sequence homology at the far C-terminal end of the protein. Although the presence of this far C-terminal region is universally conserved in phage protelomerases, the purpose of the domain is not yet understood. So far, only two studies regarding the far C-terminal domain (farCTD) have been undertaken, and both were in the *Klebsiella oxytoca* phage ΦKO2 system.

Current analyses of the ΦKO2 system show that the far C-terminal region is not required for catalytic activity, but that its absence has a dramatic negative effect on enzyme function. *In vitro* studies of the protelomerase of the *Klebsiella oxytoca* phage ΦKO2 (Tel-KO2, also called TelK) have shown that although a truncated protelomerase lacking this far C-terminal region is sufficient to form hairpins from synthetic double-stranded DNA oligomers containing the protelomerase recognition site, the efficiency of the enzyme is severely affected by shortening or removal of the far C-terminal region. Constructs lacking the far C-terminal region function at less than $1/10^{th}$ the rate of full length enzyme (Wai Mun Huang, personal communication, unpublished data).

Other protelomerase analyses are not explained by the current understanding of the catalytic region. The structure of C-terminally shortened protelomerase Tel-KO2 (lacking the far C-terminal region) bound to the central 44 basepairs of its target site was recently described (*64*). The structure indicates that the catalytic portion of the enzyme extends only to the central 42 base pairs of the normally 56 base pair target site. However, shortening the DNA target site has a dramatic negative effect on enzyme function (Wai Mun Huang, personal communication, unpublished data). *In vitro* assays measuring the ability of protelomerase to create hairpins from double-stranded oligomers containing the target site show that shortening the length of the DNA target site by as few as three base pairs per side reduces percent enzyme activity to only 25% of that of full-length protelomerase acting on full length (56 base pair) target site. Shortening the target site further (seven basepairs per side, resulting in a 42 basepair target site) reduces enzyme function even more, to less than 1% activity. The solved structure does not explain how the external regions of the DNA affect enzyme efficiency. Interestingly, enzyme constructs lacking the far C-terminal region were able to partially restore enzyme efficiency for shortened DNA oligomers in *in vitro* assays. Together, the data suggest that the far C-terminal region of protelomerase Tel-KO2 may interact with the distal regions of the DNA palindrome, the catalytic region, or both in a manner that influences protelomerase catalytic ability. Secondary structure predictions of the far C-terminal region of Tel-KO2 and other protelomerases estimate a region of flexibility but the varying length and limited sequence conservation of the C-terminal region make it difficult for automated sequence alignment programs to compare the far C-terminal regions. No further information has been reported about the structure or function of the

far C-terminal region of this or any protelomerase.

To better understand the role of the far C-terminal region in the functioning of phage protelomerases, we determined which portions of the isolated Tel-KO2 far C-terminal region are well-structured and we report the solution structure of this region. To verify the structure, we also solved the solution structure of a homologous region of a second phage protelomerase. We report that the structures form compact domains that adopt a similar fold and belong to the double stranded RNA binding domain (dsRBD)-like superfamily. We present an alignment of these domains with the far C-terminal regions of all known phage protelomerases, identifying putative homologous far C-terminal domains (farCTDs) and identify conserved residues. We propose by structural homology and limited sequence conservation that the protelomerase farCTD belongs to the small but growing group of proteins which recognize DNA in a sequence-specific way by inserting a three-stranded β-sheet in the DNA major groove.

Experimental procedures

DNA, plasmids, and cloning

DNA encoding the Tel-KO2 far C-terminal region was amplified from established vectors encoding full-length Tel-KO2 (cloning of vectors described in (*29*)). The length of the initial Tel-KO2 farCTD construct (residues 530-640 in the context of the full-length protelomerase) was selected to overlap the amino-terminal region of the protelomerase that had been studied using X-ray crystallography (*64*). Comparative ^{15}NHSQC and ^{1}H^{15}N-NOE (heteronuclear NOE) experiments combined with a standard suite of NMR assignment experiments (described below) were used to identify the structured portion of the Tel-KO2 farCTD. The length of the construct selected for

structure determination (residues 581-640) was predicted to have a short N-terminal unstructured region. The size of the protein fragment used to study the VHML protelomerase farCTD (Tel-VHML, residues 449-509) was selected to reflect the solved structure of the Tel-KO2 far CTD as determined by direct sequence and predicted secondary structure homology. The encoding DNA was amplified from PCR primers designed by Wai Mun Huang and synthesized as oligonucleotides by the University of Utah DNA/Peptide core facility.

The resulting DNA fragments encoding either Tel-KO2 or Tel-VHML farCTD were ligated into NdeI/BamHI sites of T7 promoter-driven pET16b vectors (Invitrogen) to create a template that encoded a fusion protein with an amino-terminal 6X histidine affinity tag followed by a thrombin protease site. By design, the purified proteins contained an additional four amino-terminal residues (G-S-H-M) resulting from the thrombin cleavage site. The expression constructs were verified by DNA sequencing at the University of Utah DNA Sequencing Core facility, and resulting plasmids were transformed into *E. coli* BL21 (DE3) Gold Competent cells (Novagen.)

Protein expression and purification

Tel-KO2 or Tel-VHML protelomerase farCTD fusion proteins were isolated from cells grown in two-liter quantities of M9 minimal media supplemented with 130 µg/ml ampicillin and either 2g/liter ^{13}C-glucose and 1g/liter $^{15}NH_4Cl$ or 1g/liter $^{15}NH_4Cl$ alone. Cells harboring the Tel-KO2 plasmid grew at 37°C to a cell density of approximately A[590] = 0.8. Protein expression followed induction by 0.7 mM (final concentration) isopropyl-1-thio-β-D-galactopyranoside (IPTG) for 12–16 hours at room temperature. Cells harboring the Tel-VHML farCTD fusion protein grew at room temperature until

cell density had reached A[600]=0.55. Protein expression was induced with 0.25mM IPTG for 16 hours at 18°C. In both cases, cells were harvested by centrifugation and stored frozen at -80°C.

Purification of Tel-KO2 and Tel-VHML protelomerase farCTDs was similar, with only minor exceptions. For the Tel-KO2 farCTD, the frozen pellet from one liter of culture was thawed and suspended in 20 ml of lysis buffer containing 25% (w/v) sucrose, 50 mM Tris–HCl (pH 8.0), supplemented with protease inhibitors (0.13 mM benzamidine and 0.6 mM phenylmethylsulfonyl fluoride (PMSF)), 25 mM β-mercaptoethanol (BME), and 1mg/ml lysozyme (Sigma). The resulting lysate suspension, containing 0.6 mg of cell pellet/ml, was incubated at 0°C for approximately 30 minutes. Following sonication (4 X 20 seconds with a micro-probe), the suspension was treated with RNase (150 µg/ml) and DNase (5 µg/ml in 5 mM Mg_2SO_4) at 0°C for one hour. The following were then added to reach the noted concentration before an overnight incubation at 0°C: Thesit (Boehringer Mannheim) 0.6% (v/v), NaCl (2 M), benzamidine (0.13 mM) and PMSF (0.625 mM.) The suspension was clarified by centrifugation (25,000 rpm and 4°C in an SW40 rotor) to remove cell debris.

Soluble protein was purified by nickel sepharose chromatography (Amersham Pharmacia) as follows: clarified lysate was applied (as two batches) to a 12–15 ml column of Ni-NTA agarose (Qiagen) equilibrated with a buffer consisting of 10% (v/v) glycerol, 50 mM Tris–HCl (pH 7.5), 10 mM BME, and 0.1 mM benzamidine. The protein-loaded column was then washed with 10 column-volumes of equilibration buffer which had been supplemented with 500 mM NaCl and 20 mM imidazole. The protein was then eluted from the column with equilibration buffer supplemented with 500 mM

NaCl and 800 mM imidazole. Following elution, fractions containing the protein were pooled and successively diluted and reconcentrated using 10,000 MWCO Amicon® Centricon® Centrifugal Filter Devices (Millipore) to remove excess NaCl and imidazole, reaching a final buffer concentration of 100mM NaCl and 50mM imidazole in 50mM sodium phosphate, pH 7.4. The 6x His tag was removed using biotinylated thrombin, and NaCl was added to reach 250µM. The cleavage reaction mixture was then purified first with a streptavidin column to remove the protease, and then a nickel column to remove the cleaved tag as well as any uncleaved protein. Recovery was nearly 100%, as indicated by Bradford assay (using Pentax BSA as standard.) Recovered protein was concentrated using a fresh 10,000 MWCO centricon (Millipore), with successive dilutions with NMR buffer followed by concentration to a final buffer composition of 25mM sodium phosphate, pH6.4, 10mM NaCl in 90% H_2O and 10% D_2O. Final protein yield, as indicated by Bradford assay, was 0.350mL of 1.0mM protein.

The purification of the Tel-VHML farCTD was similar, with the following differences: the concentration of protease inhibitors originally added to the lysis buffer were 0.15mM benzamidine, 0.75 mM PMSF; following nuclease digestion, protease inhibitors were supplemented to reach original concentration and the detergent and sodium chloride were added to reach final concentrations of 0.25% Thesit (Boehringer Mannheim) and 2.5M NaCl. Column purification differed as follows: following lysate centrifugation, the clarified supernatant was diluted with 50mM Tris, pH 7.4 to a total salt concentration of 1M NaCl, and imidazole was added to 10mM before loading onto a nickel sepharose column. The column was washed first with 50mM Tris, pH7.4, 0.5M NaCl then with a low concentration imidazole wash (as before, but containing also 20mM

imidazole) before elution with 0.8 M imidazole in 50mM sodium phosphate, pH 7.4 with 0.5M NaCl. Pooled fractions were concentrated using a 3000MWCO Amicon® Centricon® Centrifugal Filter Device (Millipore). Final protein yield, as indicated by Bradford assay, was 0.450mL of 1.0mM protein.

NMR spectroscopy: data collection and resonance assignments

NMR spectra were recorded on either a Varian Inova 500 MHz or 600 MHz NMR spectrometer equipped with a triple-resonance $^{1}H/^{13}C/^{15}N$ probe (cryogenic probe, 600MHz; room temperature probe, 500MHz) and z-axis pulsed-field gradients. To optimize resolution and minimize broadening, Tel-KO2 experiments were conducted at 30°C. Tel-VHML experiments were conducted at 25°C. NMR data were processed using FELIX 2004 software (Felix NMR, Inc., Accelrys, San Diego), and the SPARKY software program (T.D. Goddard and D. G. Kneller, University of California, San Francisco) was used to assign resonances using standard approaches with tools developed within this program. Assignment of main-chain atoms, β carbons, and β protons was accomplished from analysis of triple resonance experiments including CBCACONH(3D), HNCACB(3D), HBHACONH(3D), and HNHA(3D)(*66*). Aliphatic side-chain atoms beyond the β position were assigned from resonances in additional spectra, including CCONH(3D), HCCONH(3D), HCCH TOCSY(3D), and HCCH COSY(3D). Aromatic side-chain assignments were made using CBHD, CBHE, CHSQCTOCSY, and CTHMQCTOCSY(3D) experiments. Side-chain amide resonances were assigned from intraresidue NOEs and HNCACB and CBCACONH spectra. The backbone torsion angle restraints were derived from chemical shift information as evaluated by the TALOS software program (*67*). NOE connectivities were obtained from 3D [$^{1}H,^{13}C,^{1}H$] NOESY

and 3D [^{1}H,^{15}N,^{1}H] NOESY experiments.

Structure calculations

Following resonance assignment of backbone and side-chain atoms, the tools in the SPARKY software program were used to identify NOE correlations and intensities from relevant NOESY spectra. Using the resulting NOE correlation data and torsion angle restraints generated as described above, NOE assignments and initial structures calculations were generated using the automated NOE assignment and torsion angle dynamics capabilities of the CYANA software program (version 2.1) (*68*). Briefly, using the criteria of chemical shift agreement, network anchoring, and consistency with an initial structure, a total of 100 randomized conformers were "folded" into 3D structures by introducing NOE constraints in a step-wise manner. Each conformer underwent seven cycles, with 10,000 steps of torsion angle dynamics per cycle. The lowest energy CYANA calculated structured were further refined in XPLOR-NIH using a simulated annealing protocol that utilizes the hydrogen bond restraints and assignments determined by the CYANA program. PROCHECK-NMR (*69*) and the validation programs supplied at the PDB deposition site (http://deposit.pdb.org/adit/) were used to substantiate the structures. PDB files will be deposited into the PDB database prior to publication. Structures were visualized and figures created using PyMOL (DeLano Scientific) (*70*) and MOLMOL (*71*).

Use of sequence/structure databases and sequence alignment

Structural homology searches and structural alignments were accomplished using the Dali and DaliLite programs (*65*). Alignment of complete protelomerase sequences

was accomplished using CLUSTAL W (*72-75*). Coordinates of the dsRBDs of the DNA-binding domains of the integrase protein from conjugative transposon Tn916 (*76*), I-PpoI homing endonuclease (*77*), AtERF1 protein (*78*), and lambda integrase(*60*), and of the dsRBDs of *Xenopus laevis* XlrbpA dsRBD2 (*79*), *Drosophila* staufen dsRBD3 (*80*), *Aquifex aeolicus* RNase III (*81*) used in structure comparison calculations were obtained from the Protein Data Bank (PDB).

Results and discussion

Identification of the structured portion of theTel-KO2 far C-terminal region

NMR was used to evaluate the suitability of the construct for structure determination, and to design the size and conditions of the construct for study. The length of the initial Tel-KO2 farCTD construct (residues 530-640 in the context of the full-length protelomerase) was selected to include a short region of the amino-terminal region of the protelomerase that had been studied using X-ray crystallography (*64*). ^{15}NHSQC studies of the initial farCTD construct yielded mixed results. A small number of well-resolved, well-distributed peaks were dominated by a central region of strong but poorly resolved peaks (see Figure 2.1). ^{15}NHSQC profiles with this type of pattern can indicate that a protein is not well-folded or that parts of the protein are unstructured, since the comparative ^{15}N and ^{1}H frequencies of backbone amides of individual amino acids and unfolded peptides are not significantly different. It is the different magnetic environments created by a folded protein which results in altered chemical shifts for individual amide resonances. Poorly resolved peaks can also indicate that multiple amide bond vectors are in very similar environments. The Tel-KO2 far C-terminal region has a

Figure 2.1 Analysis of the full far C-terminal domain of TelKO2 to determine suitability of construct for spectroscopic study.

A) Sequence profile showing the unusually high number of negatively charged residues.

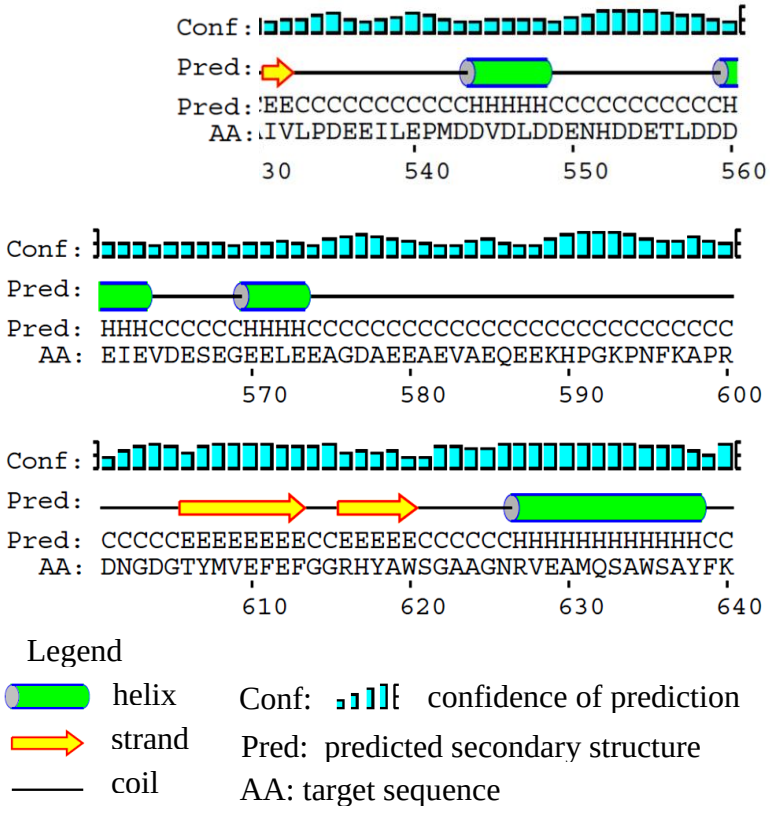
B) Secondary structure prediction as generated by the PSIPRED program (*82, 83*). The structure predicts a number of helices in the N-terminal half of the peptide, but these are reported with a low level of prediction confidence.

C) ^{15}NHSQC of the full far C-terminal domain. Although a number of peaks are well-dispersed and well-resolved, the central portion of the spectrum (bordered by a dashed box) contains a group of peaks with poor dispersion and resolution.

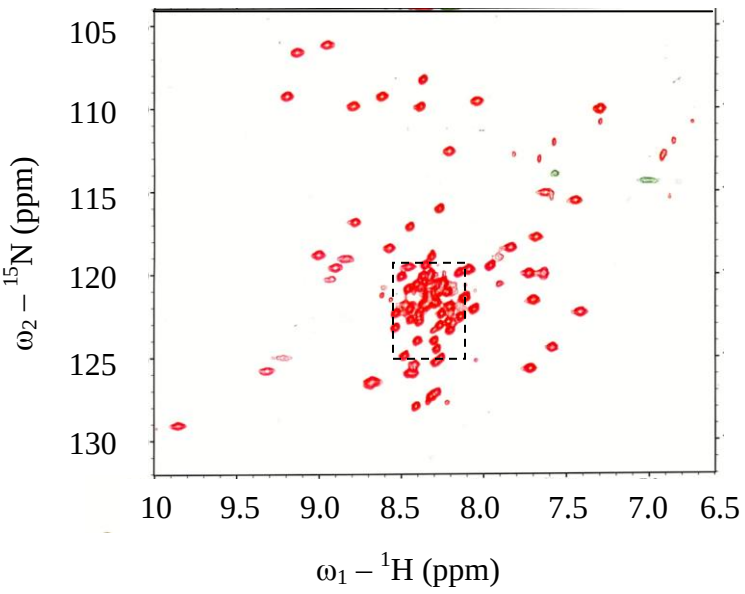
A.

Residue	Count
A Alanine	11
R Arginine	3
N Asparagine	4
D Aspartate	15
C Cysteine	0
Q Glutamine	2
E Glutamate	22
G Glycine	10
H Histidine	4
I Isoleucine	2
L Leucine	5
K Lysine	4
M Methionine	4
F Phenylalanine	4
P Proline	5
S Serine	5
T Threonine	2
Y Tyrosine	3
V Valine	6
W Tryptophan	2

B.



C.



unique compositional profile, being composed of a large number of negatively charged residues. Because of proportionately large number of similar amino acids (the composition of the protein construct is approximately 20% glutamate and 14% aspartate by residue type) and because the secondary structure analysis predicted a number of α-helices (see Figure 2.1), it was unclear based on ^{15}NHSQC studies alone whether the poorly resolved central peaks were indicative of a partly or poorly folded protein, or whether the region of low resolution was a result of the unique construct profile. HNNOE experiments were used to indicate whether the construct was likely to be well-folded. The ^{1}H-^{15}N heteronuclear NOE (HNNOE) experiment can be used as a probe in determining the structural state of a protein. Alone, HNNOE data is not sufficient to completely characterize molecular dynamics, but provides a qualitative indication of the mobility of individual amide bond vectors. As such, it can provide a useful assessment for which residues within a polypeptide are structured and which are not (*84*, *85*). HNNOE experiments indicated that while the central, largely unresolved grouping of the ^{15}NHSQC resonances were from residues undergoing rapid motions characteristic of an unstructured polypeptide, the resolved frequencies had values that indicated that the residues were relatively immobile, characteristic of residues within a well-folded structure. These results were supported by the predicted secondary structure calculations and by initial backbone assignments (*vide infra*.)

To determine whether the folded portions localized to a distinct region within the construct, a ^{15}N^{13}C-labelled protein was purified, and resonances were assigned using a variety of standard triple-resonance experiments. Unstructured protein regions typically yield very strong peaks, and these peaks dominated the spectra of all NMR experiments.

Thus, although many of the resolved peaks on the ^{15}NHSQC could be assigned, the obfuscation of the peaks within triple-resonance spectra by the overpowering resonances from the unstructured protein regions prevented a complete structure determination of this construct. The analysis did, however, result in an identification and mapping of the folded regions of the protein, which were localized to the extreme carboxy-terminal side of the far C-terminal construct. Figure 2.2 shows the comparative ^{15}NHSCQ and ^{15}N-HNNOE profiles and the construct residues that were assigned.

Based on identification of the folded portion of the far C-terminal region along with the secondary structure prediction analysis, constructs with increasing amino-terminal truncation were designed. The encoded proteins were expressed and purified, and their ^{15}NHSQC profiles were compared to determine the optimal construct length for structure determination. The construct selected for structure determination, consisting of residues 581-640, was still predicted to contain an N-terminal unstructured region, but was selected nevertheless due to its better purification efficiency. The resulting purified protein also contained an additional four amino-terminal residues (G-S-H-M), resulting from the thrombin cleavage site engineered to facilitate purification.

Analysis of the ^{15}NHSQC of the resulting construct showed fewer resonances than expected, as well as a range of peak widths with broadening for some individual peaks. Sedimentation equilibrium studies of the first KO2 protelomerase farCTD construct (residues 530-640) had shown a possible multimeric equilibrium (data not shown) so a series of solution and experimental conditions were evaluated by NMR to define the optimal conditions for structure determination. Sample conditions were varied to evaluate the effects of protein concentration, pH, and salt and buffer concentrations.

Figure 2.2 Determination of structured regions of Tel-KO2 (530-640).

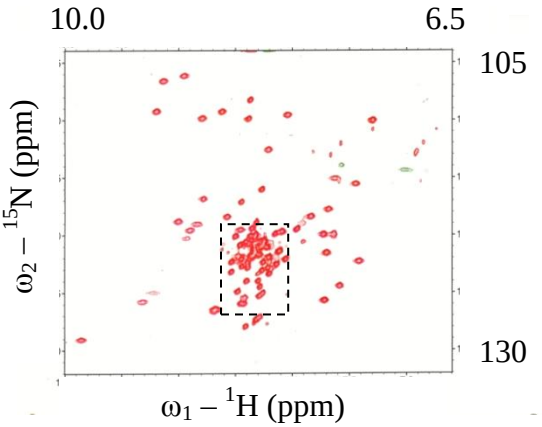
A) ^{15}NHSQC of the full far C-terminal domain. Although a number of peaks are well-dispersed and well-resolved, the central portion of the spectrum (bordered by a dashed box) contains a group of peaks with poor dispersion and resolution.

B) ^{15}N-HNNOE of full-length C-terminal region of Tel-KO2 (530-640). Red corresponds to positive peaks, and green corresponds to negative peaks. The ^{15}N-HNNOE indicates that the majority of the peaks in the central region of the spectrum belong to residues with a molecular motion similar to individual amino acids, and are most likely contained in an unstructured region of the peptide, while the well-dispersed peaks belong to residues that are in a structured region.

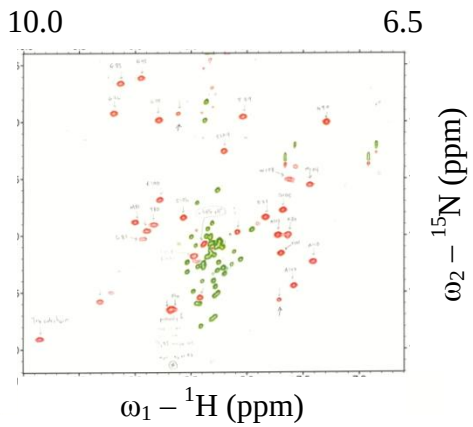
C) Sequence of TelKO2 (530-640) indicating residues for which amide resonances were assigned using a standard suite of triple resonance experiments. All residues with assigned amides (indicated by bold, underlined text) are located in the far C-terminal end of the peptide.

D) ^{15}NHSQC of the selected N-terminally truncated construct. The majority of the resonances representing amides in unstructured regions have been removed, while the amides from structured regions are all present.

A.



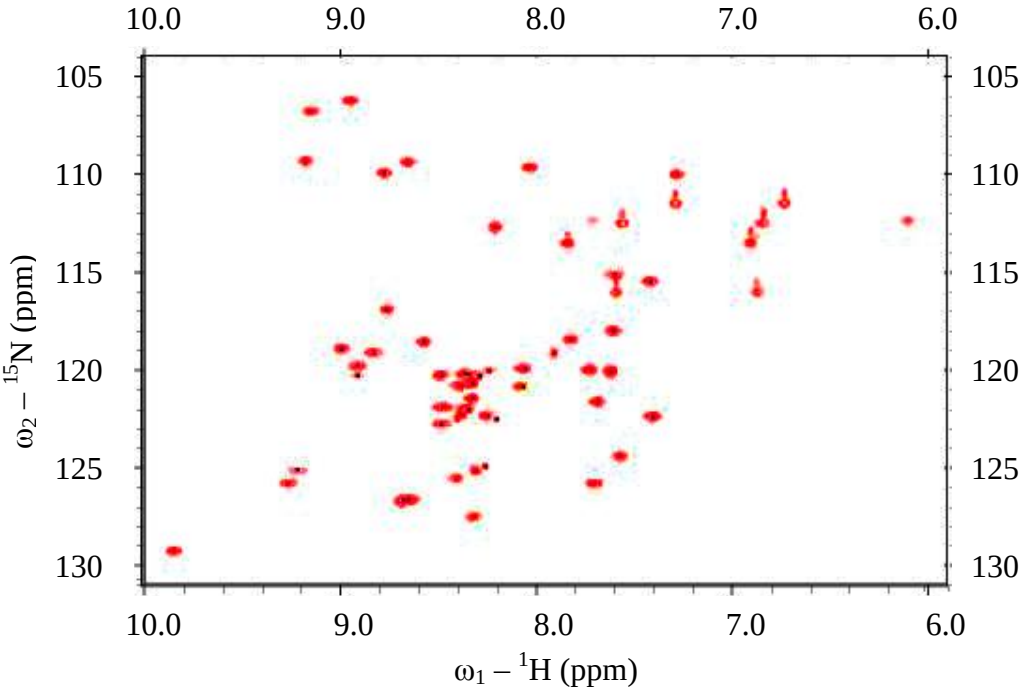
B.



C.

MGSSHHHHHHSSGLVPRGSHMVLPDEEILEPMDDVDLDDENHDDETLDDDEIEV
DESEGEELEEAGDAEEAEVAEQEEKHPGKPNFKAPRDN**GDGTYM**VEFEF**GGR**HY
AWSGA**AGNRV**E**AMQSAWSA**YFK

D.



The sample temperature and NMR field strength (500 and 600 MHz) were also varied since these affect spectral appearance in the presence of exchange. While the overall pattern of the ^{15}NHSQC was maintained, different conditions resulted in a surprising variability of the resolution of peaks in different regions of the spectrum. Because it was believed that higher temperatures and lower protein concentration would most likely shift the monomer/multimer equilibrium toward the monomeric form, conditions were selected which mimicked the ^{15}NHSQC pattern changes produced under these relative conditions. Lower experimental temperatures are usually better for sample stability, and can also protect the sample from rapid solvent exchange at exposed regions, which would interfere with the detection of amide resonances. Additionally, higher protein concentrations generally facilitate structure analysis because of better NMR signal to noise ratio. Conditions were selected which created a pattern that mimicked that observed at higher temperatures and lower concentrations without resorting to those extremes for experimental conditions.

Although the majority of the resonances were assigned and the structure determined, several expected resonances were unaccounted for (see Table 2.1). These included several resonances that were indeed present, but simply overwhelmed and obscured by the strong signals issuing from unfolded region of the peptide, as well as resonances that were presumably weakened or absent due to excessive broadening. Among the former were overlapping signals from residues in the unstructured region of the protein, including the HB* and HG* resonances of residues E582, E585, and E587. Among the latter were several resonances within a compact region of the protein (residues 616-620.) Loss of signal in this region is likely caused by local dynamics or

Table 2.1:

Tel-KO2 farCTD581-640 missing or unassigned resonances

Residue	Missing resonance[a]	Note
GLU 582	1HB/2HB, 1HG/2HG	In unstructured N-term tail
GLU 585	1HB/2HB, 1HG/2HG	In unstructured N-term tail
GLU 587	1HB/2HB, 1HG/2HG	In unstructured N-term tail
LYS 589	1HG, 1HE/2HE	In unstructured N-term tail
LYS 593	1HG/2HG, 1HE/2HE	Beginning of structured region
LYS 597	1HB/2HB	
PHE 611	H, HZ	
PHE 613	HZ	
ARG 616	HA, 1HB/2HB, 1HG/2HG, 1HD/2HD	Region of third β-strand
HIS 617	H, HA, HD2	Region of third β-strand
TYR 618	H, HA, HE1/HE2	Region of third β-strand
ALA 619	HA	Region of third β-strand
TRP 620	H, HD1	Region of third β-strand
GLY 622	H	Region of third β-strand
PHE 639	HZ	

29 missing chemical shift assignments. Assignment completeness 90.5%.

Atom description follows PDB nomenclature.

intermediate exchange rather than solvent exchange, especially given the high proportion of missing HA resonances (four out of five residues in this region) which are not typically subject to solvent exchange-based signal attenuation. Given previous data suggesting a possible multimeric equilibrium of the far C-terminal region, it may be that this portion of the protein is the site of multimer interaction. Additionally, the structure shows that this region contains an unusually high number of phenylalanines and tyrosines, which may have contributed to signal attenuation due to ring-flipping motions about the Cβ-Cγ bond axis.

Several amide resonances within the 616-620 region were absent from the ^{15}NHSQC. Traditional molecular graphics programs such as PyMOL (70) and MOLMOL (71) use the Kabsch-Sanders algorithm (86) to determine secondary structure elements, which essentially looks for specific hydrogen bonding patterns involving amide resonances. Due to the missing amide resonances in this region, secondary structure was not determined in these programs. However, analysis of solved structures has lead to the observation that the existence of secondary structure leads to a characteristic observable shift of Cα, Cβ and Cγ resonances compared to those observed for residues in random coil conformation (*87*). The values of these shifts indicate a propensity for regions of the structures to populate extended or α-helical conformations. Typically, Cα shifts are better predictors of alpha helical regions, while Cβ and Cγ shifts are better indicators of β-strand or extended conformation. A consensus of the assigned Cα, Cβ, and Cγ shifts for the residues in this region show that this region, although underdetermined, is most likely composed of β-strand/extended region (data not shown.) In the case of the residues with missing ^{15}NHSQC amide resonances, other intraresidue frequencies were observable

and assigned from other data sets, such as the side-chain frequencies from the aromatic or aliphatic region of CHSQC experiments. In no case were all resonances for a given residue missing or unobservable. Nevertheless, due to the missing resonances, the structure is unavoidably somewhat underdetermined, particularly in the 616-620 region.

To verify the structure and to identify themes of conservation among phage protelomerases, a homologous phage protelomerase far C-terminal region was selected for structural study. A sequence comparison of known phage protelomerases revealed that the protelomerase of phage VHML (Vibrio Harvey Myovirus Like) of the free-living aquatic *V. harveyi* proteobacterium was the most distant relative of TelKO2 within the phage clade, so this protelomerase was selected. Sequence conservation of the two regions was low (16.4%, with only 9.8% identity and 6.6% homology) as determined by BLAST comparisons (*88, 89*). Secondary structure prediction for the two domains was similar, however, with two contiguous β-strands followed by a carboxy-proximal α-helix. The VHML far C-terminal region contains fewer phenylalanines and tyrosines, and early tests indicated that the VHML construct designed to be homologous to the Tel-KO2 far C-terminal region did not appear to have an unstructured region or suffer from issues of peak broadening or signal attenuation caused by intermediate exchange or other local dynamics, making it a good candidate for structure determination.

Structure of Tel-KO2 and Tel-VHML far C-terminal domains

The 2D homonuclear and 3D 13C- and 15N-edited NOESY experiments provided 625 and 1098 conformationally restrictive NOE distance restraints for Tel-KO2 and Tel-VHML far C-terminal regions, respectively. The NMR structure ensembles have a root mean square deviation of 0.51Å (Tel-KO2 farCTD) and 0.43Å (Tel-VHML farCTD) for

backbone atoms in structured regions. All prolines were in the trans conformation, as confirmed by 13Cβ and 13Cγ chemical shifts typical of trans-Prolines as part of the CYANA program (version 2.1) (68). A full summary of structural statistics is given in Table 2.2. Both the Tel-KO2 and Tel-VHML far C-terminal regions are well within acceptable ranges in all points of measure: on the basis of agreement between individual structures within ensembles, good geometries, low residual target function energies, and lack of NOE violations, the structures of both constructs are well-determined. However, the Tel-VHML far C-terminal region, which was not hampered by the broadening issues of local dynamics or intermediate exchange as the Tel-KO2 structure was, is better defined, and it verifies and improves on the Tel-KO2 structure. Compared with the 29 missing or unassigned resonances for the Tel-KO2 construct, the Tel-VHML construct has only two unassigned resonances (tyrosine 57 HD1/HD2 and methionine 64 1HE/2HE/3HE) The Tel-VHML structure has more assigned and fewer unassigned NOESY peaks than its Tel-KO2 counterpart, a lower average RMSD to mean coordinates, and more favorable Ramachandran statistics. This is most likely due to the absence of the N-terminal unstructured region in the Tel-VHML construct and better determination of the C-terminal region due to a lack of exchange-based broadened NMR signals that confounded the Tel-KO2 study. As with the Tel-KO2 construct, a consensus of Tel-VHML Cα, Cβ, and Cγ shifts predicts three β-strands before the C-terminal α-helix. For the Tel-VHML construct, however, and all expected peaks were accounted for in the region of the third predicted β-strand. The resulting structure confirms the presence of the third β-strand, and by homology, verifies that the Tel-KO2 construct also contains a three stranded β-sheet. The resulting Tel-VHML and Tel-KO2 structures thus

Table 2.2

Comparison of structural statistics of Tel-KO2 and Tel-VHML farCTD structures

Parameter	value	
Structure	Tel-VHML	Tel-KO2
Residues	449-509	581-640
Constraints used for structure calculation		
Short-range NOEs |i-j|≤1	1554	934
Medium-range NOEs 1<|i-j|<5	208	137
Long-range NOEs |i-j|≥5	416	288
Dihedral angle constraints[a]	90	82
Hydrogen bonds[b]	24	14
Average Deviations from idealized geometry (RMSD(dev)) from XPLOR[c]		
Bonds (Å)	.002(.000)	.003(.000)
Angles (˚)	.435(.003)	.441(.004)
Impropers (˚)	.278(.012)	.259(.009)
Average		
CYANA assignments/constraint (CYANA)	1	1
CYANA target function value, $Å^2$	0.29	0.02
Average RMSD to mean coordinates[d]:	res: 452-509	res: 593-640
Avg. backbone to mean (Å)	0.43	0.51
Avg. heavy atom to mean (Å)	1	1.03
RMSD from ideal geometry[z]		
Bond lengths (Å)	0.012	0.012
Bond angles (˚)	1.5	1.4
Ramachandran analysis for structured regions[e]:	procheck	procheck
Most favored regions (%)	96.6	92.3
Allowed regions (%)	3.4	7.7
Generously allowed regions (%)	0	0
Disallowed regions (%)	0	0

[a]Dihedral angle constraints generated using TALOS (*67*).
[b]Determined from an agreement of structures resulting from initial CYANA calculations (*68*)
[c]The statistics (average (SD)) calculated for the bundle of the 20 best-energy conformers.[c]
[d]Average of the 20 best-energy structures calculated using the program XPLOR using upper limit distance constraints and H bonds from CYANA and dihedral angle constraints from TALOS
[e]Determined using PROCHECK-NMR (*69*)

adopt the same fold, a compact domain with a βββα topology in which the three β-strands form an anti-parallel β-sheet which packs against and folds around the C-terminal alpha helix. The structures have been called the far C-terminal domains (farCTDs.) The solved Tel-KO2 and Tel-VHML farCTD structures are each represented by an ensemble of the 20 lowest energy conformers in Figure 2.3).

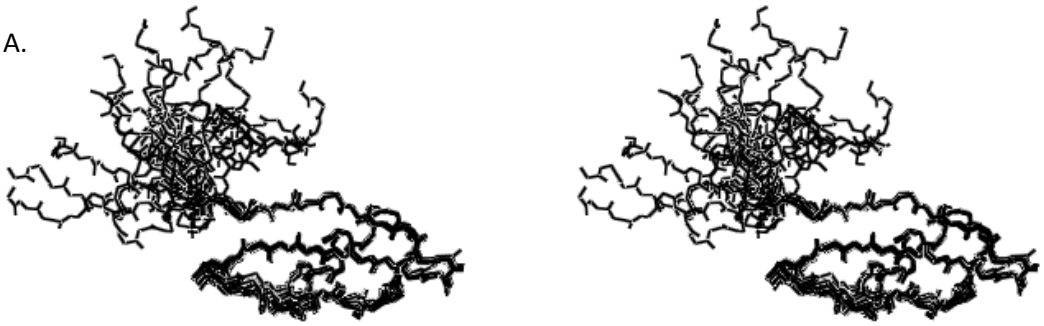
Although the farCTDs from the KO2 and VHML protelomerases belong to the same fold, they are nonidentical. The two domains share only limited sequence identity and differ with respect to the length and curvature of the helix. The C-terminal helix in the Tel-KO2 farCTD consists of 13 residues, while the helix in the Tel-VHML farCTD consists of 18 residues. The N-terminal region of the Tel-KO2 farCTD appears to be completely unstructured, while the N-terminal region of the Tel-VHML farCTD is structured and somewhat extended, but does not appear to conform to common secondary structural elements. The relevance of this extended region has yet to be established.

The hydrophobic cores of the two proteins are also different. As shown in Figure 2.4, there are more aromatic residues in the Tel-KO2 farCTD core, especially near the C-terminal side of the helix. The aromatic residues appear to be stacked and angled in a way that would enhance the stability of the folded structure. In particular, the W635/F639 and F596/F610 pairs are stacked in a nearly perfectly planar conformation in all five of the best-energy structures, with average distances of slightly less than 5Å between their centroids. F613 is approximately perpendicular to the W635/F639 pair, which would also provide additional stability to the stacked pair (see Figure 2.4.) In the Tel-VHML farCTD structure, only three residues (W455/W483/Y502) are involved in the aromatic network on the C-terminal edge of the alpha helix, and are arranged in

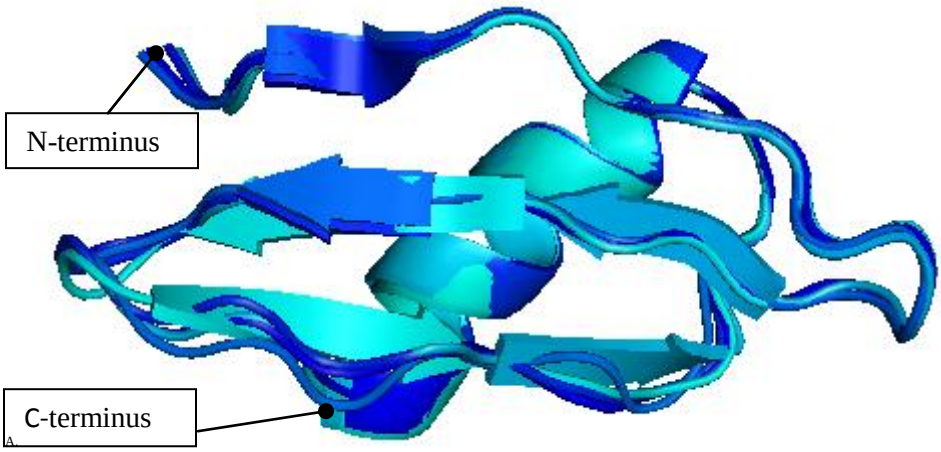
Figure 2.3 NMR ensembles of Tel-KO2 and Tel-VHML farCTD structures

NMR ensembles of the 20 lowest-energy structures of Tel-KO2 farCTD (A) and Tel-VHML farCTD (C) shown as backbone diagram in stereo. NMR ensembles showing ribbon representation of the five lowest energy structures of the Tel-KO2 farCTD ((B), structured region only, residues 593-640) and Tel-VHML farCTD (D).

A.



B.



C.

D.

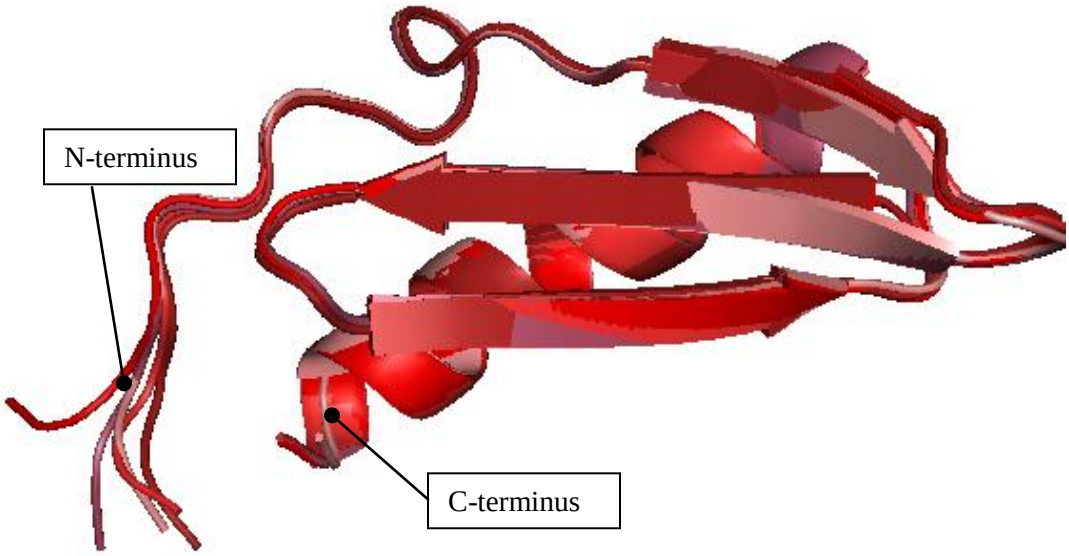


Figure 2.4 Comparison of the hydrophobic cores of the Tel-VHML and Tel-KO2 farCTDs.

A) Ribbon diagrams of the top five structures of the Tel-VHML farCTDs. The hydrophobic residues which make up the interior are shown as sticks. Aliphatic residues are colored wheat while aromatic residues are shown in white.

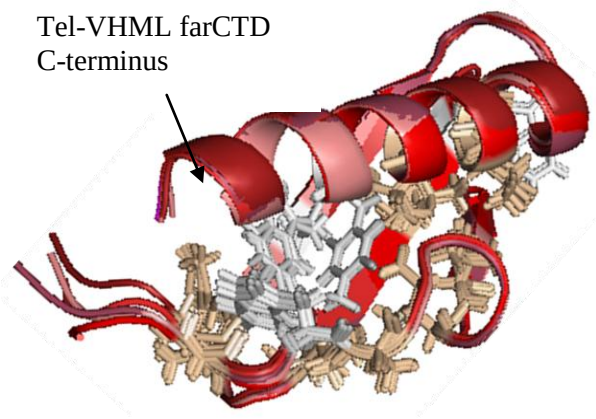
B) The aromatic stacking network of the Tel-VHMLfarCTD. The ten best energy structures are shown, and are represented as ribbon diagrams of the backbone. The side-chains of relevant residues are shown as sticks. Residues W455 (yellow), W483 (green), and Y502 (blue) are oriented in an orthogonal conformation to each other.

C) Ribbon diagrams of the top five structures of the Tel-KO2 farCTD. The hydrophobic residues which make up the interior are shown as sticks. Aliphatic residues are colored pale blue, while aromatic residues are shown in white.

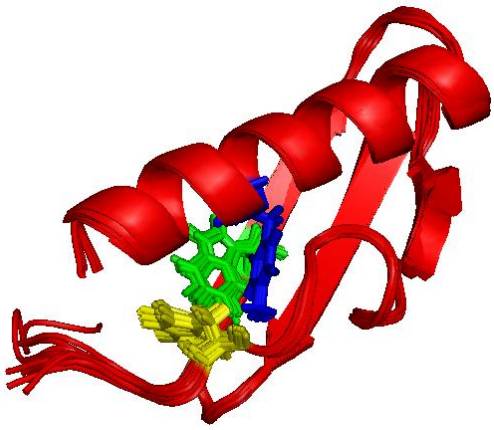
D) The aromatic stacking network of the Tel-KO2 farCTD. The ten best energy structures are shown, and are represented as ribbon diagrams of the backbone. The side-chains of relevant residues are shown as sticks. The W635/F639 stacking pair is shown in red. F613, which is orthogonal to this pair, is shown in yellow. The F596/F611 pair is shown in green.

The interior of the Tel-KO2 farCTD contains more aromatic residues, with an extensive network of stacking interactions both planar and perpendicular.

A.

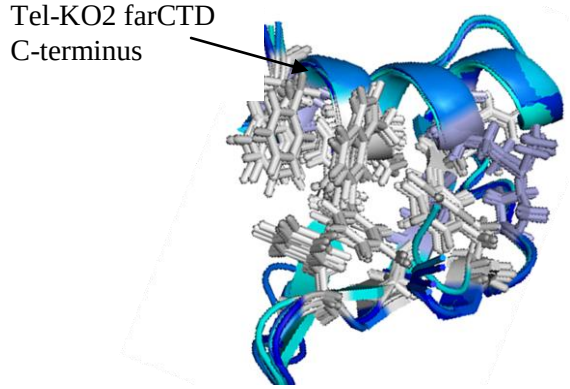


B.

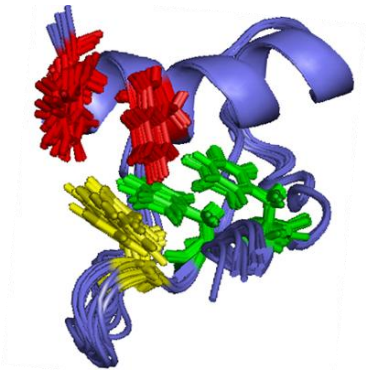


W455 (yellow), W483 (green), and Y502 (blue) in orthogonal conformation

C.



D.



W635/F639 stacking pair (red) perpendicular to F613 (yellow). F596/F611 stacking pair (green)

orthogonal conformations to each other. The Tel-VHML farCTD appears to be additionally stabilized by the interaction of the C-terminus of the longer helix with the somewhat extended N-terminal region. The homologous Tel-KO2 farCTD N-terminal region lacks observable structure.

Because of the very limited sequence conservation, differing secondary structure predictions, and the varying length of unstructured linker regions between the farCTDs and their N-terminal protelomerase regions, automated sequence alignment programs have difficulty in recognizing the conserved domain as being homologous among protelomerases. Most prediction programs predict two, rather than three β-strands, some identifying strands β2β3 (as with Tel-KO2 farCTD) and some identifying strands β1β2 (as with Tel-VHML farCTD) which further complicates alignment (see Appendix A for a comparison of Tel-KO2 and Tel-VHML secondary structure prediction and comparative Cα shifts.) Having solved the Tel-VHML and Tel-KO2 farCTD structures, we used the DaliLite program to overlay the structures, and identified several conserved and similar residues (*65*). The hydrophobic interior of the two proteins was especially well-conserved, resulting in a positionally homologous network of hydrophobic residues (both aromatic and aliphatic) that is shown in Figure 2.5.

Aside from the hydrophobic core, the structural overlay revealed additional regions of sequence similarity which appear to be conserved among the other phage protelomerases (see Figure 2.6). The loop between the β1β2 strands, for instance, consists of a GD-rich region which is highly conserved among the other putative protelomerase farCTDs. The loop between the β2β3 strands is shorter than the β1β2 loop, and contains a conserved glycine. The short loop between the β3 strand and the

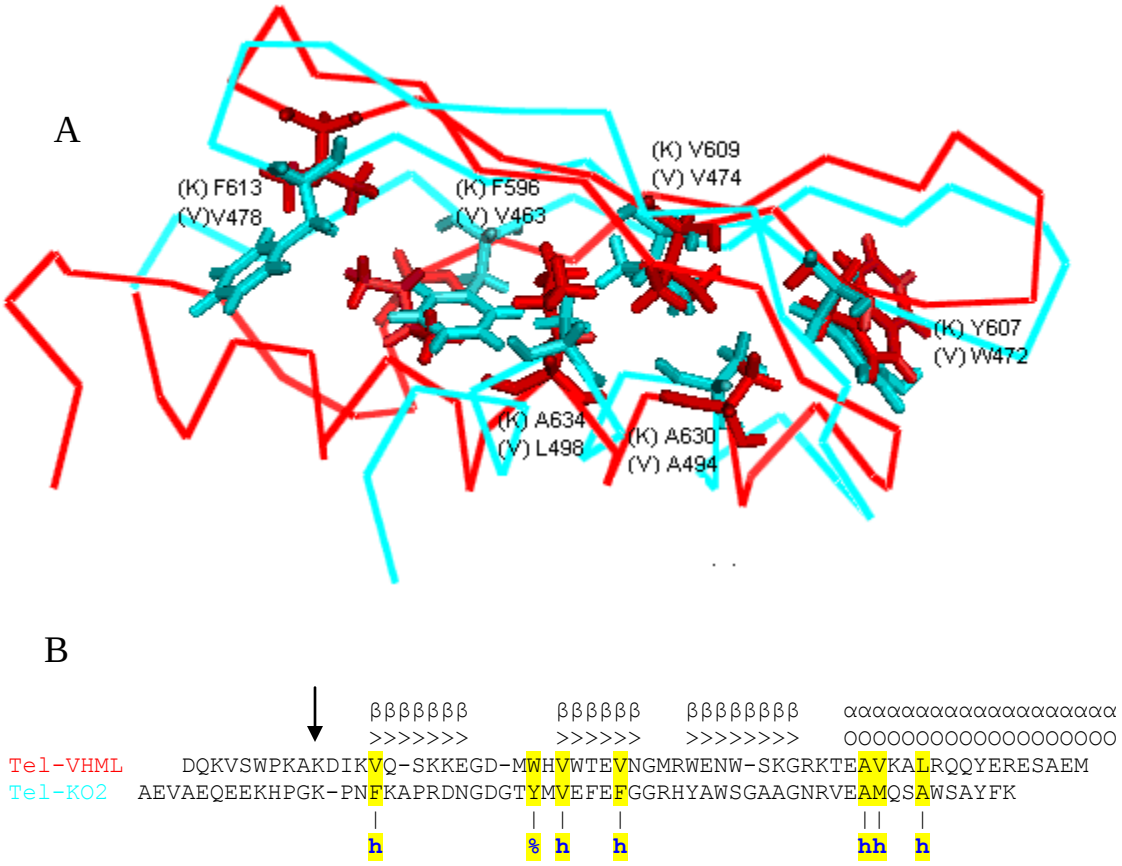


Figure 2.5 Conserved network of hydrophobic residues in the interior of phage protelomerase farCTDs.

A) Cartoon showing an overlay of the solved farCTDs. The Tel-KO2 farCTD is shown in cyan, and the Tel-VHML farCTD is shown in red. Hydrophobic residues that are conserved by type and location are represented as sticks. Figure text denotes the comparative residues with (V) indicating that a residue belongs to Tel-VHML and (K) indicating a Tel-KO2 residue. The two structures have an RMSD of 2.1 in comparative regions as calculated by a DaliLite structure comparison (*65*).

B) A sequence alignment of the two domains indicates the location of the similarly conserved residues. A consensus sequence is shown beneath the alignment, in which (%) indicates an aromatic residue and (h) represents a hydrophobic residue (either aromatic or aliphatic). The symbols above the sequence alignment indicate the location of the secondary structure elements identified by structure determination (β indicates β-strand or extended conformation, while α indicates α-helix). An arrow indicates the first residue within the sequences for which DaliLite structure comparisons were calculated.

Figure 2.6 Conserved residues of the phage far C-terminal domain.

A) and B) Structural overlay showing the externally-facing residues which are conserved according to position and type. The overlaid backbones of Tel-VHML and Tel-KO2 are shown in red and cyan, respectively. Views A) and B) are shown with a 90° vertical rotation to each other. Text indicates the identity and sequence number of residues in the context of the full-length protelomerase, with (V) indicating the residue belongs to Tel-VHML and (K) indicating the residue belongs to Tel-KO2.

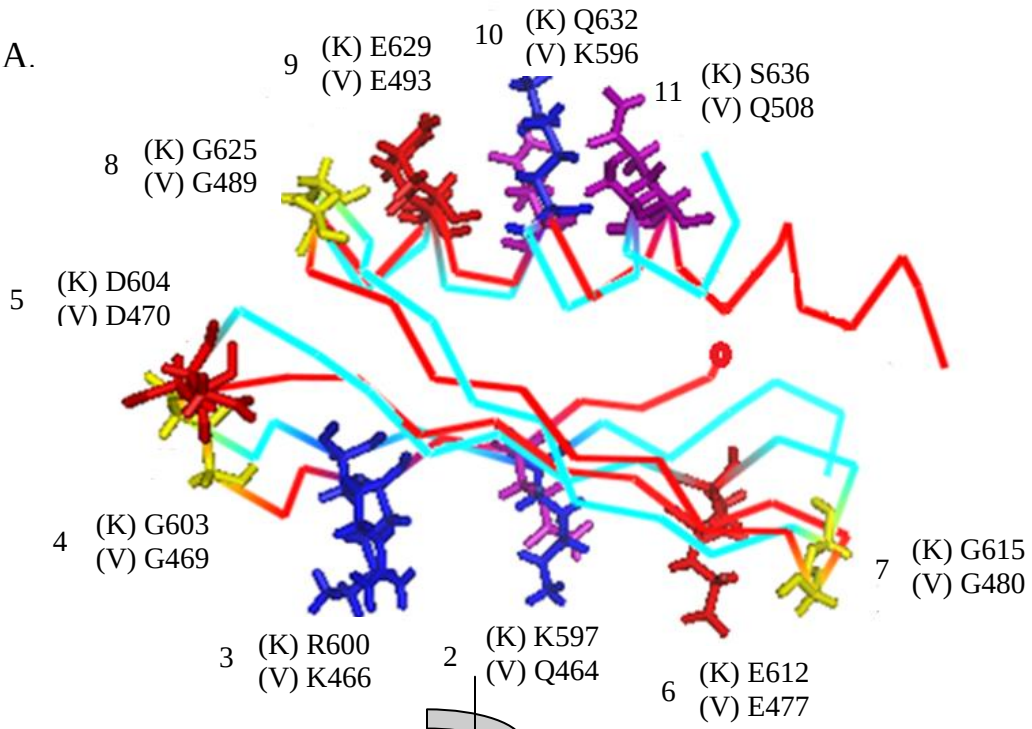
C) An alignment of all known and putative phage protelomerase farCTDs as compared to the solved Tel-KO2 and Tel-VHML structures. The numbers to the right of each aligned sequence indicate the C-most residue of that sequence in context of the full-length protelomerases. Residues which appear to be conserved through a majority of the protelomerases are shown in red and underlined, and are numbered in both the aligned sequences and the displayed overlays (1-11). Underneath the alignment is displayed: first, an indication of the degree of conservation among the domains; next, a consensus sequence; and finally, the location of the conserved residues within the aligned Tel-VHML and Tel-KO2 structures. Symbols indicating consensus sequence, conservation, and structural position are as follows:

<u>For the consensus sequence:</u>
h: hydrophobic (either aliphatic or aromatic)
u: small (Gly, Ala)
%: aromatic
^: polar/charged
±: charged
+: positively charged
‘: negatively charged

<u>For the degree of conservation:</u>
* total conservation
: high conservation
. moderate conservation

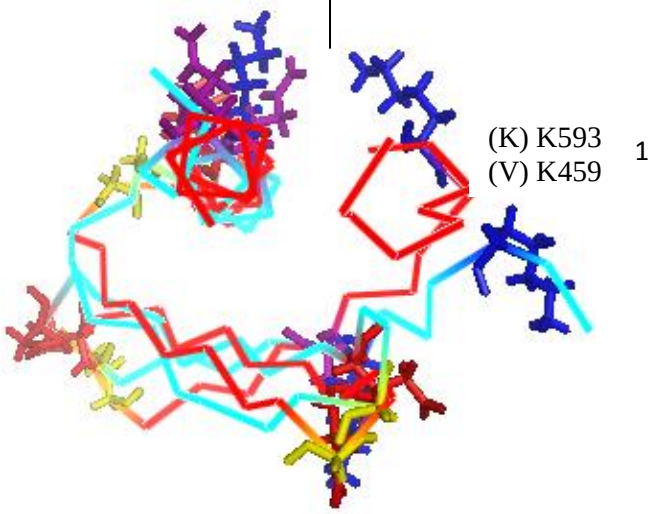
<u>For the location of the conserved residues:</u>
i: the side-chain extends into the interior of the protein
α: the side-chain extends externally from the face of alpha helix
e: the side-chain extends externally from the face of the beta sheet
l: the residue resides in a loop region

A.



90°

B.



C.

```
                1    2   3  45       6  7          8    9  10  11
      ---------------ββββββββ------ββββββββ----ββββββ-----αααααααααααααααααααα
VHML  QEEDQKVSWPKAKDIKVQ--SKKEGD-MWHVWTEVNGMRWEN-WS-KGR-KTEAVKALRQQYERESAEM  509
KO2   AEVAEQEEKHPGK-PNFKA-PRDNGDGTYMVEFEFGGRHYAW-SGAAGN-RVEAMQSAWSAYFK      640
N15   EEGPEEHQPTALK-PVFKP-AKNNGDGTYKIEFEYDGKHYAW-SGP-ADSPMAAMRSAWETYYS       631
PY54  SDNASDEDKPEDK-PRFAAPIRRSED-SWLIKFEFAGKQYSW-EG-NAESVIDAMKQAWTENME      628
VP882VPAAEKQPKKAQK-PRLVAH-QV-DDEHWEAWALVEGEEVAR-VKIKG-TRVEAMTAAWEASQKALDD  538
Halm  VAAAVPKEVAEAK-PRLNAHPQ--GDGRWVGVASINGVEVAR-VGNQAG-RIEAMKAAYKAAGGR     520
VCampVVETKPKDETVIK-PKMKGH-KE-DDGTWLVDVTIEDKSWQISVGKEPKNVMDAFKLAWNEFvyrkalpe……
Conservation:     * :....  :  .*. : . . . :  .       :   ..*:. ::.  .
Consensus:        K-P^h^u--+^-GDG-%-h-h-h-G--%-------u^--h'Ah+-A%^-%-------
Location:           i^ie   e  lll i i i i l  i       α    αiiα i α
```

alpha helix is also composed of small residues. The external face of the alpha helix shows considerable conservation according to residue type. When viewed through the helix, a series of charged and polar residues align in a surprisingly linear fashion, especially given the variability in the curvature of the helix. Judging from the aligned putative protelomerase farCTDs, this pattern is nearly universally conserved, and includes, starting from the N-terminal base of the helix, a small residue, a negatively charged residue, and two successive polar/charged residues. Perhaps equally notable as the conservation within the loops and on the face of the alpha helix is the complete lack of conservation on the external face of the β-sheet. There is, however, notable conservation at the beginning and end of strand β1, which consists of basic or polar residues, as well as a universally conserved lysine just prior to the start of the same strand.

The calculated electrostatic surface potentials of the Tel-KO2 and Tel-VHML structures are very similar, as calculated by the Adaptive Poisson-Boltzmann Solver (APBS) add-in within the PyMOL program (*70, 90*). Both structures display a positive surface along the β1/helix face, and a negative surface along the β2/helix face. The calculated surface potentials of these faces are shown for both proteins in Figure 2.7.

With the structure of the farCTD solved and regions of conservation identified by structural and sequential alignment, an understanding of the functionality of this domain will be aided by the identification of its binding partner. Initial studies aimed at accomplishing this directive both by surface plasmon resonance and NMR were inconclusive. However, a careful comparative analysis of the farCTD with the sequence and structure of proteins with similar domains gives a good indication of the probable

Figure 2.7 Electrostatic Surfaces of the Tel-KO2 and Tel-VHML farCTDs show similarly charged surfaces along homologous faces..

Red indicates a negative charge, while blue indicates a postitive charge. White indicates a neutral charge. The surfaces are generated at partial transparency to show the secondary structure cartoon beneath. The cartoon is colored in a spectrum with the N-terminus beginning in blue and the C-terminus finishing in red.

A. Tel-KO2 farCTD oriented to show the most negatively charged surface, along the β3/helix face.

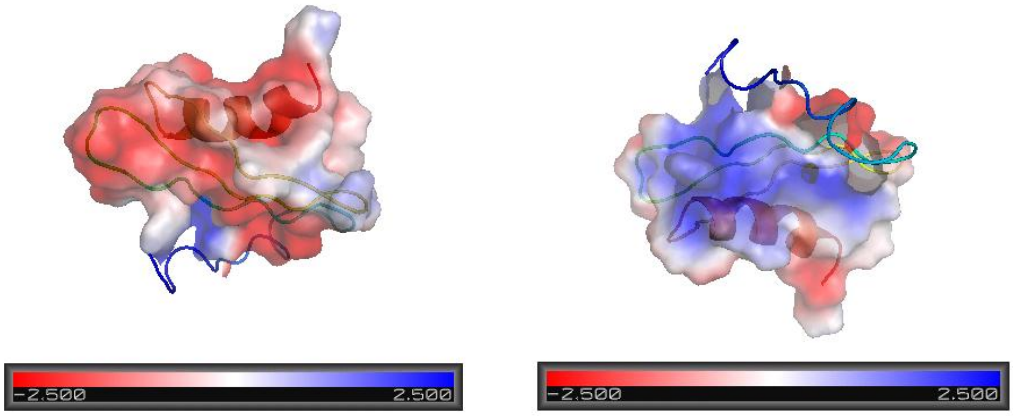
B. Tel-KO2 farCTD oriented to show the most positively charged surface, along the β1/helix face.

C. Tel-VHML farCTD oriented to show the most negatively charged surface, along the β3/helix face.

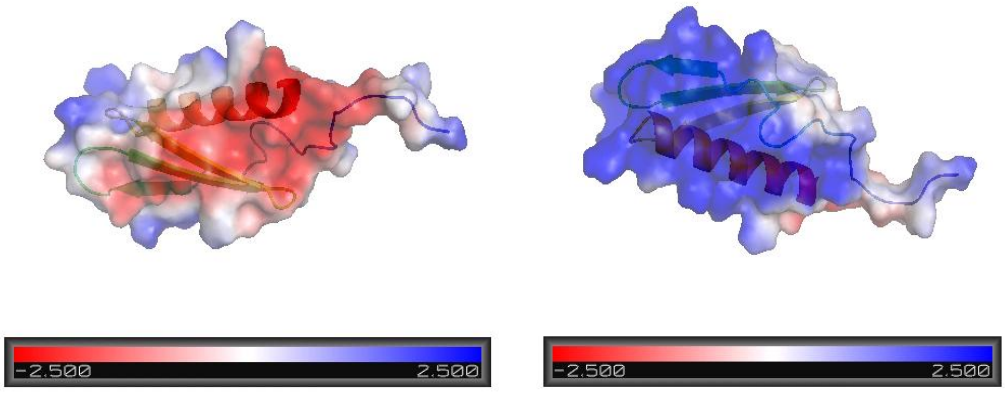
D. Tel-VHML farCTD oriented to show the most positively charged surface, along the β1/helix face.

The two views of each protein are rotated approximately 180˚ relative to each other along the horizontal axis.

A. Tel-KO2 negatively (left) and B. positively (right) charged surfaces



C. Tel-VHML negatively (left) and D. positively (right) charged surfaces



class of binding partner of the farCTD, which will facilitate the design of future studies.

A Dali structure search revealed that the protelomerase farCTD structures most closely resemble a fold known as the double-stranded RNA-binding domain (dsRBD), the founding member of the double-stranded RNA-binding domain-like superfamily (*91*). The dsRBD-like superfamily also consists of the homologous–pairing domain of Rad52 recombinase and the ribosomal S5 protein, N-terminal domain. The structural statistics of the protelomerase farCTDs compare favorably with those of other recently published dsRBDs (*92*). Proteins which contain dsRBDs can be separated into classes based on their known binding partners. dsRBDs are known to bind RNA, DNA, and other proteins. An analysis of the conserved features of each class of dsRBD will help to indicate which macromolecular class is the likely partner of the protelomerases farCTD, and will aid in the design of future studies.

The double-stranded RNA binding domain

The classic dsRBD is an αββββα fold in which an anti-parallel three-stranded β-sheet packs against a carboxy-proximal α-helix and is further stabilized by an additional amino-proximal α-helix. The dsRBD was identified in 1992 by St. Johnston et al., who noted repeated regions of sequence similarity in the *Drosophila melanogaster* gene, staufen, and a *Xenopus laevis* gene of then unknown function (*93*). The group used the repeats in both systems to identify a minimum consensus sequence, which they used to search the protein sequence database. In addition to the repeats in staufen (required for the localization of maternal mRNAs) and the *X. laevis* gene, the product of which was later named Xlrbpa, (for *Xenopus laevis* RNA-binding protein A, a homolog of human TRBP involved in formation of the RISC complex) additional proteins containing the

consensus sequence were identified, several of which were known or thought to bind double-stranded RNA. These included human dsRNA-activated inhibitor (DAI), human trans-activating region (TAR) binding protein (TRBP), and RNase III. The thousands of examples of this domain that have since been identified are from widely diverse origins ranging from archaea, bacteria, and viruses to eukaryotes, in which they are found in both the nucleus and cytoplasm of both plants and animals. Although all dsRBDs share a structural fold, dsRBDs that bind different classes of macromolecules have distinct characteristics that reflect their different modes of macromolecular recognition. Based on principles of sequence conservation, general architecture, and hierarchy of important observed binding regions between the farCTD and other dsRBDs, it is most likely that the farCTD binds DNA. However, it is possible that the domain may also interact with other protein elements, either of protelomerase or other origin.

dsRBD as RNA-binder

As the eponymous and founding member of the superfamily, RNA-binding dsRBDs define the classical architectural fold (αββββα) and preferentially bind dsRNA over dsDNA and RNA/DNA hybrids. Sequence conservation has been extensively studied and is well characterized (see Figure 2.8). It extends across all four elements of secondary structure and in the β1β2 and β3α2 loops, and is especially high in the three regions involved in direct RNA interaction (*94*). Classifications of dsRBDs have been made in terms of sequence, the number of copies of dsRBDs within the dsRBD-containing protein, the presence of other intraprotein domains such as a catalytic domain, and RNA-binding ability (*95*). The five separate dsRBDs in PKR, for instance, are classified as to whether or not they bind dsRNA at all (dsRBDs 1, 3, and 4 do) as well as

Figure 2.8 Comparative consensus sequences of dsRNA-binding dsRBDs and DNA-binding dsRBDs with known and putative protelomerase farCTDs.

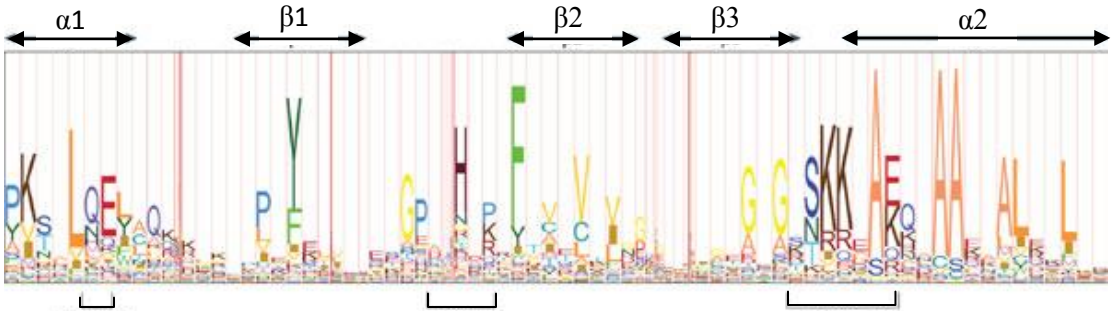
A) The consensus sequence for RNA-binding dsRBDs derived from the sequence logo of the hidden Markhov model (HMM) created by Tian et al. (*94*). Reprinted by permission from Macmillan Publishers Ltd: [NATURE REVIEWS | MOLECULAR CELL BIOLOGY] (Tian et al, 5, 1013-1023 (December 2004) doi:10.1038/nrm1528 citation), copyright (2004). The HMM was created from all dsRBDs then available in the InterPro database (1,428 total.) Adjacent residues are separated by a line, the thickness of which indicates the probability of an insertion of random residues. The width of each residue indicates the probability of having a residue at that position (narrow residues tend to be deleted at their positions compared to wide residues.) The height of a residue indicates how conserved it is among all dsRBDs. Secondary structures of the dsRBD are indicated above the sequence logo. The three dsRNA interaction regions (regions 1, 2 and 3) are marked below the sequence logo. Intervening sequences between α-helices (α) and β-strands (β) are known as loops: loop 1 is between helix 1 and strand 1, loop 2 between strand 1 and strand 2, loop 3 between strand 2 and strand 3, and loop 4 between strand 3 and helix 2. A consensus sequence is shown below the RNA-binding dsRBD HMM, with residues that are also conserved among protelomerase farCTDs highlighted.

B) An alignment of known and putative farCTD sequences is shown, with a farCTD consensus sequence displayed both above and below the alignment. Within the alignment, residues conserved among farCTDs are indicated in red text and underlined, while secondary structure (determined by solution or homology) is highlighted, and is indicated above the alignment: (α, alpha helix; (β, extended or beta strand conformation.) In the consensus shown above the alignment, residues conserved between farCTDs and RNA-binding dsRBDs are highlighted. In the consensus sequence displayed below the alignment, residues conserved between DNA-binding dsRBDs are highlighted.

C) A consensus sequence and alignment of DNA-binding dsRBDs. Secondary structure is indicated as described in B. Within the β-strands, externally-facing residues are indicated by colored text. Conserved residues are underlined. The consensus sequence is displayed above the alignment, with residues that are also conserved among protelomerase farCTDs highlighted.

All three consensus sequences contain alternating hydrophobic residues in the region of strand β2 which probably reflects conservation of core-stabilizing residues, as well as a small (glycine or proline) residue at the beginning of strand β1 (G or P) and the end of strand β3. Outside of these regions, only one other residue is well-conserved between RNA-binding dsRBDs and protelomerase farCTDs (an alanine in the middle of the C-terminal α-helix.) None of the points of conservation between farCTDs and RNA-binding dsRBDs are in the regions noted for RNA-interaction. Outside of regions of total domain conservation already mentioned, similarity between the protelomerase farCTD and DNA-binding dsRBD consensuses is greatest just N-terminal to the beginning of and at the C-terminal end of the strand β1, which contain conserved positive and primarily positive or polar residues, respectively. In the DNA-binding dsRBDs, these residues interact with the DNA phosphate backbone.

A. RNA-binding dsRBD concensus and hidden Markhov Model of 1,428 dsRBDs




```
RNA-binding dsRBD consensus sequence:
PK^-L^E-----+----P-Y-------GP--H-P-F-h-V-h-G/P-----G-G-SKK--AE^-AA--AL---L--
```

B. farCTD sequence alignment and consensus sequence

```
   farCTD consensus sequence:
      ------------K-P^h^u--+----GDG---%-h-h-h-G---%----G--u---h'Ah+-A%^-%------
      --------------ββββββββββ---------βββββββββ---βββββββββ----αααααααααααααααααααα
VHML  QEEDQKVSWPKAKDIKVQ--SKKE---GD--MWHVWTEVNG-MRWEN-WSK-GR-KTEAVKALRQQYERESAEM
KO2   AEVAEQEEKHPGK-PNFKA-PRDN--GDG--TYMVEFEFGG-RHYAW-SGAAG-NRVEAMQSAWSAYFK
N15   EEGPEEHQPTALK-PVFKP-AKNN--GDG--TYKIEFEYDG-KHYAW-SGP-ADSPMAAMRSAWETYYS
PY54  SDNASDEDKPEDK-PRFAAPIRRS--ED---SWLIKFEFAG-KQYSW-EG-NAESVIDAMKQAWTENME
VP882 VPAAEKQPKKAQK-PRLVAH-QV---DDE--HWEAWALVEG-EEVAR-VKIKG-TRVEAMTAAWEASQKALDD
Halm  VAAAVPKEVAEAK-PRLNAHPQ----GDG--RWVGVASING-VEVAR-VGNQAG-RIEAMKAAYKAAGGR
VCamp VVETKPKDETVIK-PKMKGH-KE---DDG--TWLVDVTIED-KSWQISVGKEPKNVMDAFKLAWNEFvyrk…

   farCTD consensus sequence
        ----------K-P^h^u--+^---GDG---%-h-h-h-G---%----G--u---h'Ah+-A%^-%-------
```

C. DNA-binding dsRBD consensus sequence and alignment:

```
                  +-G/P----+^----G----%-h-h------G----hG-------------------------
      ----------------βββββββ------βββββββ-------βββββββ--αααααααααααααααααααα
1gcc   …AVTAAKGKHYR--G--VRQRPW---G---KFAAEIRDPAKNGARVWLGTFETAE-DAALAYDRAAFRMR…
Tn9  MSEKRRDNRGRILKT-G--ESQRK---DG---RYLYKYIDSF--GEPQFV-YS…
                                                …WKLVATDRVPAGKRDAISLREKIAEL
λINT     MGRRRSHERRDLPPNLYIRNN---G---YY--CYRDPRT-G-K-EFGL—GRDRRIAITEAIQANIELFS…
i-Ppo1                                                …ALTNAQILAVIDSWEETVGQFP…
…LGGGLQGTLHCYEIPLAAPYGVG-FAKN----G-PTRWQYKRTINQV---VHRWGSHTVPFLLEPDNINGKTCTA…
```

which bind RNA with the greatest affinity (*96, 97*).

Beyond a preference for dsRNA, the question of binding specificity by the dsRBD must be addressed in diverging but related aspects: RNA sequence specificity, RNA helix geometry specificity, and RNA "secondary structure" specificity. Although most data suggests that dsRBDs do not impart binding specificity, increasing numbers of studies imply degrees of binding preference. At this juncture, it is perhaps only possible to say that while most dsRBDs bind most dsRNA, some dsRBDs bind dsRNA more tightly, and do so with varying exclusivity. Appendix B contains a literature review of the current understanding of the ability of dsRBDs to impart binding specificity to the proteins in which they are located.

Multiple structures of dsRBDs in complex with dsRNA have revealed common recognition and binding patterns. As reviewed by Tian et al., in all cases, including the second dsRBD in *X. laevis* XlrbpA with a non-physiological dsRNA (pdb accession code 1DI2) (79), *D. melanogaster* staufen dsRBD3 in complex with an RNA hairpin (pdb accession code 1EKZ) (80), *A. Aeolicus* RNaseIII dsRBD in complex with dsRNA (pdb accession codes 1RC7 and 1YYW) (81), and *S. cerevisiae* RNase III (Rnt1p) in complex with the 5' terminal RNA hairpin of snR47 precursor (pdb accession code 1T41), the domain binds along a single face of the dsRNA (*94*). Protein to RNA interactions are primarily to RNA phosphate and 2'-hydroxyl groups rather than to sequence-specific contacts, and are often water-mediated. Along the face of the RNA, successive minor-major-minor grooves are contacted by protein helix α1 (2'-hydroxyls contacted), the β3–α2 loop (backbone phosphates contacted), and the β1–β2 loop (2'- hydroxyls contacted) respectively. Figure 2.9 shows an overlay of three dsRBDs which have been solved

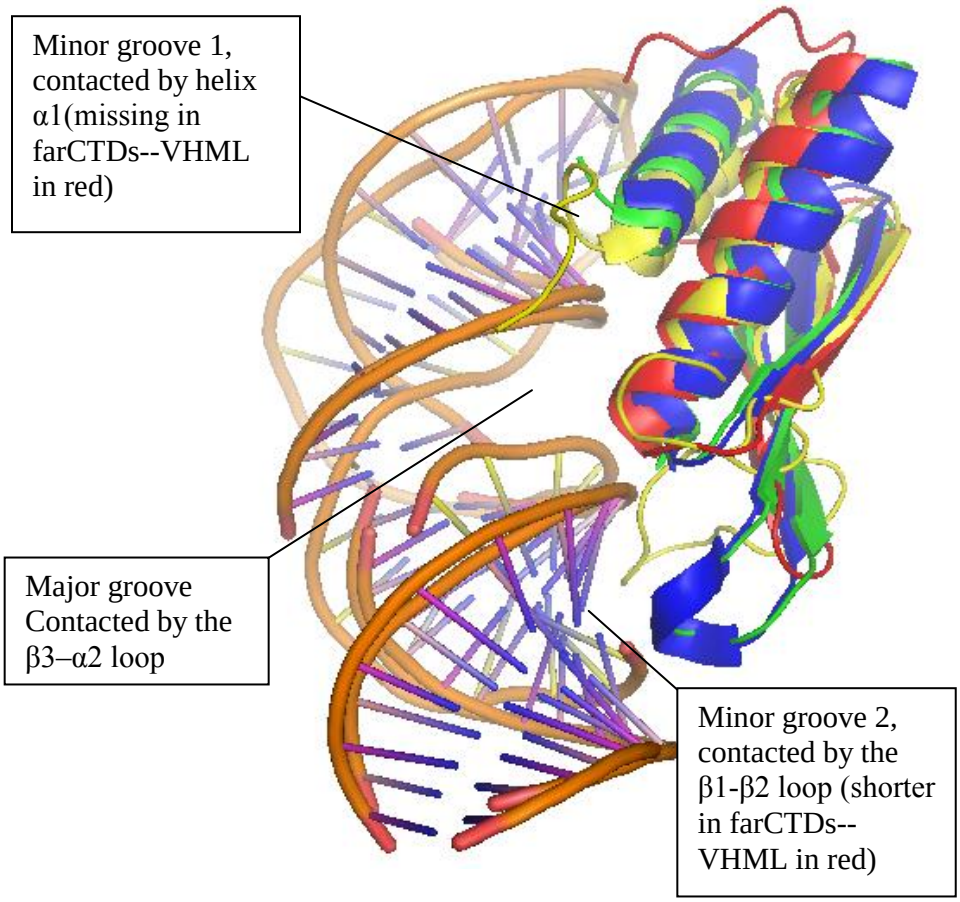


Figure 2.9 Comparison of VHML farCTD with reported structures of RNA-binding dsRBDs bound to dsRNA.

Overlay of the solved dsRBDs bound to dsRNA with Tel-VHML farCTD as calculated in a DaliLite structure comparison. *Xenopus laevis* XlrbpA dsRBD2with a non-physiological dsRNA (blue) (*79*), *Drosophila* staufen dsRBD3 in complex with an RNA hairpin (yellow) (*80*), *Aquifex Aeolicus* RNaseIII dsRBD in complex with dsRNA (green), and Tel-VHML farCTD (red). The structures show the common recognition modes as indicated.

bound to dsRNA, along with our Tel-VHML farCTD structure.

Of the three important RNA interaction motifs of the known RNA-binding dsRBDs, only the third, the loop between strand β3 and helix α2, is comparatively structurally well-conserved in the farCTD. Among RNA binders, however, the highly conserved sequence in this region consists of an SKK*AE, which is not present in the farCTD. The first interaction motif, helix α1, is not present in the solved structures of the protelomerase farCTD. It is possible that a helix forms in the unstructured region of the farCTD on binding or that it interacts with a helix from a remote region of the protein, but such interactions are not described for known RNA-binding dsRBDs, for which the secondary structural elements are always contiguous. The second interaction motif, the loop between strands β1 and β2, is the site with the most variability in length for RNA binders, but is nevertheless usually longer than the loop observed in the protelomerase farCTD (*94*). The RNA binding dsRBDs contain a highly conserved GP*H sequence in this region, which in the farCTD is substituted by a DG-rich sequence (*94*). The β1β2 turn in DNA-binding dsRBDs, on the other hand, contains a glycine but lacks other notable conservation.

dsRBDs as protein binders

Some dsRBDs are known to interact with protein elements, although this function is much less well characterized than RNA recognition. The majority of protein-binding dsRBDs are grouped with RNA-binding dsRBDs, and the degree to which the domain is dedicated toward RNA versus protein binding is sometimes a matter of controversy, as in the case of the two dsRBDs of rat ADAR2 (*98*). However, the dsRBDs in several proteins have been shown to act as dimerization domains, allowing dsRBD-containing

proteins to form homodimers or to dimerize with heterologous proteins which also contain dsRBDs (*99-103*). dsRBDs have also been shown to interact with other, non-dsRBD protein domains (*104, 105*). Despite the classification of some dsRBDs as protein-binding domains, however, the identification of the principles behind protein binding has remained elusive.

A comprehensive or definitive analysis of the characteristics of protein-binding dsRBDs has yet to be accomplished, either by sequence conservation or locational mapping on the domain. St. Johnston originally identified two types of dsRBD based on level of sequence conservation (*93, 106*). Type A dsRBDs maintained conservation over the entire length of the domain, while type B dsRBDs maintained sequence conservation primarily in the C-terminal region, but less so in the N-terminal sides, with only moderate conservation in the second RNA-interacting region (the β1β2 loop.) While some of the identified type B dsRBDs (such as Xenopus XlrbpA dsRBD3 and human TRBP dsRBD3) have since been shown to bind protein elements, no consistent sequence consensus of this relatively small group has been identified, other than a comparative lack of conservation in N-terminal regions. A consistent or definitive mapping of important protein-binding regions is also lacking. An early study of the dsRBDs of PKR identified the hydrophobic residues on one side of an amphipathic helix as the dimerization site, but a later structure showed that this side of the helix interacted with the three-stranded β-sheet and was buried in the protein, and the mutations (A or L to E) were probably more likely to compromise the entire structural fold rather than simply interrupting dimerization (*107, 108*).

RNA-independent interaction of dsRBDs has been documented primarily by

biochemical means, and although a few structures showing dsRBD-protein interactions have been solved, more are needed before a consensus can be described. The current structures provide examples showing different binding surfaces being utilized for protein interaction. Although each of these structures provides only a single example of a protein-binding face, the interactions are worth noting as possibilities for the function of the protelomerase farCTD. The first example shows the only example of a dsRBD/dsRBD interaction. The structure of the human DGCR8 core, containing two dsRBDs and a C-terminal region (pdb accession code 2yt4), was recently determined at a resolution of 2.6 Å (*109*). In complex with its partner Drosha, DGCR8 cleaves primary microRNA (pri-miRNA) substrates into precursor miRNA and initiates microRNA maturation. As the structure shows, the two dsRBDs interact through the β2β3 loop of dsRBD1 and the C-terminal end of dsRBD2's helix α2 (see Figure 2.10). Both domains exhibit many more interactions to a central helix, which is sandwiched between the two dsRBDs, primarily by the C-terminal regions of their second helices (α2). Both dsRBDs also show interactions with other protein elements through one side of their first helices, α1. Interestingly, the regions of the dsRBDs which show interactions to each other and to the sandwiched helix are different than those traditionally associated with dsRNA binding, which may indicate that in this case dsRNA binding and protein binding by a dsRBD need not be mutually exclusive. A model showing dsRNA interaction with an intact DGCR8 core was proposed based on a superimposition of the DGCR8 dsRBDs with the Xlrbpa–dsRNA complex structure (*79*). Although the dsRNA modeling is compelling, DGCR8 was not solved in the presence of its RNA substrate, and more data is needed to support the biochemical evidence that suggests the protein does not undergo

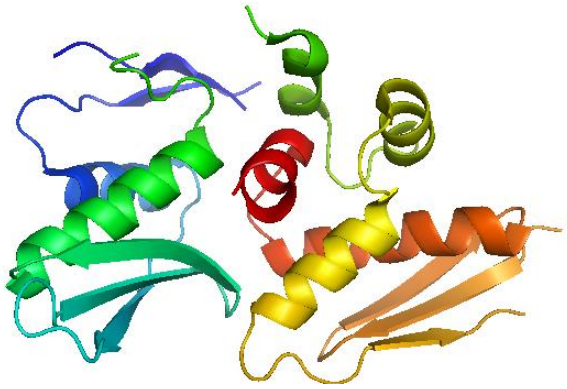
Figure 2.10 Crystal structure of the human DGCR8 Core showing dsRBD/protein contacts.

A) The DGCR8 core colored in spectrum with the N-terminus beginning in blue and the C terminus finishing in red.

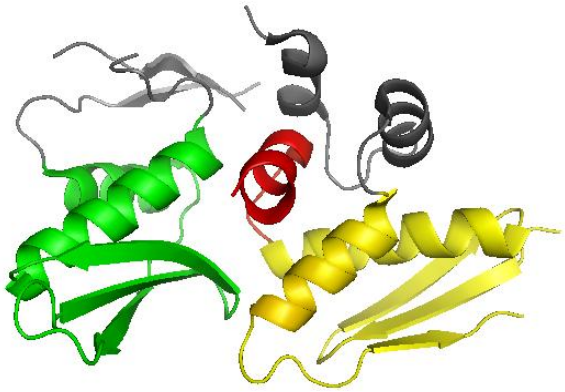
B) The DGCR8 core showing the two dsRBDs and the central helix: dsRBD1 is colored green, and dsRBD2 is colored yellow. The central helix is shown in red. Both dsRBDs show interactions to the central helix with the C-terminal end of their second alpha helices (helix α2.) Loop β2β3 of dsRBD1 interacts with the C terminal end of dsRBD2s helix α2. The dsRBDs also make protein contacts with their first helices, α1, dsRBD1 with β-strands from elsewhere in the core, and dsRBD2 with the central alpha helix.

C) The model created by Sohn et al. by superimposing the Xlrbpa–dsRNA complex structure onto the dsRBDs of DGCR8. The molecule is rotated approximately 90˚ around a horizontal axis from the views in A) and B), and shows that the RNA-binding surfaces are different than those that bind protein.

A.

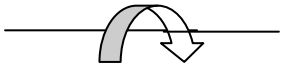


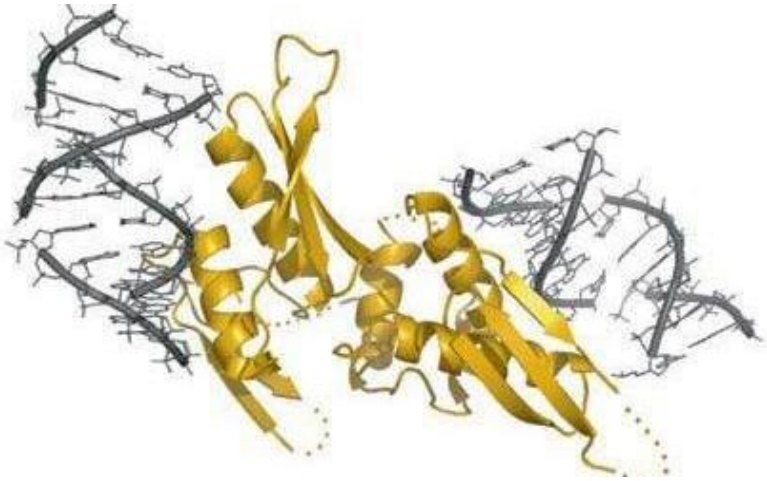
B.



dsRBD1(green) dsBD2 (yellow)

C.





dsRBD1 dsBD2

extensive remodeling in order to act catalytically with its partner. Nevertheless, the structure provides an example of protein-interacting surfaces of a dsRBD. Proteins with dsRBDs are often highly modular, containing flexible or extended linkers in between domains which has resulted in a less complete understanding of protein binding elements. The family of bacterial RNase III enzymes, for example, contain an N-terminal catalytic domain and a C-terminal dsRBD attached by a flexible linker. Previous crystal structures, such as the *Mycobacterium tuberculosis* RNase III, solved as a 2.1 Å homodimer, did not contain credible electron density for the carboxy terminal dsRBD, leaving the authors to speculate that the domain was highly mobile with respect to the nuclease domain (*110*). A recent crystal structure of *Aquifex aeolicus* RNase III was solved in the presence of dsRNA and showed both the endonuclease and dsRBD engaged with the dsRNA, but there was no interaction between the two domains, which are separated by a seven residue linker (*111*). The *Thermotoga maritime* RNase III structure, however, solved at 2.0 Å resolution, shows a second example of a dsRBD/protein interaction. The *T. maritime* structure was solved as a homodimer without RNA, and shows an apparent interaction between the carboxy-terminal dsRBD and its amino-terminal nuclease domain. The *T. maritime* RNase III structure was deposited by the Center for Structural Genomics (Wilson, I.A., deposition: 2002-09-13, release: 2002-11-13, publication forthcoming; DOI: 10.2210/pdb1o0w/pdb; pdb accession code 1O0W) and shows that in this case, the primary protein interacting surface of the dsRBD is the face of the second helix α2, which interacts with the nearly parallel helix of the endonuclease domain (see Figure 2.11.) This series of structures, like the model for the DGCR8 core, not only provides an example of a protein-interacting face of a dsRBD, but

Figure 2.11 Interactions of the dsRBD of bacterial RNase IIIs

A) and B) show the dsRBD-endonouclease interaction of the *Thermatoga maritime* RNase dimer (pdb accession code 1O0W.) C) and D) show the dsRBD-RNA interaction of the Aquifex aolicus RNase III (pdb accession code 2NUG.)

A) One monomer is shown in gray, while the other is colored in spectrum with the N-terminus beginning in blue and the C terminus finishing in red. In the spectrum-colored monomer, the dsRBD is on the upper left-hand side and colored in red and orange

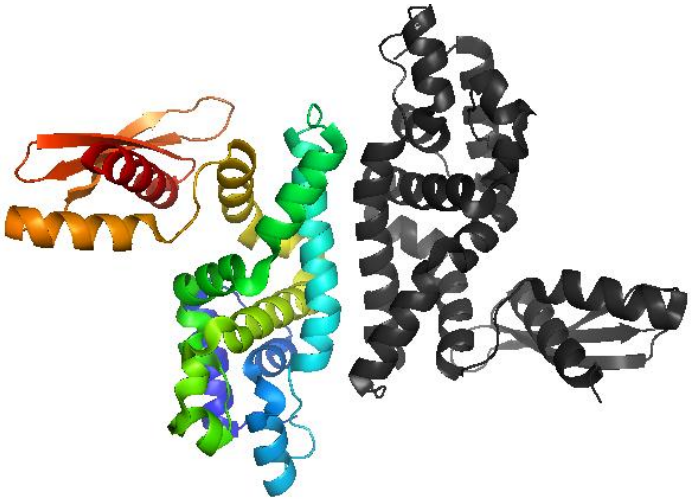
B) A close-up of one monomer of the RNase III from *T. maritime*. The dsRBD is colored in red and orange. Extensive interactions occur between the face of the C-terminal α-helix (red) and the nearly parallel helix of the endonuclease domain (yellow).

C) RNase from *A. aolicus*, crystallized with RNA. The enzyme is oriented with the catalytic domain positioned similarly to the catalytic domain of *T. maritime* RNase shown in A), and is colored similarly, with one monomer in gray and the other shown in spectrum. The dsRBD of the spectrum-colored monomer is colored in orange and red. The dsRBD (shown in red and orange) has obviously changed orientation (as described in the manuscript presenting the structures.)

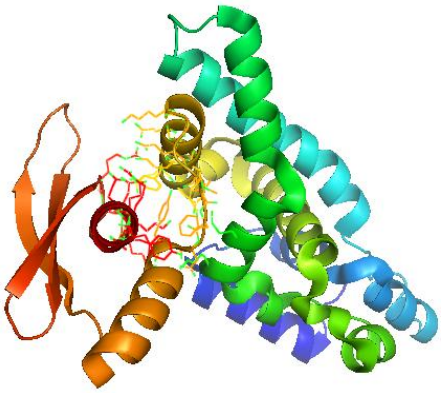
D) The *A. aolicus* structure has been rotated to show that the dsRBD (red) is not bound to the catalytic domain (green) but the linker has allowed it to perform its expected function (RNA binding.)

Thermatoga maritime RNaseIII

A.

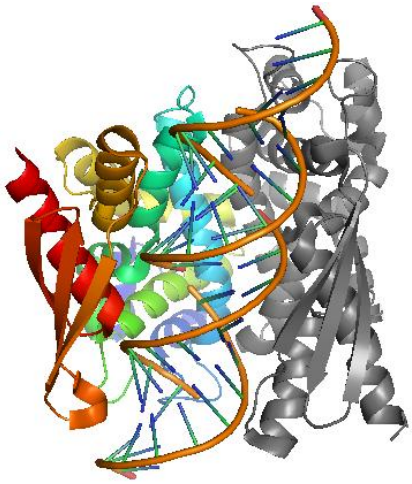


B.

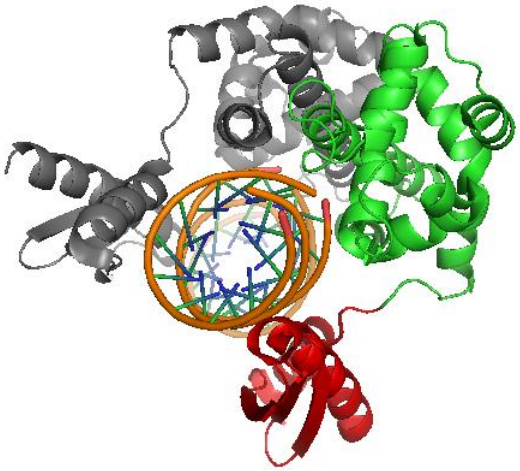


Aquifex aolicus RNase III

C.



D.



implies that a single dsRBD can bind more than one type of macromolecule. Despite the growing understanding of the role of dsRBDs in protein interactions, too few examples are known. Further biochemical and structural analysis will be required before determining principles such as sequence conservation and binding patterns can be identified. Thus, while the possibility of farCTD/protein interaction cannot be ruled out, such a functional role is also not conclusively supported by comparison to known structures.

DNA-binding dsRBDs: Nonclassical dsRBDs

A small number of proteins containing a non-classical dsRBD have been recently reported to bind dsDNA in a novel way, and comprise a new DNA-binding fold. The proteins include *Arabidopsis thaliana* ethylene responsive factor domain 1 (AtERF1-DBD) (*78*), the amino terminal DNA-binding domain of the conjugative transposon Tn916 integrase (Tn916-Int-DBD) (*76*), and the arm-binding domain of *Escherichia coli* bacteriophage λ integrase, sometimes called the N-domain. (*60*). All structures were reported as protein/DNA complexes. The domains present as a variation on the classical dsRBD in that the first helix, α1, is missing, resulting in a contiguous, anti-parallel, three-stranded β-sheet which packs against a carboxy-proximal α-helix (βββα.) A fourth protein, a homing endonuclease from the slime mold *Physarum polycephalum* (I-*Ppo*I) (*77*), has been found to bind DNA in a similar way, but differs from the other structures in the architectural arrangement of its secondary structural elements: the helix against which the three-stranded β-sheet binds is not directly carboxy-proximal to the β-strands, but originates instead from a noncontiguous, amino-relative location in the protein.

For all four proteins, the DNA binding surface is the face of a β-sheet composed

of three contiguous, anti-parallel strands whose comparative lengths (among the proteins) are similar but not identical. The β-sheet inserts into the major groove of its cognate DNA, where alternating residues project toward the interior of the protein or directly out from the face of the β-sheet where they interact site-specifically with DNA bases or with the phosphodiester backbone. The dsRBD is oriented within the major groove such that strand β1 (N to C) is parallel to the proximal DNA backbone which is oriented in a 5'to 3' direction. Strand β3 (N to C) is proximal to the complementary strand's backbone, which aligns 3' to 5' in a parallel manner (see Figure 2.12). The β-sheet adopts a conserved curvature such that the DNA-facing side of strands β1-β2 forms a concave surface, while the β2-β3 DNA-facing surface is convex. This hybrid curvature maximizes the DNA contacting area by molding the protein to the shape of the major groove.

The location of the residues within the β-sheet that are important for base recognition varies among the four proteins. A typical three-stranded β-sheet is slightly wider than can be easily accommodated within the DNA major groove, and the four proteins have different strategies to overcome that problem (*112*). The Tn916-Int-DBD is rotated slightly within the major groove so that strands β2-β3 insert more deeply and contain all but one of the residues responsible for base-recognition, while strand β1 abuts the proximal DNA backbone and contains small externally facing residues that accommodate backbone proximity. The AtERF1-DBD is also rotated slightly within the major groove, but in the opposite direction of the Tn916-Int-DBD, inserting strands β1-β2 more deeply, with small residues on the C-terminal end of strand β3 to accommodate the proximal DNA backbone. The I-Ppo1 structure shows that the central strand of the β-

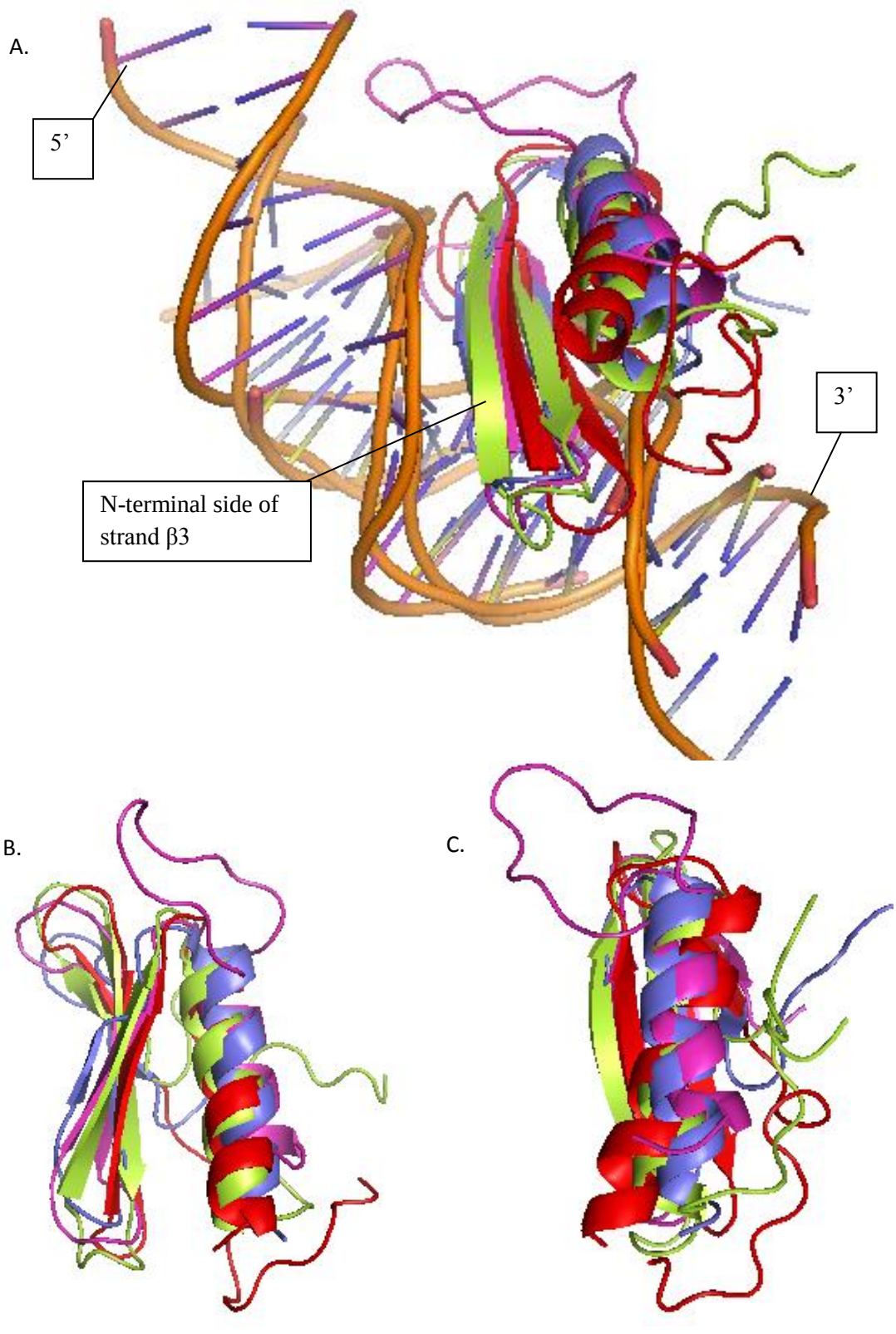
Figure 2.12 Comparison of VHML farCTD with reported structures of DNA-binding dsRBDs bound to DNA.

Overlay of the solved DNA-binding dsRBDs bound to dsDNA with Tel-VHML farCTD as calculated in a DaliLite structure comparison. 1z1g: λ-integrase arm-binding domain bound to accessory (arm) DNA (slate) (*60*); 1tn9: amino terminal DNA-binding domain of the Tn916 integrase protein with its dsDNA binding partner (*76*, *112*) (magenta), 1gcc: Arabidopsis thaliana ethylene responsive factor domain 1 and its DNA binding partner (*78*) (lime-green); Tel-VHML farCTD (red). The structures show the common recognition modes as indicated.

A) A view showing the insertion of the β-sheets into the major groove

B) A view showing the similarity of the alignment of the β-strands and of the lengths of the intervening loops.

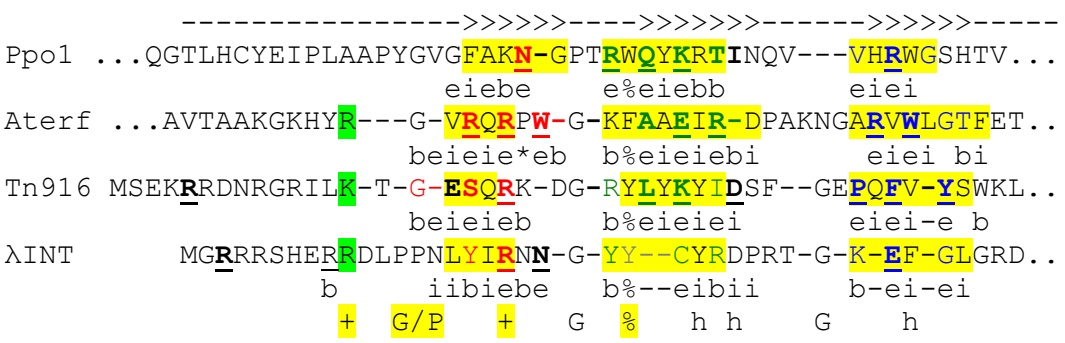
C) A view showing the orientation of the helices compared to the β-sheets. Much more variation is present in helix length and orientation compared to RNA-binding dsRBDs.


A.
5’
3’
N-terminal side of strand β3
B.
C.

sheet plays the main role in base-specific recognition, with three of the four externally facing residues making base-specific contacts, and the fourth contacting the phosphate backbone. One residue in each of the first and third strands in the sheet (β3 and β5 in context of the full protein) also make a base-specific contact, and portions of the first and third strand are highly twisted, allowing residues from both to insert into the major groove, which would otherwise only be able to accommodate recognition residues from two of the three strands. A recent solution structure of the λ-Int-N domain, however, shows the opposite; none of the residues projecting outward from the central strand make base-specific contacts with its DNA binding partner. Instead, residues from the first and third strands are primarily responsible for base recognition, as well as a residue which originates from the β1β2 loop. Strand β1 is highly twisted, and also contains a bulge, which appears to allow residues from both strands β1 and β3 to insert into the major groove (*113*).

Unlike comparisons of RNA-binding dsRBDs, the significance of sequence alignments of DNA-binding dsRBDs is in the lack of conservation in key binding regions (see Figure 2.13.) Sequence conservation of DNA-binding dsRBDs is extremely limited (the AtERF1 and Tn916-Int-DBD domains, for example, retain only 12% sequence homology when their secondary structural elements are aligned) especially among the externally-facing residues of the β-sheet. This lack of conservation reflects the ability of DNA-binding dsRBDs to recognize and bind different DNA sequences through base-specific major groove contacts. Only one study has compared the sequences of DNA-binding dsRBDs. In 2000, Connolly et al. reported a comparative analysis of the Tn916-Int-DBD/DNA complex with the previously published AtERF1 and I-PpoI structures

A.



B.

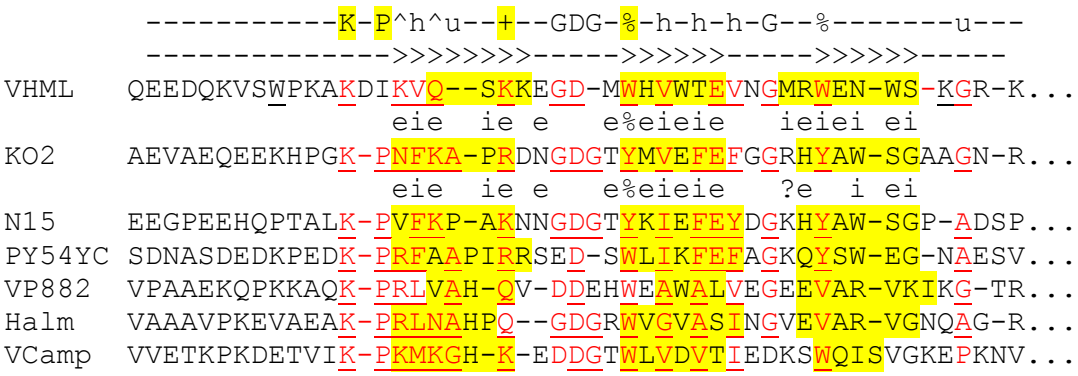


Figure 2.13 Comparison of β-strand residues of DNA-binding dsRBDs and protelomerase far CTDs.

A) Sequences of DNA-binding dsRBDs are aligned according to secondary sequence, with the orientation of residues (internally or externally facing) indicated beneath each sequence. Within each sequence, externally-facing residues are indicated with colored text, and residues important for base-specific interaction (as described in the original manuscripts) are shown in bold text and are underlined. Symbols representing orientation of residues are as follows: interior facing residue, (i); backbone-interacting residue, (b); other external facing residues, (e); conserved, interior-facing aromatic residue in strand β2, (%). A consensus sequence is located below the alignment.

B) Sequences of Tel-VHML and Tel-KO2 are shown, with orientation of residues in β-strands indicated beneath each sequence as in A. The putative protelomerase farCTDs are also shown, and are aligned according by the relatively high conservation of hydrophobic residues found at roughly alternating positions within the strands. Residues with relatively high conservation are indicated with red text and are underlined. A consensus sequence is located above the alignment.

(*112*). After noting the minimal sequence conservation Connolly instead noted four related, positionally conserved, DNA-protein contacts. These included two interfacial contacts to the phosphodiester backbone (a glycine prior to strand β1 and a positively charged residue at the beginning of strand β2) and two base-specific hydrogen bonds (originating from the second external residue on strand β1, and from a charged residue in strand β2.)

The λ-Int-N-domain structure published after Connolly's study reduces conservation even further. The conserved anchoring residues noted in Connolly's study (the glycine located just proximal to strand β1 and the basic residue which is the first externally facing residue of strand β2) are replaced in the λ-Int-DBD structure by Pro14 and Tyr23, respectively, and there are no specific base contacts originating from strand β2. However, the addition of the λ-Int-N-domain to the fold highlighted conservation that was not as obvious in the comparison of the three prior structures. Not present in the I-Ppo1 structure, but conserved among the three proteins which maintain the dsRBD fold is a basic residue located before strand β1 which appears to interact with the phosphodiester backbone in a manner to further anchor the domain within the major groove. The structures also all contain a conserved basic or polar residue at the end of the strand β1 which is poised to interact with the phosphodiester backbone.

In terms of overall architecture, there is much more variability in the DNA-binding structures than in RNA binding dsRBDs. The length of the individual strands which make up the β-sheet varies from protein to protein, and the presence and level of twist among strands varies significantly as well. In AtERF1, a beta bulge in strand β3 serves to accommodate a proper fit against the proximal DNA strand, while λ-Int-N-

domain has a β-bulge in strand β1. In Tn916, an extra insertion between strand β3 and the alpha helix creates a loop which specifically interacts with bases just 3' of the dsRBD binding site, while the λ-Int-N-domain has recently been shown to have an extended region N-terminal to the dsRBD which becomes structured on binding, making two additional contacts in each of the major and minor grooves. The extended segment forms a short 3^{10} helix and then wraps around the DNA strand, inserting into the minor groove (*113*).) Among the four DNA-binding dsRBDs, the length and orientation of the alpha helix to the β-sheet also display significant variability (see Figure 2.12.) Despite these differences, the general size, conformation, and fit of all structures within the major groove are analogous, and clearly define a structural fold and binding motif.

Protelomerase farCTD/dsRBD class comparison

Both the observed similarities and variablilities of different aspects of the Tel-KO2 and Tel-VHML structures are more consistent with DNA-binding dsRBDs than with RNA-binding dsRBDs. The protelomerase farCTDs share the overall architecture of DNA-binders (βββα), with comparable lengths for individual β-strands and intrastrand loops (see Figure 2.12.). The β1β2 loop in particular (important in RNA recognition for RNA-binders) is shorter in both DNA-binding dsRBDs and the protelomerase farCTD than in RNA-binding dsRBDs (Figure 2.8). The length and orientation of the alpha helix to the β-sheet show much more variability in the βββα structures than those observed in RNA-binders. The paradigm of two of the three β-strands of DNA-binding dsRBDs being more significant in DNA recognition is also reflected in the protelomerase farCTDs, with the β2β3 strands of the Tel-VHML being longer and more well-defined, while in the farCTD of Tel-KO2 strands β1β2 are better defined. The sequence

alignment and secondary structure prediction calculations of the putative protelomerase farCTDs with those of the solved protelomerase farCTDs indicates that they also will also reflect the DNA-binding dsRBD characteristics, including identical architecture, similar strand length, and variability in helix length. Like the DNA-binding dsRBDs, all solved and putative protelomerase farCTDs have a conserved basic residue prior to the start of strand β1, and an additional basic residue positioned to be homologous to the second externally-facing residue of strand β1. All known and putative protelomerase farCTDs except that of Tel-VHML have a conserved proline antecedent to strand β1, reflective of the λ-Int-N-domain (see Figure 2.13).

It is especially interesting that the λ-Int-N-domain belongs to the protein with which protelomerases share a catalytic mechanism and significant core domain homology. Lambda integrase is often considered the archetypal member of the tyrosine recombinase family, and it is this family that is most closely related to protelomerases in general. In fact, based on the similar overall genome organizations and close relationships of phages N15, KO2, and PY54 with *E. coli* phage λ and other lambdoid phages, it has been suggested that these protelomerase-containing phages should be classified as a lambdoid phage subgroup (*44*). Lamdba integrase has even been noted to produce covalently closed ends at *att* sites in *in vitro* experiments where mismatches are present in the core region (*114*). The λ-Int-N-domain is N-terminal to the two-lobed catalytic and core-binding domains which make up the λ-integrase central region and is attached by a flexible linker. The close relationship of λ-integrase to protelomerases, and their similarity in architecture suggests that the dsRBDs of protelomerase and λ-integrase may serve a related function. The λ-integrase N-domain recognizes and binds accessory

DNA (called ARM sites) outside of the core-binding site during integration and excision ((*61*) for a recent review.) Lambda integrase forms a four-part multimer, and the N-domains (also called ARM-binding domains) appear to influence equilibrium of the recombination reaction by shaping the complex via their interaction with each other in addition to their interaction with accessory DNA (*60*). The function of the λ integrase N-domain not only suggests that the protelomerase farCTD binds DNA, but provides a precedent for a possible second, protein based interaction in its role.

Sequence conservation between a DNA binding dsRBD and a putative phage protelomerases farCTD

Conservation of key residues between a known DNA-binding dsRBD and a putative phage protelomerase farCTD along with accompanying cognate DNA sequence conservation further provide compelling indication that the farCTD binds DNA. The Tn916-Int-DBD is the DNA-binding dsRBD for which the most structure/function information is known. Tn916 is a well-characterized conjugative transposon encoding tetracycline resistance that excises and integrates in a manner similar to *E. coli* bacteriophage λ. The protein that accomplishes both the strand cleavage and rejoining aspects of transposition is the transposon-encoded integrase (Int) which is composed of two domains (*115*). The C-terminal, catalytic domain is, like the minimally enzymatically active N-terminal region of Tel-KO2, a member of the tyrosine family of site-specific integrases. It recognizes and cleaves inverted repeats at the transposon/chromosome junction. The catalytic domain is delivered to the site of recombination by the site-specific interaction of the N-terminal domain with its cognate DNA, which binds distally to the point of recombination (*116*) at inverted repeat sites

within the transposon arm (*115, 117, 118*). The size of the N-terminal DNA binding domain (called Tn916 Int-DBD) has been experimentally determined, and a structure of the isolated domain bound to its cognate DNA has been reported (*76*). Like the λ-Int-ARM binding domain, the Tn916-Int-DBD belongs to the dsRBD structural fold.

In 2000, Connolly et al. reported a series of rational mutagenesis and binding affinity assays to determine the specific protein residues and DNA bases along the binding surfaces of the Tn916 Int-DBD and its cognate DNA that are responsible for affinity. A series of fluoroscein-labeled and natural fluorescence assays monitored the effect of substitutions and point mutations on affinity (*112*). The residues whose mutations resulted in the highest affinity decrease were also identified as being responsible for base-specific interaction with the cognate DNA through unambiguous hydrogen bonding according to the solved structure. The residues responsible for these base-specific interactions were located on the external face of the β-sheet on all three strands. Additional important phosphate/hydrogen bond and base/residue van der Waal interactions were also noted. The base-specific hydrogen bonding interactions included R20 in strand β1, K28 in strand β2, and Y40 in strand β3. Phosphate contacts included G16 (amide hydrogen) and K21 in β1 and W42 in strand β3. Base van der Waals contacts, such as those originating from L26 in strand β2 and F38 in strand β3 were noted, but X to A mutation caused a much lower drop in DNA binding affinity.

A putative protelomerase farCTD with otherwise minimal conservation to the Tn916-Int-DBD contains nearly all of the residues noted in the Tn916-DBD mutagenesis study to significantly decrease binding affinity to cognate DNA when mutated. PY54 is a temperate phage which infects *Yersinia enterocolitica*, a gram-negative bacterium

belonging to the Enterobacteriaceae family. During its lysogenic phase, PY54 is harbored within *Y. enterocolitica* as a linear plasmid prophage with covalently closed hairpin ends (52). A 77kDa protein PY54 protein with sequence homology to the phage N15 protelomerase has been found to process recombinant plasmids containing a 42bp palindrome natively found upstream from the PY54 protelomerase-like gene on ORF 32 of the PY54 phage genome. The sequence of the 42bp palindrome reflects the sequence of the processed hairpin ends and also has some similarity to the protelomerase binding sites of Tel-N15 and Tel-KO2. *In vivo* protelomere creation has been shown to require both the protelomerase-like gene and the 42kb palindrome, as well as 15 bp inverted repeat sequence which flanks the palindrome at differing distances (66bp left, 45bp right), for efficient processing (51). Sequence alignment of the PY54 protelomerase (Tel-PY54) with solved phage protelomerases shows a probable farCTD with secondary structure prediction that is similar to those predicted for the Tel-VHML and Tel-KO2 farCTDs. Based on the PY54 secondary structure prediction and on sequence alignment with both solved protelomerases and DNA-binding dsRBDs, residues with probable positional homology to the Tn916-Int-DBD residues were identified. Table 2.3 shows the comparison of the Tn916 Int-DBD and the putative PY54 protelomerase farCTD along with the published values for affinity decrease caused by sequence mutation of the Tn916 Int-DBD. Figure 2.14 shows the sequence alignment of the Tn916 Int-DBD β-strands with the putative Tel-PY54 farCTD.

The DNA oligomer used in the solution structure of the Tn916-Int-DBD/cognate DNA complex, was the consensus sequence of the five Tn916-Int-DBD binding sites (5'-GAGTAGTAAATTC-3' (coding strand) (*116*).) The Tn916-Int-DBD differs from the

Table 2.3

Comparison of Tn916 Int-DBD residues important for DNA binding affinity to probable positionally homologous residues of putative PY54 protelomerase farCTD.

Tn916 mutation	Fold affinity decrease	Residue location	Contact type	PY54 farCTD comparison
Y40A	14	Strand β3	Base H-bond	Y606
K54A	10	Extended loop	Phosphate H-bond	None
K28A	11	Strand β2	Base H-bond	K598
R24A	8	Strand β2	Phosphate H-bond	S594
K21A	3	β1β2 loop	Phosphate H-bond	R590
R20A	2	Strand β1	Base H-bond	R589
W42A	2	Strand β3	Phosphate H-bond	W608
L26A	1	Strand β2	Base vdW contact	L596

Figure 2.14 Comparison of the external facing residues of the β-sheet of transposon Tn916 Integrase DNA-binding domain (Tn916-Int-DBD) to the probable homologous residues of putative protelomerase farCTD Tel-PY54.

A) The externally-facing residues of the Tn916-Int-DBD β-sheet are shown in the context of the three strands which make up the sheet. The strand number is indicated above the sequences, and internally-facing residues are indicated by the small letter (i), while the externally facing residues are numbered as described in the manuscript reporting the solved structure, and the interior-facing anchoring aromatic residue located in strand β2 is indicated by a % symbol. Externally facing residues shown in Connolly's study to result in a significant decrease in cognate DNA binding affinity when mutated are highlighted, and the residue number in context of the full-length protein is indicated below the sequences of the β-sheets. The homologous residues for putative protelomerase farCTD Tel-PY54 were identified by secondary structure prediction and by sequence alignment with solved Tel-KO2 and Tel-VHML farCTDs. Numbers of the probable positionally homologous residues are indicated below the sequence. The fold affinity decrease caused by X to A mutation in Connolly's Tn916 study is indicated below both sequences.

B) Sequence alignment of the Tn916-Int-DBD with putative the farCTD of protelomerases PY54 along with the farCTDs of Tel-VHML and Tel-KO2 showing the reasoning for the identification of probable homologous residues between PY54 and Tn916-Int-DBD. The secondary structure elements based on the solved structures (Tn916-Int-DBD, Tel-VHML, Tel-KO2) or the predicted secondary structure (putative Tel-PY54) are indicated by highlighting (β-sheet in yellow, α-helix in grey). The residues in the TN916-Int-DBD noted in Connolly's mutation analysis to significantly reduce DNA binding affinity on mutation are indicated by bold, red type. Residues noted for specific base interaction (either hydrogen bond or van der Waals) are underlined.

C) Sequence alignment of all known and putative protelomerases. The secondary structure elements (either of solved structures or likely homologous structures) are highlighted as in (B.) and are indicated above the sequence alignment (β for β-strand, α for α-helix. Conserved residues are indicated in red text.

D) Secondary structure prediction of the far C-terminal region of putative protelomerase Tel-PY54 as calculated by the PSIPRED program. The prediction is similar to those of the solved farCTDs of Tel-KO2 and Tel-VHML, in which only two of the three β-strands which make up the sheet are predicted.

A.

	β1	β2	β3
	1i2i3	%1i2i3i	1i2i3
Tn916	KTGESQRK	RYLYKLID	PQFVYSW
	16 20 21	24 26 28	40 42
PY54	RFAAPIRR	SWLIKFEF	KQYSW
	585 589 590	594 596 598	606 608
Affinity decrease	2 3	8 1 11	14 2

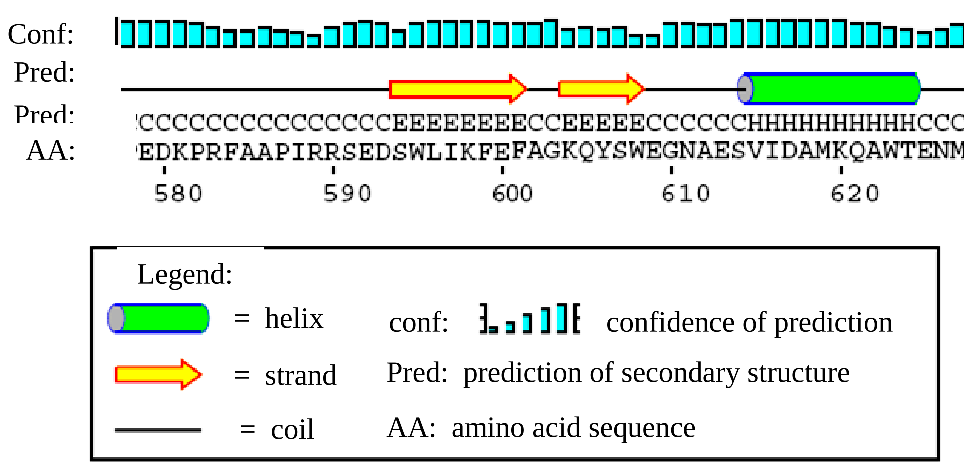
B.

```
              --------------βββββ------ββββββ-----βββββββ-----ααααααααααααααααααα
1Tn9    EKRRDNRGRILK-T--GESQRKDG--RYLYKYIDSFGEPQFVYSWKLVATDRVPAGKRDAISLREKIAEL
Py54   SDNASDEDKPEDK-PRFAAPIRRSED-SWLIKFEF-AG---KQYSWEGNAESVIDAMKQAWTENME
KO2    AEVAEQEEKHPGK--PNFKAPRDNGDGTYMVEFEF-GG---RHYAWSGAAGNRVEAMQSAWSAYFK*
VHML       DQKVSWP-KAKDIKVQSKKEGD-MWHVWTEV-NG--MRWENWS-KGR-KTEAVKALRQQYERESAEM
λINT       -B1B2 ERRDLPPNLYIRNNG--YY—CYRDPRTG--K-EF-GL-GRDRRIAITEAIQANIELFSGHI…
```

C. farCTD sequence comparison

```
             ---------------βββββββ------ββββββ----βββββββ-----ααααααααααααααααααα
VHML    QEEDQKVSWPKAKDIKVQ--SKKEGD-MWHVWTEVNGMRWEN-WS-KGR-KTEAVKALRQQYERESAEM
KO2     AEVAEQEEKHPGK-PNFKA-PRDNGDGTYMVEFEFGGRHYAW-SGAAGN-RVEAMQSAWSAYFK
N15     EEGPEEHQPTALK-PVFKP-AKNNGDGTYKIEFEYDGKHYAW-SGP-ADSPMAAMRSAWETYYS
PY54YC  SDNASDEDKPEDK-PRFAAPIRRSED-SWLIKFEFAGKQYSW-EG-NAESVIDAMKQAWTENME
VP882   VPAAEKQPKKAQK-PRLVAH-QV-DDEHWEAWALVEGEEVAR-VKIKG-TRVEAMTAAWEASQKALDD
Halm    VAAAVPKEVAEAK-PRLNAHPQ--GDGRWVGVASINGVEVAR-VGNQAG-RIEAMKAAYKAAGGR
VCamp   VVETKPKDETVIK-PKMKGH-KE-
DDGTWLVDVTIEDKSWQISVGKEPKNVMDAFKLAWNEFvyrkalpekp*
                    * :   ...      *  : . .    : .       :    ..*:.  ::.   .
           -----------K-P^h^u--+--GDG-%-h-h-h-G--%-------u---h'Ah+-A%^-%-------
                        i^ie   e  lll i i i i l  i        α    αiiα i α
```

D. PY54 secondary structure prediction



other known DNA-binding dsRBDs in the presence of an extended loop between the third strand of the β-sheet and the terminal alpha helix. The solved structure of the DBD DNA complex shows that the extended loop inserts in a region just outside that recognized by the dsRBD's β-sheet, which binds across the sequence 5'-GTAGTA-3'. Four bases within the cognate DNA oligomer are specifically recognized by externally facing-residues from the three-stranded β-sheet (see Figure 2.15). Lys28 recognizes two base-steps through its side-chain amine group, donating a hydrogen bond to G3's O6 and N7 atoms and to T4's O4 atom. Nucleotide substitutions resulting in the loss of either of these contacts (such as G3A or T4C) causes a tenfold reduction in Tn916-Int-DBD-DNA affinity levels. Tyr40 also recognizes two base-steps through its side-chain, with the hydroxyl group accepting a hydrogen bond from C21's N4 amine group, and donating a hydrogen bond to A20's N7 atom. Nucleotide substitutions that disrupt this interaction by eliminating either the N4 amine group of C21 (G6A/C21T) or potential contacts to the purine ring at base-step 7 (T7G/A20C) cause 26 and 16-fold reductions in respective affinity levels.

Although the precise location of the binding partner of the putative PY54 protelomerase farCTD cannot be definitively determined without further experimentation, the PY54 palindromic DNA recognition site provides an indication of one possibility. The DNA sequence recognized by the dsRBD-like portion of the Tn916-Int-DBD is also present in the palindromic PY54 protelomerase binding site. Just downstream of the PY54 palindrome's central six base-pairs (the last of which is a 3'-G) is the sequence 5'-TAGTA-3', while just upstream of the central palindrome is an imperfect match (5'-TAGTC-3' for one strand and 5'-CAGTA-3' for the other.) (see

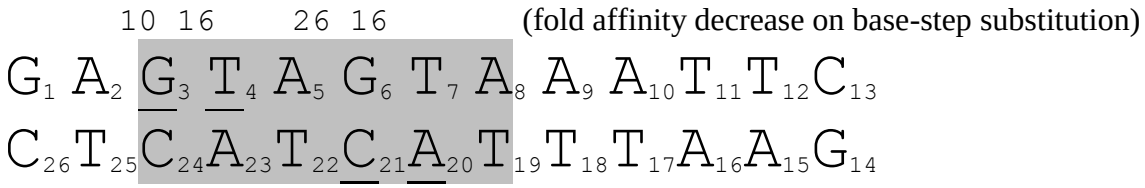
Figure 2.15 Effects of base-step substitution of Tn916 Int-DBD cognate DNA on complex affinity, and comparison to PY54 protelomerase target site.

A) The sequence of the oligomer used in both NMR structural studies and in the intrinsic fluorescence quenching studies of the Tn916 Int-DBD-DNA complex reported by Connolly et al. (*112*). The highlighted portion indicates the region of the oligomer in which the dsRBD binds, excluding the interaction of the extended loop between strand β3 and the alpha helix. Bases which are specifically recognized by side-chains of externally facing residues of the β-sheet are underlined.

B) As reported by Connolly et al., individual base-step substitutions within the oligomer are shown and their corresponding effect on Tn916-Int-DBD binding affinity is indicated. The decrease in binding affinity is reported as fold decrease compared to the wild-type oligomer, and is listed above each base-step substitution shown. The Tn916-Int-DBD residue that interacts with the wild-type base is shown above or below the substituted base, along with the percentage of conformers from the solved structure that were seen to interact with the wild-type base. A description of the wild-type residue/base interaction is shown below each substituted oligomer.

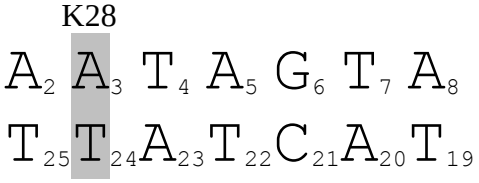
C) Protelomerase PY54 target site. The palindromic region (excepting the central six base-pairs) is underlined. Regions with sequence similarity to the dsRBD-binding portion of the Tn916-Int-DBD cognate site are highlighted.

A. Oligomer used in Tn916-Int-DBD studies



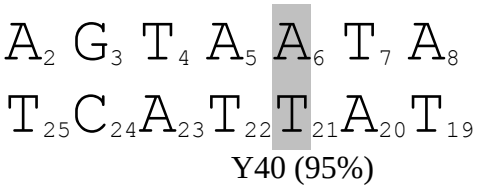
B. Base-step substitutions and their effect on Tn916-Int-DBD affinity.

G3A substitution （10 fold reduction）



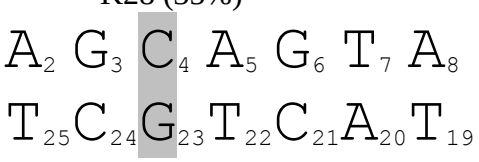
K28's amine group donates an H bond to G3's O6 (75%) and N7 (65%)

C21T substitution (26 fold reduction)



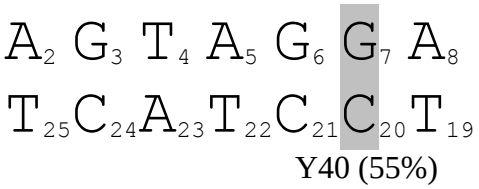
Y40's hydroxyl accepts an H bond from C21's N4 amine group

T4C substitution (16 fold reduction)



(K28's amine group donates an H bond to T4's O4 atom)

A20C substitution (16 fold reduction)



Y40's hydroxyl donates an H bond to A20's N7 atom

C. Py54 protelomerase palindromic target site (*51*)

5′-A**TAGTC**ACCTATTTCAGCATACTACGC-GCG**TAGTA**TGCTGAAATAGGTTACTGT
3′-TATCAGTGGATAAAGTCGT**ATGAT**GCG-CGCATCATACGACTTTATCCA**ATGAC**A

Figure 2.15).

The homologous β-sheet residues for other known and putative protelomerase farCTDs do not match any of the other known DNA-binding dsRBDs, so their likely cognate DNA binding sequences cannot be as easily identified. However, the pattern of a 5-6 base sequence just downstream of the central six base-pairs of the palindromic protelomerase target site being imperfectly repeated at the far 5'-end of the palindrome is also present in the other known protelomerase binding sites (see Figure 2.16). This sequence pattern of the palindromic protelomerase binding site provides an intriguing possibility for future studies seeking to identify the binding partner of the protelomerase farCTD in an effort to elucidate the function of this interesting domain.

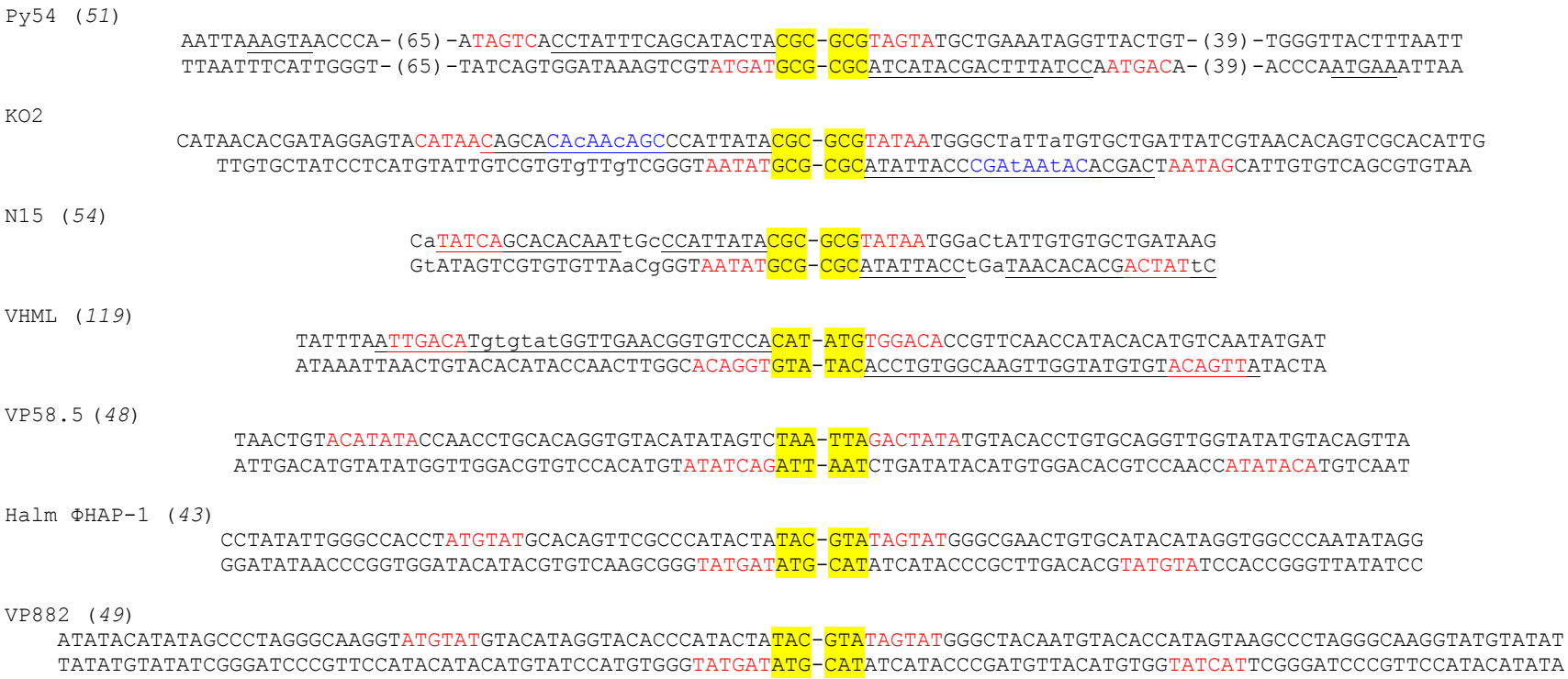


Figure 2.16 Comparsion of the binding sites of known phage protelomerases.

The pattern of a 5-6 base sequence just downstream of the central six base pairs of the palindrome and an imperfect match at the 5' end of the palindrome is repeated for all known and putative protelomerase target sites. Blue text on the Tel-KO2 target site indicates the region of naked DNA. Red text indicates the 5-6 pair sequences which may be the site of farCTD binding. Lowercase font indicates an imperfect section of a palindromic target site.

CHAPTER 3

CONCLUSION

Summary

This dissertation has focused on understanding the structure and function of the far C-terminal region of phage protelomerases. Our work has identified the structured portion of the far C-terminal region of two phage protelomerases and has shown that the regions form a well-defined, homologous domain. The domain has been called the far C-terminal domain (farCTD). It is composed of a contiguous, three-stranded, anti-parallel β-sheet which wraps around and binds against a C-terminal alpha helix. Sequence alignment and secondary structure predictions show that all known and putative phage protelomerases contain C-terminal regions which will almost certainly form homologous domains. The farCTD belongs to the structural superfamily known as the double-stranded RNA-binding domain (dsRBD), although the protelomerase farCTD differs from the canonical dsRBD in lacking an additional N-terminal helix. This modified architecture (βββα) is identical to that of a small subgroup of the dsRBD superfamily which is known to bind DNA in a sequence specific manner through the insertion of the three-stranded β-sheet into the DNA major groove.

Future directions

The structure of the farCTD paves the way for a more comprehensive understanding of its role in the functioning of the protelomerase enzyme. The mechanism of hairipin telomere formation is beginning to be understood, and a crystal structure of the *Klebsiella oxytoca* phage ΦKO2 protelomerase core domain lacking the farCTD has been reported, but a number of questions remain, and it is likely that an understanding of the farCTD will be closely interconnected in answering these questions (*29, 64*).

What is the binding partner of the farCTD?

In determining the function of the farCTD, it seems unlikely that the farCTD would be able to perform any kind of enzymatic function, especially as an isolated domain. Given the catalytic studies and the precedence of the function of other dsRBDs, it is far more likely that the role of the farCTD is to bind to and interact with other macromolecules. Although initial binding studies were inconclusive, sequence and structure comparisons of the farCTD and full-length protelomerase with other dsRBD-containing proteins indicate that the protelomerase farCTD likely interacts with DNA.

Compared with consensuses of DNA and RNA binding dsRBDs, the protelomerase farCTD is most like DNA-binders in terms of overall architecture, loop and strand length, and presence and absence of sequence conservation in key regions. Specifically, the farCTDs mimic the βββα topology, shorter loops, and more variable strand length, helix length, and helix angle of the DNA binding dsRBDs. Sequence comparisons of the solved and putative farCTDs with a consensus of DNA and RNA binders also show that the farCTD is more closely related to DNA binding dsRBDs.

Significant conservation exists in the three RNA interacting regions of RNA-binding dsRBDs, and none of the conserved residues within these motifs are conserved in the farCTDs. For DNA-binding dsRBDs, there is a significant lack of conservation among important DNA-recognition regions (the external face of the β-sheet) reflecting the ability of DNA-binders to base-specifically recognize short DNA sequences. Instead, there is conservation in regions poised to interact with the DNA backbone, including positively charged and polar residues prior to and at the end of strand β1. When aligned by the hydrophobic residues conserved among all three dsRBD classes, solved and putative farCTDs follow the paradigm described by DNA-binding dsRBDs. There is a lack of conservation among the residues of the β-strand likely to be externally facing, while the positive residues which in the DNA-binders are poised to interact with the DNA backbone are universally conserved among the putative farCTDs. While it is difficult to predict with absolute certainty which residues of the strands of putative farCTDs will be externally facing due to the possibilities of β-bulges and other strand anomalies, there seems to be very limited conservation for farCTDs in these regions.

The homology of full-length protelomerase to its closest relatives also suggests that the farCTD binds DNA. Protelomerases share sequence homology and a catalytic mechanism with type IB topoisomerases and the site-specific recombinase family known as tyrosine recombinases (*22*, *27*, *28*). The archetypal member of the tyrosine recombinase family is λ-integrase from the temperate coliphage λ. Like the solved protelomerase Tel-KO2, λ-integrase has a multi-lobed central region and a small domain attached by a flexible liker. In λ-integrase, this N-terminal domain, a member of the DNA-binding family of dsRBDs, sequence specifically binds DNA at sites distal to the

site of recombination and shapes the recombination complex via interaction with other N-domains in the reactive multimer (*60*). The close relationship of λ-integrase to protelomerase in terms of sequence conservation, mechanism, and domain architecture suggests that the dsRBDs of protelomerase and λ-integrase may serve a related function. The precedent of the function of the λ-integrase N domain not only suggests that the protelomerase farCTD binds DNA, but provides a precedent for a second, protein based interaction in its role. Several experiments suggest themselves in determining whether this theory is correct for the farCTDs. Initial farCTD/DNA binding site studies were inconclusive. However, it may be that the removal of the N-terminal region of the farCTD reduced DNA binding, as observed in the N-domain of phage λ-integrase, in which removal of only two of the unstructured N-terminal residues virtually abolished arm-type binding and recombinase activity (*120*). A recent solution structure of the λ integrase ARM-binding domain has shown that the previously described "unstructured" N-terminal tail in fact wraps around the DNA strand and inserts in the minor groove, providing additional stabilizing DNA contacts (*113*). Since our DNA affinity experiment used the shortened protein construct Tel-KO2 (593-640), the unstructured region was entirely absent.

A number of possibilities exist in testing this hypothesis. An obvious starting place is to repeat the previously attempted experiments which were inconclusive. At least one BIACORE experiment using a much shortened DNA protelomerase target site was performed by the Huang lab using the TelKO2 protelomerase, but due to proprietary interests, the experiment set up and results could not be described. Since that time, structures of both the core domain and the farCTD have been reported, and the increased

understanding of the Tel-KO2 will aid in experiment design and success. Once the likely binding site has been confirmed, chemical shift mapping experiments can be repeated with optimized experimental conditions and procedures. Initial attempts could follow the successful procedure used to map chemical shifts of λ-integrase N-domain (1-64) with and without a 14-basepair oligomer containing the P'1 recognition site ((*120, 121*) for descriptions and protocols.)

While the kinetic effect of protelomerase and target site truncation data in the KO2 system implies that the farCTD binds within the DNA palindrome, it may also serve another role in the phage life cycle. The known protelomerase-containing phages are temperate phages, and replicate stably within their DNA hosts as linear plasmid prophages during their lysogenic phases. However, the phages also have a lytic phase, and exist as circularly permuted phage DNA with overhanging, cohesive ends. The conversion of the genome from phage to prophage form upon infection has been described. The phage genome circularizes via the cohesive ends and is then cleaved by protelomerase to form a linear, hairpin-capped molecule. The conversion from linear prophage to circularly-permuted DNA suitable for loading into phage heads within the lytic cycle is less well understood. A recent study of the course of lytic development within the *E. coli*/N15 system showed that protelomerase is required for both plasmid replication and lytic development (*122*). The authors present data consistent with a model in which the differential processing of the head-to-head circular genomic dimer known to be the intermediate of normal prophage replication would result in monomeric circular molecules. The authors suggest that if protelomerase molecules processed a replicated dimer through a synaptic complex in which the protelomere dimer junctions

were held in close proximity and allowed to act in concert, the reaction mechanism could result in unit-length genome monomers by the switching of religation partners of the complementary protelomere junctions. This model is reminiscent of the integration/excision of the phage lambda genome through the function of the lambda integrase enzyme. Protelomerases could follow the lambda integrase prototype of bivalent binding in which the dsRBD (farCTD in protelomerases) binds outside of the core recognition site to aid in strand exchange, potentially in the flanking inverted repeat sites which have been identified for most protelomerase-containing phages (*22, 48, 54*).

How does the farCTD enhance the creation of hairpin telomeres?

Although the core portion of protelomerase Tel-KO2 (consisting of the N-terminal, catalytic, and C-terminal domains) is sufficient to produce hairpin telomeres, prior studies have shown that the presence of the farCTD has a substantial effect on the efficiency of the protelomerase. When acting on full-length DNA binding site, Tel-KO2 constructs lacking the farCTD are only able to reach partial completion of hairpin resolution compared to full-length protelomerase, even at a 15X longer reaction time. Despite this observation, the precise role of the farCTD in hairpin production is unknown. The domain may function to increase DNA binding affinity or specificity. Alternatively, it could bind to DNA in such a way that modifies the shape of the DNA, perhaps facilitating unwinding of coiled DNA, allowing more room for rearrangement of the hairpin-forming central region, or possibly promoting dissociation of the core domains from each other or from the DNA itself.

How does the protelomerase dimer dissociate?

The mechanism and timing of protelomerase dimer dissociation is unknown, especially with regard to hairpin formation. The reported Tel-KO2 structures contained variations of suicide substrates so that hairpin formation was prohibited and the enzyme was trapped in an intermediate state. These structures included a nicked palindromic DNA substrate with an orthovanadate ($VO4^{3-}$) to mimic the scissile phosphate of a pentavalent transition state of DNA cleavage, and a covalent TelKO2-DNA complex, representing the phosphotyrosine intermediate following DNA cleavage. As a result, less is known about the interaction of the protelomerase enzyme with the final hairpin DNA product. A modeling experiment was conducted to roughly determine whether the Tel-KO2 dimer observed in the crystal structures could accommodate juxtaposed hairpin ends. Several previously solved tetrapeptide DNA hairpins were modeled into the approximate location of the original DNA substrate within the Tel-KO2 dimer to see if the structure could accommodate covalently formed hairpin ends. Of the several solved tetrapeptide DNA hairpins used, only one (pdb accession code 1qe7) fit the observed TelKO2 crystal structure conformation without steric clash ((*64, 123*), supplemental information.) The authors suggest hairpin formation could occur spontaneously following DNA cleavage through favorable entropic shift, and that the resulting steric effects could trigger hairpin dissociation. It is also possible that steric restrictions disfavor hairpin formation until after Tel-KO2 dimer dissociation. An intriguing possibility for the role of the farCTD exists in the possible DNA binding site 3' of the central six base pairs of the protelomerase core target sequence. Binding in this location, the farCTD could serve to disrupt the binding of the catalytic region of the protelomerase

to the target DNA either to create room allowing hairpin formation to occur, or to disrupt dimer association after hairpins have formed.

What is the fate of protelomerase once hairpin product has formed?

A recent single-molecule study suggests that λ integrase forms a stable product complex that doesn't readily disassemble (*124*). Likewise, it has been suggested that protelomerases may also remain bound to the resolved chromosome termini, which could help to protect chromosome ends (*64*). Resolution of duplicated telomere sequences by full-length Tel-KO2 has been shown to be remarkably efficient (*64*). However, Tel-KO2 has a low reported enzyme turnover activity. In fact, published studies of purified Tel-KO2 acting on binding site-containing substrate in *in vitro* assays indicate such slow turnover so as to indicate stoichiometric activity. The authors of the study suggested that the enzyme may become inactive at the end of each round of activity, or else require other cellular factors in order to act catalytically (*29*) The low turnover does not appear to be the result of tight end-product binding, however, as purified DNA hairpin does not act as a competitor for the hairpin creation reaction nor, the authors reported, did it appear to be caused by the enzyme becoming unstable during reaction conditions (reported, data not shown.) It is interesting to note that in unpublished studies by the Huang lab, TelKO2 protelomerase lacking the farCTD was sufficient to produce hairpin, but the formation was not only much slower, but seemed unable to process all substrate. Under similar conditions, hairpin formation for the native enzyme is complete in two minutes while the shortened enzyme results in only 50% conversion after 30 minutes. Absence of the farCTD significantly inhibited reaction progress and completion, and may

suggest that the farCTD aids in enzyme turnover. Understanding the function of the farCTD will likely help answer the question of protelomerase turnover.

Like Tel-KO2, the *Borrelia burgdorferi* protelomerase (often called ResT) does not appear to turn over during in vitro analysis. In trying to determine whether the slow turnover was due to slow product release or product inhibition of reaction, it was discovered that ResT can fuse hairpin products to form double-stranded DNA (*125*). Interestingly, the authors state that the properties of hairpin fusion suggest that the reaction reversal does not occur after an initial hairpin creation reaction while the enzyme lingers in a post-resolution synaptic complex, but rather that fusion occurs independently, or after product release from a forward reaction (*126*). The details of this reverse reaction in context of the low turnover of the enzyme will be informative to the functioning of farCTD-containing protelomerases. Unlike the *B. burgdorferi* protelomerase (ResT), Tel-KO2 telomere end fusion activity has not been observed, and is not detectable *in vitro* even under high hairpin product concentration (*64, 127*). Secondary structure prediction indicates that *B. burgdorferi* does not likely have an extended linker/farCTD. It may be that the farCTD interacts with hairpin product in a manner to inhibit or prevent this reverse reaction.

How do the distal DNA target site ends contribute to enzyme efficiency?

The crystal structures of the Tel-KO2 involved both a shortened protein construct (Tel-KO2 4-535, lacking the flexible region and the farCTD) and a shortened DNA oligomer (the central 44 basepairs, lacking the six most distal base pairs on either side.) The protein extends only to the distal ends of the shortened DNA, with the C-most

regions folding back toward the center of the oligomer. The structure implies that the core domain does not contact the most distal basepairs, and does not explain how the distal DNA ends contribute to protelomerase activity. Unpublished data by the Huang group has shown that shortening an oligomer containing the Tel-KO2 protelomerase DNA target site by only three base pairs on either side of the recognition sequence (resulting in a 50 vs. 56 basepair recognition site) reduces the activity of full-length protelomerase (as measured by creation of hairpin protelomeres) to 25% compared with activity on full-length DNA target. Further shortening the DNA target has an even more drastic effect on enzyme activity. A 42 basepair oligomer (only one basepair longer on either side than the oligomer in the crystal structure) reduces enzyme activity to less than 1% compared to activity on full-length DNA target. Interestingly, a shortened protein (Tel-KO2 1-531) partially restores enzyme activity on shortened 42 base-pair target sites, returning activity to 15% of that of wildtype enzyme acting on full-length target. Combined, these two unpublished experiments suggest that the farCTD may interact with the distal DNA ends.

Integrase lambda shows a similar pattern of influencing enzyme activity. Weak core-site binding of lambda integrase relies on the high-affinity arm-DNA binding of the N-domain for proper delivery of the core (catalytic) domain to its recognition sequence. However, the absence of accessory (arm) DNA inhibits core site binding compared to truncated lambda integrase mutants lacking the N-domain. Addition of accessory (arm) DNA reverses the intrinsic inhibition of core-binding by the N-domain (*128*).

What correlation do phage protelomerase farCTDs have to bacterial protelomerases?

The two bacterial protelomerases are shorter than the phage protelomerases. Most automated alignments do not show an obvious homology of the C-terminal region of either bacterial protelomerase to the farCTD of phage protelomerases. However, the existing homology between phage protelomerase farCTDs is also sufficiently limited that automated alignment programs were unable to properly correlate them. While no experimental data exists, the alignment data for *Borrelia* protelomerases with phage protelomerase farCTDs is compelling. At this time, no structural data has been obtained for any protelomerase within the *Borrelia* family (usually called resolvase, or resT.) Whether or not the farCTD correlates specifically with the *Borrelia* protelomerases C-terminal domain, a greater comprehension of protelomerase function in general could provide insight into possible future treatments for a number of significant human pathogens, most notably Lyme disease, a currently chronic and debilitating condition which is one of the fastest growing infectious diseases in both the United States and Europe.

What scientific interest do phage protelomerases hold?

Recently, a genetic engineering technique was described in which addition of an Tel-N15 recognition site and a to a normally circular plasmid results in a stably replicating linear construct within a Tel-N15 expressing host strain. The resulting linear plasmid's hairpin ends are rendered safe from RecBCD nucleases both *in vitro* and *in vivo (18).* The authors suggest that this new technique will be especially useful in a number of applications, including creating large linear constructs, assembling and

studying artificial chromosomes, and as a means to clone and study large linear viral genomes.

Aside from technological application, the phage protelomerases hold significant scientific interest of their own. Tel-VHML, for example, comes from the phage *Vibrio harveyi* myovirus like, which infects *Vibrio harveyi*, a free-living, gram-negative, bioluminescent marine bacterium of the Vibrionaceae family of proteobacteria. *V. harveyi* is related to *Vibrio fischeri*, a symbiotic bioluminescent bacterium that colonizes the light organs of various marine organisms such as squid (*129, 130*). *Vibrio* classes also include *cholerae*, a human pathogen which causes cholera (*131*). *V. harveyi* was originally described in 1936 and is named after Edmund Newton Harvey, a pioneer in the study of bioluminescence (*132*). The species is ubiquitous throughout marine environments. It contains at least ten different chitinases, and may be important in nutrient cycling in aquatic environments (*131, 133*). *V. harveyi* is thought to be responsible for the milky seas effect, the widely observed but not fully understood phenomenon described by Jules Verne in his classic 1870 novel Twenty Thousand Leagues under the Sea (*134*). In this phenomenon, stretches of surface ocean water large enough to be observed from space become luminescent for up to days at a time (*135*). The currently favored hypothesis explaining the milky seas effect is that *Vibrio harveyi* associate with marine algae during algal blooms, allowing colony density great enough to allow the quorum sensing which regulates bioluminescence (*136, 137*).

Almost all strains of *Vibrio harveyi* are non-pathogenic, but some have been reported to cause devastating outbreaks of disease among populations of marine invertebrates and vertebrates, causing a variety of disease states including infectious

necrotizing enteritis, eye lesions, skin ulcers, and vasculitis ((*47*), and the references therein). Disease outbreaks have been particularly devastating among invertebrate marine aquacultures, with dramatic financial loss due to the high mortality rate (nearly 100% with some strains) of infected animals, and the persistence and survival of virulent strains, which has been attributed to the resistance to antibiotics and disinfectants of the *V. harveyi* biofilms (*46*). The disease, called vibriosis or bacterial luminosis, has been reported in prawn hatcheries, rock and spiny lobster aquacultures, sea cucumber aquacultures, and in pearl oysters ((*47, 138, 139*), and the references therein) and has most recently been associated with the rapid tissue necrosis of stony corals (*140*). *Vibrio harveyi* pathogenicity has been recognized as a substantial threat, and a significant ongoing research effort is currently underway to attempt to ameliorate vibrio pathogenesis (*139, 141-151*)

Infection with phage VHML has been heavily implicated in *V. Harveyi* virulence. VHML conferred virulence on previously naïve, avirulent strains with accompanying alteration of colonial morphology, haemolysis upregulation, and production of previously absent extracellular proteins antigenically similar to toxin components produced by virulent strains (*45*). Later, infection of Vibrio harveyi by VHML was shown to cause changes to its phenotypic profile (*152*).

The conferral of virulence by a phage of a previously naïve bacterial species is not unknown. Cholera in humans is caused by the lysogenic conversion of nontoxigenic *Vibrio cholera* or *Vibrio mimicus* strains by a phage (CTXΦ) (*153, 154*). Although *V. harveyi* is not usually characterized as a human pathogen, the medical journal *Pediatric Blood & Cancer* recently reported a pediatric oncology patient with central line sepsis

caused by a *Vibrio harveyi* infection, most likely contracted from swimming in the Mediterranean sea while on vacation (*155*). Other current research examines the possibility that some *Vibrio harveyi* strains may indeed be pathogenic to humans on exposure to bathing water, as well as by consumption of infected seafood (*156*).

Likewise, VP58.5, a 42-kb plasmid myovirus prophage has been associated with the serovar O3:K6 pandemic clonal complex (*157*). This complex is responsible for a world-wide pandemic of seafood borne illness (*158-162*). Interestingly, although the VP58.5 and VHML phages have very similar genetic profiles (gene sequence is 92% similar) they infect a different range of hosts (*48*).

Conclusion

As we come to understand the role of the farCTD in relation to the rest of the protelomerase, we will better understand how the enzyme itself functions. A more clear understanding of the protelomere-formation mechanism will likely have far-reaching effects in agriculture, aquaculture, biotechnology, and human health.

APPENDIX A

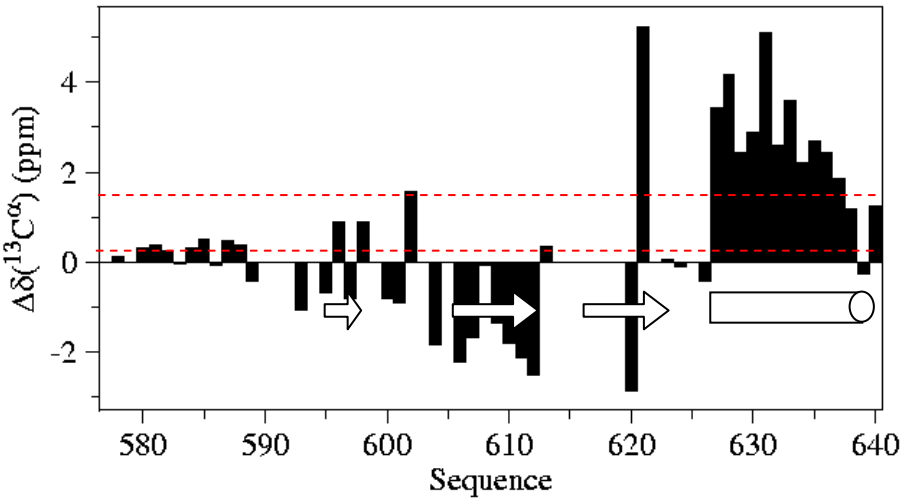
COMPARISON OF THE SECONDARY STRUCTURE PREDICTION AND $^{13}C\alpha$ CHEMICAL SHIFT INDICES FOR THE TEL-KO2 AND TEL-VHML FAR-CTDS

Figure A.1. Secondary structure prediction and comparative $^{13}C\alpha$ chemical shift indices support the solved structures of the farCTDs of Tel-KO2 and Tel-VHML.

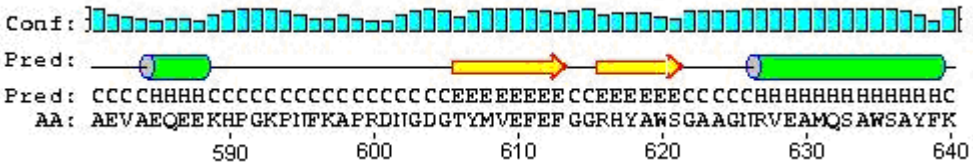
A) and C) Graphical representation of secondary structure propensity based on the raw Cα chemical shift differences (Δppm) for Tel-KO2 (531-640) and Tel-VHML (450-509), respectively. Dotted lines are positioned at 0.7 and -0.7. Values > 0.7 indicate a propensity for helical conformation; values at $-0.7 \leq 0 \leq 0.7$ indicate a propensity for random coil; values < -0.7 indicate a propensity for β-strand or extended conformation. In both cases, a cartoon indicates the secondary structure as determined by NOE structure calculations: arrows indicate extended or β-strand regions, while cylinders indicate α-helices.

B) and D) Graphical representation of secondary structure prediction based on of the farCTDs of the Tel-KO2 and Tel-VHML protelomerases. As indicated, cylinders above the displayed sequence indicate that the sequence is predicted to form an alpha helix, while arrows indicate a prediction for extended or β-strand conformation. A line indicates a prediction for random coil. The height of the bar above the graphical prediction indicator displays the relative confidence of the predictions. Predictions made using PSIPRED program (*82, 83*). Although the program predicts two β-strands for each farCTD, it identifies different strands (β2 and β3 for TelKO2 farCTD and β1 and β2 for VHML farCTD.)

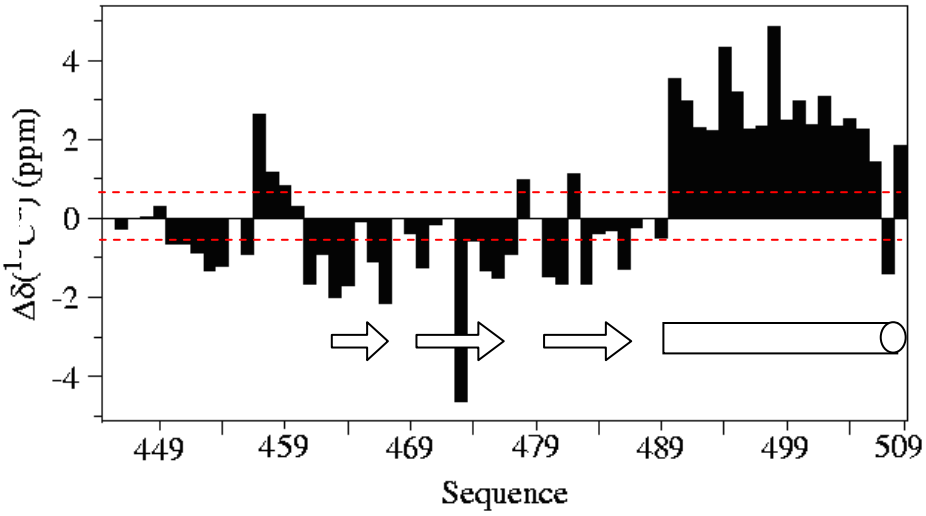
A.


4
2
0
-2
580
590
600
610
620
630
640
Sequence

B.


Conf:
Pred:
Pred: CCCCHHHHCCCCCCCCCCCCCCCCEEEEEEEECCEEEEEECCCCCHHHHHHHHHHHHHC
AA: AEVAEQEEKHPGKPIFKAPRDIIGDGTYMVEFEFGGRHYAWSGAAGIRVEAMQSAWSAYFK
590
600
610
620
630
640

C.


4
2
0
-2
-4
449
459
469
479
489
499
509
Sequence

D.


Conf:
Pred:
Pred: CCCCCCCCCCCCCEEEECCCCCEEEEEEECCCCCCCCCCCCHHHHHHHHHHHHHHHHCC
AA: DQKVSWPKAKDIKVQSKKEGDMWHVWTEVIGMRWEIWSKGRKTEAVKALRQQYERESAEM
459
469
479
489
499
509

APPENDIX B

LITERATURE REVIEW OF THE ABILITY OF DSRBDS TO RECOGNIZE SEQUENCE SPECIFICTY IN DOUBLE-STRANDED RNA

APPENDIX B

SPECIFICITY OF RNA-BINDING DSRBDS

Introduction

As the name suggests, the protein fold known as the double-stranded RNA binding domain (dsRBD) was first recognized by its ability to recognize and bind double-strand RNA (dsRNA) (*93*). Proteins containing RNA-binding dsRBDs have been found in proteins throughout eukaryotic, bacterial, and viral systems, and in one known example of archaea. RNA-binding dsRBDs are involved in many different aspects of RNA function including RNA localization, posttranscriptional gene silencing, RNA editing, and translational repression. Despite the varying roles these proteins, the domains have in common the ability to specifically recognize and bind dsRNA over single stranded RNA, DNA, or RNA/DNA hybrids.

The binding specificity debate

Beyond a preference for dsRNA, the question of binding specificity by the dsRBD is one of current debate and ongoing research. Although most data suggests that dsRBDs do not impart binding specificity, increasing numbers of studies imply degrees of binding preference in some systems. While most dsRBDs bind most dsRNA, it appears some dsRBDs bind dsRNA more tightly, with varying exclusivity. Thus, the answer to the binding specificity debate will likely be multi-tiered, and the question must

be addressed diverging but related aspects: RNA sequence specificity, RNA helix geometry specificity, and RNA "secondary structure" specificity.

Sequence specificity

It is generally believed that individual dsRBDs do not bind with sequence specificity. Solved dsRBD/dsRNA complexes show that the majority of minor groove/base contacts are water-mediated, and those that are not are to hydrogen bond-acceptors present in all four bases, leaving the authors of a comprehensive 2004 review to conclude that there is no strong evidence to indicate that interactions by any known dsRBM are sequence specific (*94*). The ability to recognize a broad range of dsRNA targets without a strict target sequence requirement would be beneficial in the roles played by some dsRBD-containing proteins. In proteins such as protein kinase R (PKR), which plays a major role in the cellular response to viral infection on activation by binding to dsRNA, or Dicer, an RNase III family member that cleaves dsRNA and pre-microRNA (miRNA) into short dsRNA fragments called small interfering RNA (siRNA), an ability to recognize dsRNA targets without a strict sequence requirement would allow a greater range in target selection and resultingly, scope of enzymatic activity. Even among proteins which ubiquitously recognize dsRNA, though, there are examples of sequence preference. Although these preferences may frequently be a result of anti-determinants, at least one example exists of a sequence determinant for dsRBD binding.

A recent study seeking to map the determinants and anti-determinants of *Escherichia coli* RNase III binding to a minimal substrate found that an AU base pair within a specific region of substrate RNA hairpin acts as a determinant for enzyme activity. In this example, the carboxamide group within a conserved glutamine in the

first α-helix of the dsRBD recognizes a uracil O2 atom in the canonical AU pair and a neighboring ribose 2'-hydroxyl group. Substitution with a CG or CI pair reduced binding (*163*). It is especially interesting that this example of substrate differentiation was discovered in a system which had generally been considered to lack target discrimination entirely (*164*).

Although no other examples of direct sequence requirement are currently known, many dsRBD-containing proteins have target preferences. For example, using site-specific affinity cleavage, Spanggord and Beal showed that dsRBD1 of PKR bound with a specific orientation on both a synthetic stem-loop RNA and VA_I RNA, an adenoviral inhibitor of PKR (*165*). Further tests on the synthetic stem-loop showed that binding was unaffected by disruptions in the helix caused by nucleotide mismatch and bulging. Other examples exist of multiple dsRBDs within a protein binding with specific orientation, or different binding modes on a dsRNA target (*92, 166*). In this and other examples, it is likely that this preference for substrates comes in part from the differing helix geometry imparted by sequence.

Helix geometry specificity

Ideal A-form geometry of dsRNA is rare. Bends, twists, and kinks within a dsRNA are inherent to a given sequence, as is the degree of flexibility of the helix itself. Because these "flaws" within a helix result in altered groove geometry, dsRNAs of different sequence may be more or less easily accessible to a given dsRBD, or allow a greater or lesser degree of induced fit on binding, all of which could contribute to dsRBD/dsRNA affinity. Additionally, although the sequence of the dsRBD is relatively highly conserved, sequence differences do exist among recognized dsRBDs and these

differences in sequence ultimately result in a greater or lesser capacity for binding dsRNAs.

Secondary structure specificity

Although observed examples of sequence specificity by dsRBDs to dsRNA molecules are limited, preference to various RNA structural features has been observed. A 2006 study by Stefl and Beal allowed the development of an NMR-based model of the complex of the two dsRBMs of rat ADAR2 with the R/G loop, a well-conserved 70 nucletotide stem-loop from the R/G site of the GluR-B pre-mRNA, which is often used as a model system for A-to-I editing. In this study, chemical shift perturbation allowed the identification of the protein and RNA surfaces involved in complex formation, and showed that the individual dsRBMs bound separate and specific portions of the target RNA. Purified, separate dsRBM1 recognized a conserved pentaloop, while dsRBM2 recognized two bulged bases adjacent to the editing site, demonstrating structure-dependent recognition by the ADAR2 dsRBMs. Rnt1p endoribonuclease, the *Saccharomyces cerevisiae* homolog of bacterial RNAse III, specifically recognizes and cleaves RNAs with short (12-14 base pairs), irregular stem-loops capped by AGNN tetraloops that are interrupted by internal loops and bulges. Two independent structural and biochemical studies of the Rnt1p dsRBD have showed that helix α2 binds to the tetraloop minor groove of either the 5' terminal hairpin of the snR47 precursor, one of its small nucleolar RNA substrates (*167*), or to various short stem-loop RNAs that mimic physiological targets (*168*). Surprisingly, although direct base interaction of the tetraloop minor groove is apparent in the solved structures, it is not to the conserved A or G. Instead, the authors believe that specificity to the tetraloop minor groove came from the

"tuning" of the angle of helix α2 by a third, carboxy-terminal helix to the specific tetraloop shape. Removal of this helix not only destabilized the protein, but removed its ability to specifically bind to short RNA hairpin substrates, although it retained the ability to bind to long dsRNA (*167, 168*). Besides the third helix, other slight structural differences were observed: helix α1 was shorter than other solved dsRBDs, and loop 1 (between helix α1 and strand β1, observed to interact with the minor groove in all solved structures) was free to penetrate more deeply. The results of both studies imply not only that subtle differences in sequence and structure can alter dsRBD binding capacity, but that interaction with other protein elements, in this case an additional carboxy-terminal helix, may help to convey specificity for an entire dsRBD-containing protein which a dsRBD alone cannot convey.

Beyond a single domain: Multiple copies of dsRBDs/other domains

As suggested by the *Saccharomyces cerevisiae* Rnt1p recognition of AGNN tetraloops, preference of binding of an isolated dsRBD versus that of an entire dsRBD-containing protein is sometimes different, and comes from elements or interactions outside of a single, standard dsRBD. It has been theorized, for example, that multiple dsRBDs within a protein may increase specificity of dsRNA targets (*97*). Within PKR, for instance, dsRBD2 and dsRBD5 may act in concert to determine the minimum PKR substrate target length (*96*). In some cases, the presence of other domains within a dsRBD-containing protein may help to increase dsRNA binding specificity. The recent structure of the catalytic *Aquifex aolicus* RNase III complex shows that the protein's catalytic domain also interacts with dsRNA, and the authors propose that there are four

main points of interaction, two from the dsRBD (loop 2 and helix α1--the minor groove-binders as described for all classical dsRBD interactions) and two from the catalytic domain. The authors theorize that the dsRBD selects and binds dsRNA, which allows the catalytic endonuclease domain to recognize sequence specificity and enact scissile bond cleavage (*111*).

While it is likely that the specificity debate will rage for some time to come, even now binding determinants and anti-determinants are already known for some dsRBDs, and certainly more will be identified in the future. Certainly, the secondary structure formation and inherent helix deformation caused by RNA sequences affect the ability of dsRBDs to bind, and the differences in individual dsRBDs also result in differential binding ability. The identification of the specifics of these binding determinants, however, is largely a subject of future research.

APPENDIX C

SUMMARY OF PROJECT ONE: HIV TRANSACTIVATION CHIMERAS

APPENDIX C

SUMMARY OF PROJECT ONE: HIV TRANSACTIVATION CHIMERAS

Introduction

During my graduate career I had the opportunity to work on two separate projects related to the topic of protein structure determination using high-field NMR spectroscopy. Both projects centered on understanding the structure of a nucleic acid-binding protein in the larger context of determining how the protein structure affects its interaction with its nucleic acid binding partner. The project for which the bulk of my dissertation work was directed was a collaboration with Professor Wai Mun Huang in the Department of Pathology at the University of Utah. This project was a structural study of a novel domain of two homologous proteins. The results of this project are described in chapters one through three of this document. My second project was a collaboration with Matija Peterlin, a Professor of Medicine, Microbiology and Immunology at the University of California, San Francisco. We sought to further understand the transactivation that occurs during replication of the HIV genome. Transactivation overcomes the stalling of the transcriptional machinery which occurs after the initiation complex has formed and has begun full genome replication for the production of new, infective virions. The process of transactivation involves the interaction of three

elements: an HIV protein called Tat, which has been previously transcribed and translated, a hijacked human regulatory protein called cyclinT1 along with its kinase partner (CDK9), and a highly structured portion of the nascent, partially transcribed HIV genome, an RNA element called TAR. Understanding the structural interaction of this trimeric complex would provide a deeper understanding of HIV transactivation. The Peterlin lab had previously published results on successful transactivation in *in vitro* studies using a chimera of shortened CyclinT1 and Tat rather than whole, individual proteins. My work in this project involved an attempt to identify an optimal chimeric construct for structure investigation, and initial studies in identifying a purification procedure for the selected chimera. Information not available to us at the start of the project and results of related studies that have since been published have shown that the minimal chimeras we studied were inadequate to describe the transactivation system and inappropriate for structural study. This appendix contains a description of my second project, including a discussion of the HIV transactivation, issues related to the project design, and an analysis of possible future directions for the project.

Background

HIV/AIDS: classification, prevalence and therapeutics

Human immunodeficiency virus (HIV), a member if the *retroviridae* family, is responsible for the illness known as Acquired immune deficiency syndrome (AIDS). Since the identification of HIV as the etiological agent of AIDS nearly 25 years ago, approximately 60 million people have become infected with HIV/AIDS, and nearly half of those have died (*169*). Currently, an estimated 33 million people are thought to be infected with HIV/AIDS (subtypes 1 and 2) according to the World Health Organization

(WHO) and the Joint United Nations Programme on HIV/AIDS (UNAIDS). A world-wide pandemic and the leading cause of death in sub-Saharan Africa, HIV infection has no known cure and can only be treated with medications meant to slow the development of immune system failure which results in life-threatening opportunistic infections. HIV replication is fundamentally error-prone due in large part to the infidelity of the HIV reverse transcriptase (RT), the enzyme which produces a double stranded DNA-form of the single-stranded (+)RNA genome packaged in mature, infective virions. The infidelity of RT leads to a high viral mutation rate (at least nine distinct HIV-1 subtypes have evolved) and developed resistance to chemotherapies is an ongoing struggle in the effort to treat those suffering from HIV/AIDS.

Because the virus uses host cell machinery during much of its life cycle, practical therapeutic targets for chemotherapy are limited. The majority of HIV medications are targeted toward inhibiting the functioning of either the Reverse Transcriptase enzyme (nucleoside or nonnucleoside analogue reverse transcriptase inhibitors (NARTIs/NRTIs and NNRTIs)) or HIV-1 protease, the viral enzyme that cleaves newly synthesized polyproteins to create the mature protein components of an infectious HIV virion. Other drug therapy targets include virus attachment, membrane fusion, cDNA integration, virion assembly, and maturation inhibitors ((*170, 171*)., (*172*) for a recent review.) Given the pandemic nature of the disease and the rapidity with which current therapeutics become ineffective, an ongoing research effort is vital in the treatment of HIV.

Drugs targeted to the transactivation stage of genome transcription are a relatively new research effort, but make an attractive target for anti-retroviral therapy for two reasons. Not only is transactivation required for efficient genome transcription (and

therefore production of new, infective virions) but two of the three members of the reactive complex involved in transactivation are viral, minimizing the potential for drug side effects caused by interaction with other host macromolecules. Research devoted to this initiative is ongoing (*173-177*). A solved structure of the interacting partners would provide valuable insight into the transactivation event and would greatly assist the efforts directed toward drug development. So far, however, the elucidation of the structure of the complex has proven difficult for a number of reasons. The goal of our project was to determine the structure of the immediate portions of the interacting members of the complex (HIV-1 Tat, human cyclin T1, HIV-1 TAR) in collaboration with and based on the studies of the Matija Peterlin group of the Department of Medicine at the University of California, San Francisco. The Peterlin group had reported a biologically active minimal chimera of HIV-1 Tat and human cyclin T1 (hCycT1) capable of overcoming the transactivation barrier of HIV-1 Tat transactivation in murine cells. A brief introduction to the transactivation components and to the work of the Peterlin group will explain the rationale behind our efforts.

HIV transactivation

Transcription of the full HIV genome in preparation for assembly of active virions proceeds in two stages. In the first, an initiation complex forms and transcription by RNA Polymerase II begins. The multimeric negative elongation factors DSIF (DRB (5,6-dichlorobenzimidazole 1-β-D-ribofuranoside)-sensitivity-inducing factor) and NELF (negative elongation factor) stall transcription of the genome near the start site, requiring an activation event to return RNA Polymerase II from an abortive to a productive complex (*178*). Activation occurs when an HIV-encoded protein called Tat binds to a

highly structured region of the long terminal repeat portion of newly transcribed RNA. This binding event recruits the binding of a human regulatory factor, pTEFb (positive transcriptional elongation factor b), an RNA Polymerase II transcriptional elongation factor which consists of the protein hCycT1 and its kinase partner, CDK9 (*179*). This three-member binding interaction results in phosphorylation events of various complex components including the negative elongation factors (*180, 181*) and the S2 position of the heptad repeats of RNA Polymerase II's carboxy terminal domain (*179*). Phosphorylation switches the initiation complex to an elongation complex, allowing complete transcription of the genome. For a review of the current understanding of this transactivation event in CD4+ T Lymphocytes and Machrophages, see (*182*). For a more general review of RNA Polymerase II transcriptional elongation, see (*183*).

Transactivation complex members

Tat. Tat is a viral regulatory protein with multiple functions and several important biological interactions. It plays a direct role in HIV induced immunodeficiency, including the promotion of viral infection by proliferation and differentiation of HIV target cells and apoptosis of T-cell lines and primary T-cells ((*184, 185*); see (*186*) for a review of other Tat-induced biological effects.) The sequence encoding Tat is comprised of two exons on different reading frames, which surround and partially overlap the genes encoding HIV's envelope gene (*187*). Mature Tat requires a double-splicing, post-transcriptional modification that results in a protein of variable length (from 86 to 104 amino acids.) While strains encoding only the first exon are fully active in transcriptional reporter assays, most infectious strains contain a longer construct, usually 101 residues in length. It has been reported that shorter, incomplete

constructs may also be biologically active (*186*). Tat was originally described as having five functional domains based on the effect of point mutations on the transactivation ability of an isoform containing 101 residues (*188*). The domains are: N domain (residues 1-20), cysteine rich (21-40) Core domain (41-48, contains a conserved Lys X Leu Gly Ile X Tyr motif), basic domain (49-57, later called arginine-rich motif (ARM)), and an enhancing auxiliary domain (58-67) (see Figure C.1). The remainder of the protein, comprised of the second exon, was deemed non-essential for transactivation. Structural (and even biochemical) studies of Tat have been confounded by the multiple cysteine residues within the cysteine-rich domain which can lead to misfolding (*189*). NMR chemical shifts and coupling constants indicate that Tat is a natively unfolded protein with only transient folding in two of the five sequence domains (*190*).

Tat performs its gene expression activation role by interacting via its basic (ARM) domain with the HIV long terminal repeat (LTR) in a region called the Transactivating Response Element (TAR). The LTR is a region of repeated sequence that flanks the functional genes in retroviruses. It is important in the integration of the virus into the host cell genome, and also acts as a primer region for viral transcription. The TAR element is especially important during the transcription of the genome in its entirety.

TAR. TAR (Transactivation Response Region) is a highly structured portion of the 5' LTR within the transcribed HIV genome. It consists of approximately 60 nucleotides that form a stable stem-loop with a 5', U-rich-bulge near the apex of the RNA stem. The TARs from the HIV-1 and 2 isoforms differ in the position of a single cytidine residue (see Figure C.2) near or within the bulge region, but are otherwise highly conserved among viral isolates. Comparative reported structures of free TAR and TAR

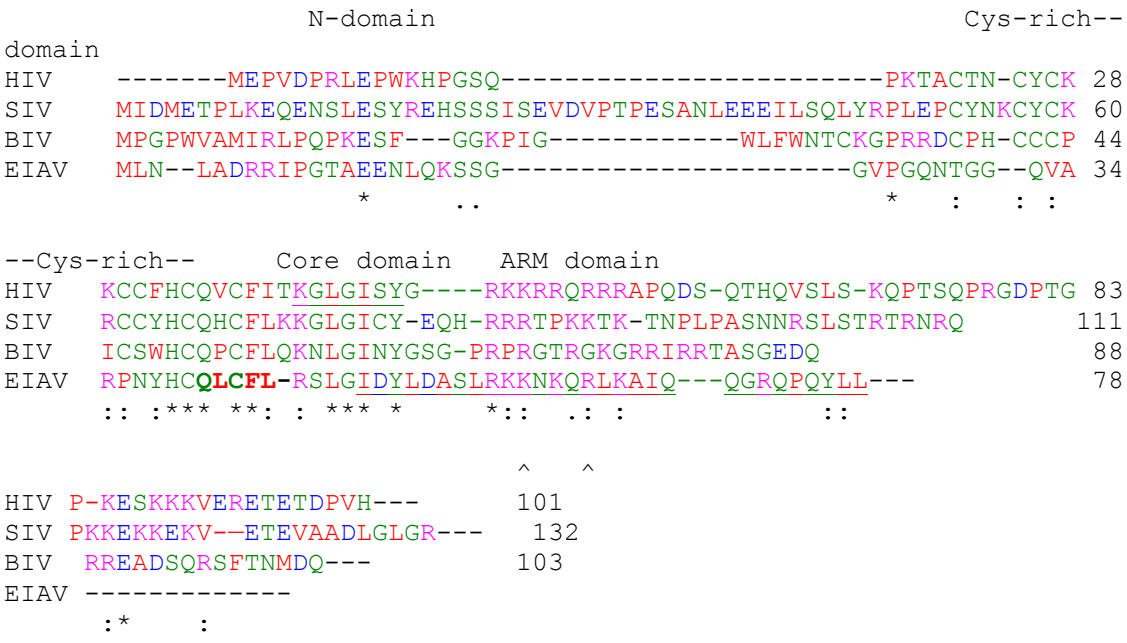


Figure C.1 Tat Sequence comparison

Sequence comparison of Tat proteins from HIV (Human Immunodeficiency Virus), SIV (Simian Immunodeficiency Virus), BIV (Bovine Immunodeficiency Virus) and EIAV (Equine Infections Anemia Virus.) The level of conservation is indicated below the alignment, with (.) indicating moderate conservation, (:) indicating high conservation, and (*) indicating universal conservation. The proteins maintain considerable conservation, especially in the regions corresponding to the HIV Core and ARM domains, with all except EIAV showing additional conservation in the region corresponding to the HIV Cysteine-rich domain.

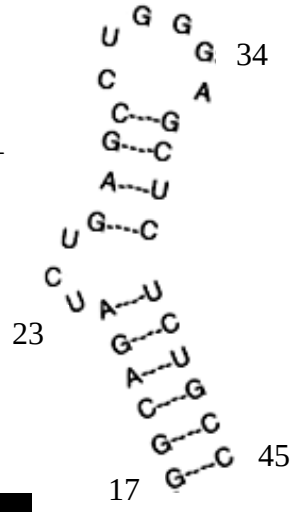
bound by small-molecule ligands, argininamide, and Tat-derived peptides show a conformational change in TAR presumably induced by binding (see Figure C.2) (*191-195*). Tat's binding position on TAR has been mapped to the U-rich bulge near the apex of the TAR RNA stem (*189*). While pTEFb cannot bind TAR on its own, the affinity of Tat for TAR is enhanced by pTEFb (*196*). The binding of hCycT1/Tat complexes to TAR requires conservation of bases in the loop of TAR not required by Tat alone, suggesting that Tat uses cooperative binding to nascent TAR to recruit hCycT1-CDK9 complexes to RNA Polymerase II (*196*).

PTEFb. Although the CDK9 portion of pTEFb is required for transactivation, it is the hCycT1 portion that appears to interact with Tat/TAR. hCycT1 is a 726 residue protein (approximately 87kDA) related to human CyclinC (196, 197). In contrast to cell cycle cyclins such as A, B, and E, which regulate cell cycle activity through their cyclin-dependent kinases, T-type cyclins mediate transcription by acting as a regulatory subunit for CDK9. Aside from a Tat-TAR recognition motif (TRM, residues 251-271), hCycT1 contains two conserved cyclin boxes (residues 1 to 250) as well as a coiled-coil region, a histidine-rich stretch, and a C-terminal PEST sequence (196). The first 272 residues are sufficient to bind to Tat/TAR in vitro and to support transactivation in vivo (196, 198). Although a shortened hCyclinT1 construct (residues 1-250) containing only the cyclin boxes can still bind to Tat, it does so more weakly than constructs that contain the TRM, and cannot transactivate HIV transcription in vivo (199). The structure of the cyclin box domain portion of hCycT1 has been solved alone (200) and with its CDK9 partner (201). The Tat/hCycT1 interaction is zinc mediated, and requires multiple Tat cysteines/histidines from within the 1-48 residue region as well as a critical cysteine

A.

```
      BULGE     LOOP
GGCUCG-UGU-AGCUCAUU-AGCUCCGAGCC     BIV
GGC-AGAUUG-AGCCCUGGGAGGUUCUCUCC     SIV-SL1
GGCCAGAU-UGAGC-CUGGGAGCUCUCUGCC     HIV2
GGC-AGAUCUGAGC-CUGGGAGCUCUCUGCC     HIV1
GCUCAGAUU------CUGCG--GUCUGAGC      EIAV
```

B.



C.

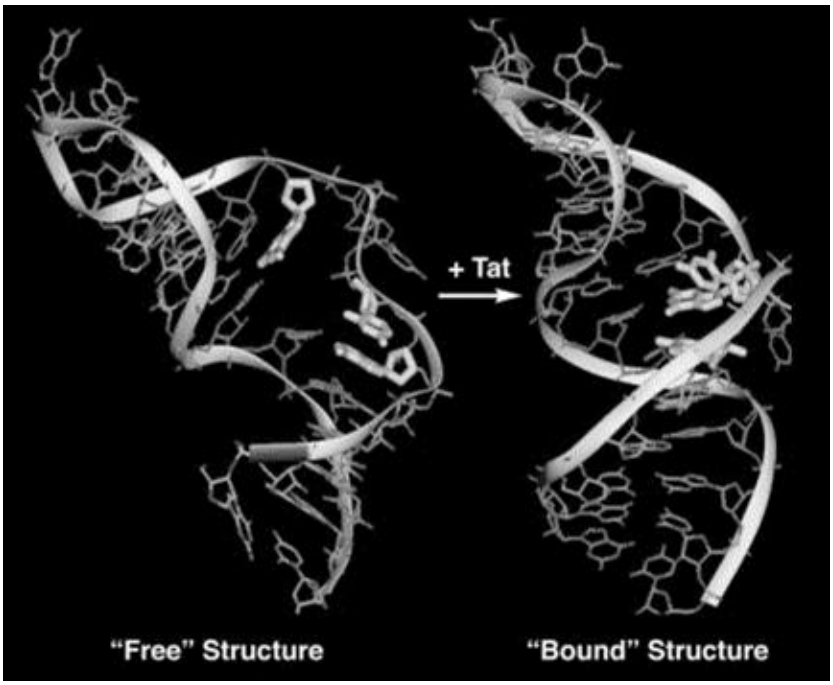


Figure C.2 HIV TAR

A. Comparative sequences of the TARs from BIV, SL1 of SIV, HIV-2, HIV-1, and EIAV. Bulge and loop regions are shown in bold type and are indicated above the alignment.

B. Secondary structure schematic of HIV-1 TAR.

C. Major groove view of free HIV-1 TAR RNA (left) and HIV-2 TAR RNA in its Tat-bound configuration (right, Tat not shown). The conformation change involves interactions between residues A22, U23 and G26 (shown as sticks) and an arginine side-chain from Tat (not shown). Reprinted from *J. Mol. Biol*. Tackling Tat, **293**, 235-254., (1999) with permission from Elsevier

residue (C261) from the TRM of hCycT1 (*202*).

Transactivation and murine cells

Multiple barriers inhibit full HIV-1 infection in murine cells, and one such barrier is the ability to produce full length genomic transcripts. Murine cells such as CHO and NIH-3T3 do not support Tat transactivation of HIV transcription (*203, 204*). Overexpression of hCycT1 in the murine cell line NIH3T3 strongly enhances transcription from an integrated proviral HIV-1 promoter, although it does not completely overcome the inability of murine cells to fully support HIV-1 infection (*202*). In murine cyclin T1 (mCycT1), position 261 is a tyrosine rather than a cysteine, and although mCycT1 can form a weak, zinc-independent complex with HIV-1 Tat, the ability of mCycT1/Tat to bind HIV-1 TAR is greatly reduced compared to hCycT1/Tat, and mCycT1 cannot support HIV Tat transactivation *in vitro* or *in vivo* (*196, 202*). The Y261C mutation of mCycT1 creates the high-affinity, zinc-dependent binding to HIV-1 Tat/TAR characteristic of hCycT1 in *in vitro* assays and supports Tat transactivation *in vivo (199, 202).* Likewise the C261A substitution in human CycT1 has been shown to abolish binding to TAR and renders the protein unable to support transactivation (*199, 205*). This variation provides an excellent system for measuring transactivation.

Defining a minimal chimera for transactivation

With the end goal of defining a minimal construct for future structural studies, Fujinaga et al. constructed a plasmid containing the HIV-1 LTR linked to the CAT reporter gene to compare the transactivation ability of a series of fusion proteins. They showed that a fusion protein of the first 280 residues of hCycT1 and full-length HIV-1_{SF2}

Tat (1-101) was capable of both binding TAR *in vitro* and transactivating transcription of the HIV LTR in murine cells at a level equal to jointly expressed unfused proteins (*205*). They further reported that a much shortened construct consisting of hCycT1 (249-280) and Tat (1-101) supported Tat transactivation, indicating that the fusion protein interacted with TAR *in vivo*, and that both hCycT1 and mCycT1 were able to supershift Tat-bound TAR in EMSA assays (*205*). Based on their recommendations, we began studies of even more truncated chimeric constructs containing shortened Tat regions and linkers. Our project goals were to express and purify the minimal chimera of hCycT1 and Tat for structural studies alone and in complex with HIV-1 TAR.

Project Results

Although we pursued purification of multiple constructs with a variety of methods of cell growth, induction, and purification schemes in an attempt to optimize yield of recombinant chimera, yield was poor, and we lost product during each of the purifications steps. Apart from the difficulties inherent to structural studies of the transactivation complex and to chimeric proteins in general, recent insights from related projects show that a well-defined structure of the minimal chimera may not be possible without the binding of TAR and CycT1, and that a construct redesign would be essential to form a structurally relevant chimeric proteint.

Purification of the cyclin T1/Tat construct was complicated by a number of factors. The construct is a minimal chimera rather than a native protein, and the core of the chimera is presumed to be the zinc-mediated interface between the proteins. Since initial studies of the purified protein showed that the chimera was underpopulated by zinc despite zinc availability during protein expression and purification, it is likely that the

two halves of the chimera are not "bound together" outside of the presence of TAR. Additionally, as a free protein, HIV-1 Tat is generally believed to be natively unstructured (*190*). Intrinsically unstructured proteins have very different characteristics than their well-folded counterparts, and often require different techniques for purification and study. As described in a recent issue of *Protein Expression and Purification*,

> Expression and purification of unstructured proteins or peptides…is not trivial. Because of the lack of well defined structure, their solubility is generally poor and it is easy to form inclusion bodies during expression or aggregates during purification. (*206*)

A recent report of a crystallographic structure of the cyclin box domain (residues 8-263) of hCycT1 on the other hand, shows that the protein, including the Tat/TAR recognition motif, is well-structured. Identification of an efficient method of expression and purification of a Tat/hCyclin T1 chimera must therefore consider the different character of its component parts, as a chimera may contain both structured (hCycT1) and unstructured (Tat) regions. Study of the HIV Tat protein is additionally complicated by its unique amino acid profile, particularly in the cysteine-rich region. Researchers who have studied both the HIV and EIAV CycT1/Tat/TAR systems noted in a recent paper that the multiple cysteines located in this short region of HIV-1 Tat make purifying the recombinant protein difficult, with complications in both expression and protein folding (*207*). Another complicating factor in purification is the arginine-rich motif (ARM) of Tat. The ARM is comprised of a number of basic residues, which presumably make the construct more soluble, but also acted to create complications. The chimeric construct binds nucleic acid extremely well, to the extent that protein quantification by calculation at A=[280] nm was impossible until after a purification step by ion exchange chromatography. Gel analysis of the elution fractions showed that ion exchange always

resulted in some protein loss. Additionally, the ability of the chimera to bind nucleic acid complicates the issue of specific TAR binding.

Transactivation studies and band shift assays by the Peterlin group indicate that constructs containing larger portions of hCycT1 were more functional than their more minimal counterparts (*205*). The transactivation assays imply that the relative success could be due to superior expression or folding, while band shift data show that larger constructs are actually more successful in binding TAR in EMSA assays. In fact, unpublished data from EMSA assays not available to us at the beginning of the project indicate that the constructs with shorter Tat regions may not form distinct or homogenous species when binding to TAR, if they bound at all. Whereas larger constructs containing the cyclin box and TRM domains of cyclin T1 as well as full length Tat (hCycT1 1-280/Tat 1-101) formed distinct bands in TAR binding assays, constructs containing only the TRM of cyclin T1 (hCycT1 249-280/Tat 1-101) formed smeared bands with indistinct borders; truncating Tat (hCycT1 249-280/Tat 1-60) resulted in very smeared bands of questionable density. Without the formation of distinct, homogeneous complexes, structural studies would be virtually impossible. Published studies that indicate that the CycT1 cyclin boxes as well as the TRM bind Tat and that both are important in positioning Tat for optimal TAR binding support this hypothesis (*205*).

Insights from the EIAV system

The recently reported structures of the hCycT1 cyclin boxes and the equine cyclinT1 (eCycT1) cyclin boxes/EIAV Tat/EIAV TAR complex provide insight into the hCycT1/Tat/TAR system, and may help in the design of future constructs for structural study. EIAV is a lentivirus that infects horses and causes Equine Infectious Anemia.

The EIAV retrovirus is similar to HIV in terms of structure and replication pathway, and like HIV, genome transcription requires the recruitment of the host cell pTEFb by an interaction of the viral Tat protein with the TAR element within the viral LTR (*208*). In the reported structures, hCycT1was solved as a fusion protein of hCycT1 (1–281) and EIAV Tat (22–78) with a 20 residue linker (GGT$(GGGS)_3$GGGTS) (*200*). The crystal structure showed density for the region hCycT1(8-263), but no density for the linker or EIAV Tat. The eCycT1 was also solved as an EIAV fusion protein, encompassing the residues eCycT1 (1-281)-GGGTS-EIAV Tat (13-69), for which density was observed for residues 5-267/4-263 (differences for individual members of the crystallographic dimer) for eCycT1 and residues 41-69 for EIAV Tat (*207*). Both the human and equine constructs contained the mutation R256W to increase protein stability.

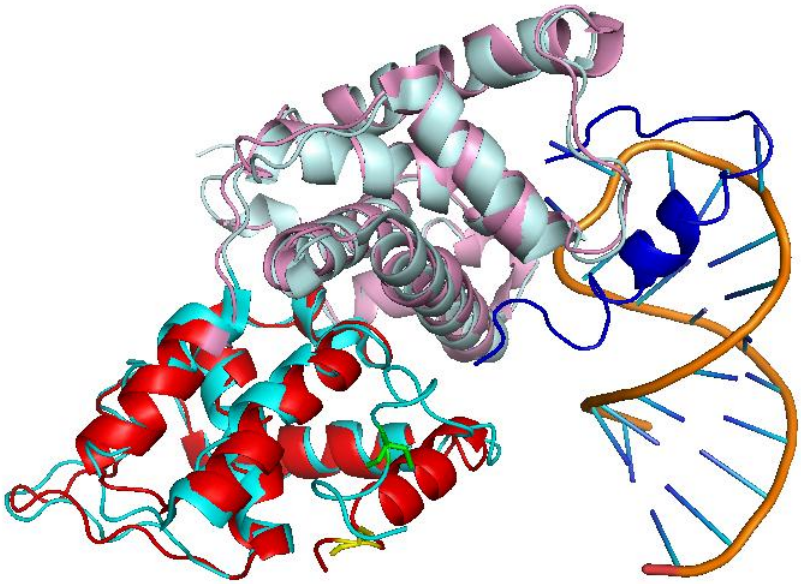
Based on the structural and sequential conservation between the two systems, it is extremely likely that the hCycT1/HIV Tat/TAR complex functions similarly to the eCycT1/EIAV Tat/TAR complex in the structural foundation that leads to transactivation. The cyclin box domains of equine and human cyclin T1 are very similar (see Figure C.3), with a sequence identity of 98% and a structural RMSD of 1.6 Å as determined by DaliLite calculations (*65*). EIAV and HIV Tat and TAR are more divergent, but contain sequence identity in key regions, indicating that the transactivation complex of HIV functions similarly. In the solved structure of the EIAV complex, the observable portion of the chimeric EIAV Tat encompasses the region that corresponds to the HIV Tat ARM as well as the sequences immediately preceeding and following it. The residues preceeding the ARM correspond to the C-terminal half of the core domain. The ARM and core domain regions contain the highest level of conservation between the HIV and

Figure C.3 The EIAV transactivating complex

A. Structural overlay of the solved cyclin boxes of equine and human cyclin T1. The hCycT1 cyclin box (pdb code 2PK2) is shown overlayed with the complex of EIAV TAR and a chimera of the eCycT1 cyclin boxes with a fragment of EIAV Tat (pdb code 2W2H). The eCycT1 cyclin box domains are colored in pale cyan (box1) and cyan (box 2) with C261 shown in sticks and colored green. The hCycT1 cyclin box domains are colored pink (box1) and red (box 2) with C261 in yellow. A portion of EIAV Tat, solved as a chimera with the eCycT1 box domain, is shown in blue, and is shown interacting with a short portion of EIAV TAR. Overlayed structures are nearly identical with a sequence identity of 98% and an RMSD of 1.6 as determined by DaliLite calculations (*65*).

B. Alignment of equine, human, and murine cyclin T1 proteins (1-300). The sequence alignment was created using ClustalW. The identity of the proteins are indicated (E, equine; H, human; M, murine.) The degree of conservation among the three proteins is indicated below the aligned sequences as follows: (*), total conservation; (:), high conservation; (.), moderate conservation. The five helices of the first cyclin box and the C-terminal helix (HC, Tat/TAR recognition motif) as determined by the solved structures of human and equine proteins are indicated and labeled below the sequences (H1-H5, HC, respectively) (*200, 207*). Above the sequence are indicated residues that in the equine protein interact with Tat/TAR in a distance shell smaller than 4 Å ((*207*), supplementary information, indicated with a T) or that in the human protein interact with Cdk9 within a 3.5 Å distance ((*201*), indicated with a C.) The proteins are very well conserved throughout, especially in the region of the first cyclin box. The residues that in the equine and human protein structures were shown to be within a 4 Å distance of Tat/TAR or CDK9 are universally conserved except for a three residue stretch located just N-terminal to helix H3, and except for position 261, which in the murine cyclin T1 is a tyrosine. The EIAV Tat recognition loop (TRL) of eCycT1 (as described in (*207*) is also indicated. The 79,80 positions are highlighted, which interact with bases in the loop region of EIAV TAR. The authors surmise that the smaller residues (PG) in HIV TAR may create a pocket for the larger G34 residue in HIV TAR which has been shown to be necessary for hCycT1 interaction (*209*).

A.



B.

```
                C     C                                    T   TT   T
E.      MEGERKNNNKRWYFTREQLENSPSRRFGLDPDKELSYRQQAANLLQDMGQRLNVSQLTIN 60
H.      MEGERKNNNKRWYFTREQLENSPSRRFGVDPDKELSYRQQAANLLQDMGQRLNVSQLTIN 60
M.      MEGERKNNNKRWYFTREQLENSPSRRFGVDSDKELSYRQQAANLLQDMGQRLNVSQLTIN 60
        ******************************:*.****************************
                                        H1

                        C    CCTTT        C    C     C CC    T   TT TTTTTTT
E.      TAIVYMHRFYMIQSFTQFHRNSVAPAALFLAAKVEGQPKKLEHVIKVAHACLHPQESLPD 120
H.      TAIVYMHRFYMIQSFTQFPGNSVAPAALFLAAKVEEQPKKLEHVIKVAHTCLHPQESLPD 120
M.      TAIVYMHRFYMIQSFTQFHRYSMAPAALFLAAKVEEQPKKLEHVIKVAHTCLHPQESLPD 120
        ******************   *:************ *************:**********
             H2                    H3                 H4       TRL

                 T       C     C    C C
E.      TRSEAYLQQVQDLVILESIILQTLGFELTIDHPHTHVVKCTQLVRASKDLAQTSYFMATN 180
H.      TRSEAYLQQVQDLVILESIILQTLGFELTIDHPHTHVVKCTQLVRASKDLAQTSYFMATN 180
M.      TRSEAYLQQVQDLVILESIILQTLGFELTIDHPHTHVVKCTQLVRASKDLAQTSYFMATN 180
        ************************************************************
                    H5

E.      SLHLTTFSLQYTPPVVACVCIHLACKWSNWEIPVSTDGKHWWEYVDATVTLELLDELTHE 240
H.      SLHLTTFSLQYTPPVVACVCIHLACKWSNWEIPVSTDGKHWWEYVDATVTLELLDELTHE 240
M.      SLHLTTFSLQYTPPVVACVCIHLACKWSNWEIPVSTDGKHWWEYVDATVTLELLDELTHE 240
        ************************************************************

                   T T       X
E.      FLQILEKTPNRLKRIWNWRACEAAKKTKADDRGTDEKTSEQTILNMISQSSSDTTIAGLM 300
H.      FLQILEKTPNRLKRIWNWRACEAAKKTKADDRGTDEKTSEQTILNMISQSSSDTTIAGLM 300
M.      FLQILEKTPSRLKRIRNWRAYQAAMKTKPDDRGADENTSEQTILNMISQTSSDTTIAGLM 300
        *********.***** **** :** ***.****:**:************:**********
                  HC
```

EIAV proteins (see Figure C.1). In the EIAV structure, the ARM-corresponding region forms a helix that inserts into the uppermost major groove of the TAR element in the stem/loop region, interacting with both bases and phosphates. Although the sequence of the TARs are otherwise divergent, EIAV and HIV TARs contain a three base stretch of sequence similarity in the stem below the loop, encompassing A6, G7, and A8 on the 5' side of the EIAV stem loop and A22, G26, and A27 of the HIV stem (spanning the 5' bulge) (see Figure C.4). The ARM-corresponding helix of EIAV Tat has a series of basic residues (Lys50, Gln54, and Lys57) on one side of the helix poised to interact with the A6-G7-A8 phosphate backbone of EIAV TAR. Two residues on the opposite side of the helix (K51 and R55) interact with bases in the loop region of TAR (G16 and U13, respectively (Figures C.4 and C.5). These residues and bases are all completely conserved in the HIV system, unlike in the comparative regions of the Simian Immunodeficiency Virus (SIV) and Bovine Immunodeficientcy Virus (BIV) Tat proteins. The eCycT1 Tyr37 residue is poised to interact with the backbone of G16 of TAR. When the solved hCycT1 structure is overlaid with the eCycT1 structure, the corresponding residue is in a nearly identical position. In the EIAV system, these interactions position residues in the loop to enable a tripartate interaction between eCycT1, EIAV TAR, and EIAV Tat. The interaction involves residues His79 and Arg80, two of the only three residues not conserved between the first 280 residues of hCycT1 and eCycT1. The differences between the EIAV and HIV TAR loops (EIAV has a tetraloop and HIV a hexaloop) and stems (HIV but not EIAV has a bulge on the 5'-side of the TAR stem) are likely accounted for by the differences in the Tat and cyclin T1 proteins. As described by the authors of the EIAV complex structure, EIAV Tat contains a four residue insert

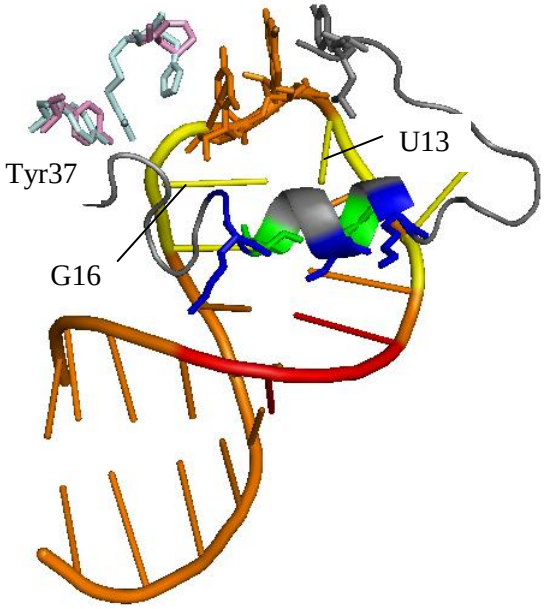
Figure C.4 Comparison of the TAR RNA stem-loop regions from EIAV and HIV-1.

A) and B) views of the interaction between the visible part of the equine Tat that was solved as a chimera with eCycT1, and equine TAR. The view in B) is rotated approximately 90˚ about a vertical axis with respect to the view in A). Both views are shown as cartoons with the TAR element colored orange. The A7-G8-A9 region is colored red. The backbone of these bases is in proximity to basic residues on one side of the Tat ARM helix, which are colored blue. Extending from the opposite side of the helix are basic residues that are colored green (Lys51 and Arg55) and that interact with bases in the loop region (G16 and U13, respectively, colored yellow.)

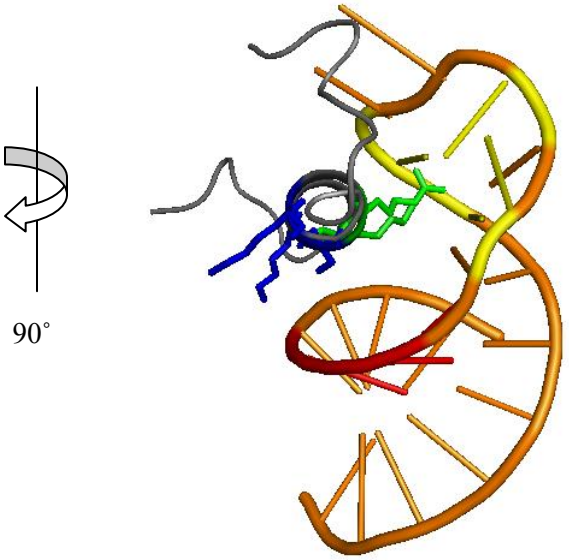
C) Alignment of the core and ARM domains of Tat from HIV, EIAV, SIV, and BIV. Residues that in the EIAV complex structure were poised to interact with the backbone of bases in the A7-A9 TAR region (and corresponding HIV Tat residues) are shown in blue text. Residues that were shown to interact with EIAV loop residues (and corresponding HIV Tat residues) are shown in green text. The four residue insert of EIAV Tat between the core and ARM domains that is not present in HIV Tat is highlighted in yellow.

D) and E) A comparison of EIAV and HIV TAR elements, respectively. The secondary structure of the TAR elements are shown. D) EIAV TAR encompasses a 22 nucleotide stem-loop structure with two U-G wobble base pairings preceding the loop section. Red circles indicate bases with backbone proximity to residues from the ARM region of EIAV Tat in the solved structure. Yellow circles indicate bases that interact with EIAV Tat in the structure. Open triangles indicate bases listed by the author to have phosphate interaction with EIAV Tat. E) HIV-1 TAR has a hexameric loop and a 3-mer bulge in the 5’ strand side of the stem. The boxed region of the TAR stemloops shows similar sequences. The arrow at the base of the HIV loop indicates the residue known to be critical to hCyclinT1 binding. D) and E) are adapted by permission from Macmillan Publishers Ltd: [NATURE STRUCTURAL & MOLECULAR BIOLOGY] (Anand, K., Schulte, A., Vogel-Bachmayr, K., Scheffzek, K., and Geyer, M., (2008) Structural insights into the Cyclin T1–Tat–TAR RNA transcription activation complex from EIAV, (supplementary information), *Nature structural & molecular biology* 15, 1287-1292.), copyright (2008)

A.



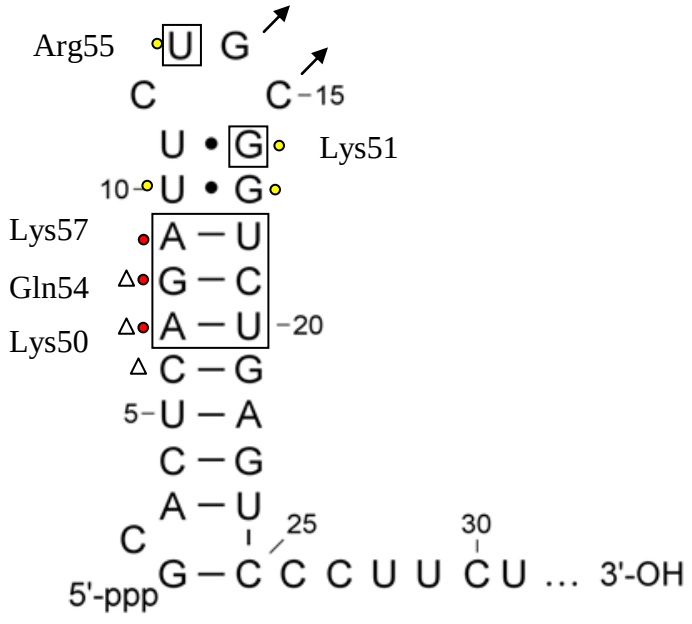
B.



C.

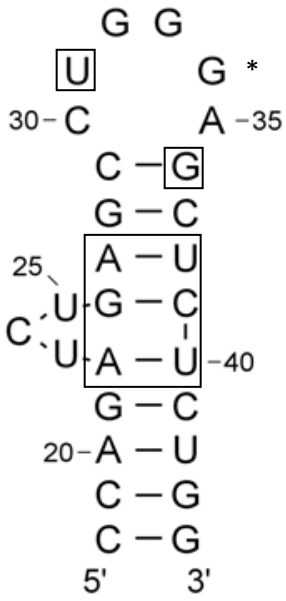
```
<--Cys-rich--      Core domain    ARM domain
HIV   KCCFHCQVCFITKGLGISYG----RKKRRQRRRAPQDS-QTHQVSLS-KQPTSQPRGDPTG 83
EIAV  RPNYHCQLCFL-RSLGIDYLDASLRKKNKQRLKAIQ---QGRQPQYLL---          78
                              RKK51 QR55+
BIV   ICSWHCQPCFLQKNLGINYGSG-PRPRGTRGKGRRIRRTASGEDQ                88
SIV   RCCYHCQHCFLKKGLGICY-EQH-RRRTPKKTK-TNPLPASNNRSLSTRTRNRQ      111
```

D.



EIAV TAR

E.



HIV TAR

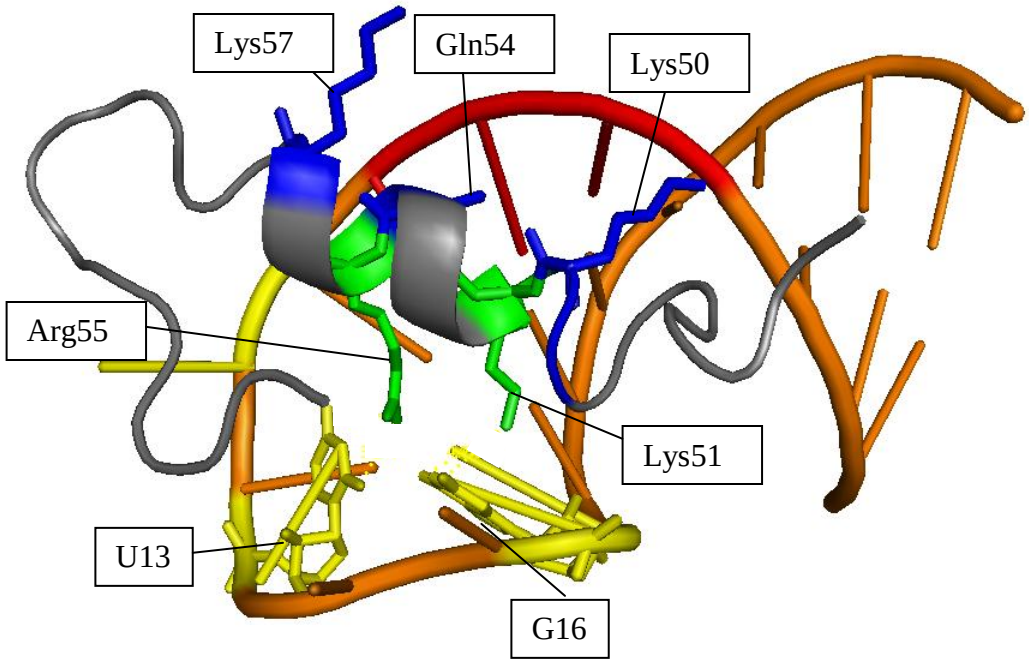


Figure C.5 The TAR interacting residues of the EIAV Tat ARM

EIAV Tat and TAR are displayed in cartoon form and colored as described in Figure C.5. All residues and bases (displayed as sticks) indicated are conserved between EIAV and HIV Tat/TAR.

(Asp45 Ala46 Ser47 Leu48) between the core domain and ARM domain that is not present in HIV Tat. The HIV TAR bulge is likely positioned to correspond to the four residue Tat deletion, maintaining the positional conformation of the conserved flanking core and ARM domains, and further contributing to specific protein/RNA interactions.

Despite the differences between the systems, the similarities indicate that the structural foundation for transactivation for HIV will mirror the EIAV system. If so, our minimal construct, which contained only residues 249-280 of hCycT1, was missing the majority of the regions that are involved in HIV Tat/TAR interaction, and was probably not adequate to form distinct, homogeneous, tightly binding complexes necessary for structural studies of the transactivating trio.

The 249-263 residue region of hCycT1 forms a helix at the C-terminal end of the second cyclin box in the solved structure (*200*). The helix, called helix Hc by the authors, corresponds to the Tat/TAR recognition motif. In the EIAV transactivating system, in addition to the critical Cys261, this helix contains two residues that were shown to interact with Tat/TAR, whereas 18 residues in cyclin box 1 are within 4Å distance shell of its Tat/TAR partners (*207*). Based on the insights from the reported structures of hCycT1 and the EIAV transactivating complex, the 249-280 residue region of hCycT1, corresponding to the single helix Hc, would not likely be sufficient to describe the transactivating complex or to enable the formation of homogenous representative units sufficient for structural studies.

Future directions

While the reported structure of the EIAV transactivating complex provides insights into the corresponding HIV system, the differences between the two (including

the bulge and larger loop of HIV TAR and the additional cysteines of the cysteine-rich domain of HIV Tat) are sufficient to indicate that a structure of the HIV system would be beneficial for the rational design of small molecules targeted to the tripartate transactivating complex. The EIAV complex structure indicates that a redesign of hCycT1/HIV Tat chimera is probably required for successful studies of the HIV transactivating complex.

EIAV Tat has far fewer cysteines in the region corresponding to the HIV Tat cysteine-rich domain, and the projected EIAV complex is expected to form a single zinc finger-like interface between EIAV Tat and eCycT1, as opposed to the transactivating complexes of other immunodeficiency viruses like HIV, which are believed to require two zinc atoms to form the Tat/CyclinT1 interface. Despite the simplification of the EIAV Tat/CyclinT1 interface, however, the authors of the paper of the reported transactivating complex structure were unable to detect density for the cysteine-rich region of Tat or the interaction between the TRM of eCycT1 and EIAV Tat. The system was constrained by the requirements of the formation of the crystallographic dimer, which required a relatively short RNA construct, and may have been the cause of the lack of density for the TRM/Tat interaction.

Perhaps a longer RNA construct could be used in the elucidation of a complex structure if NMR were used to study the complex. To reduce the signal overload inherent to NMR studies of large constructs, it may be possible to create a construct containing only the first (and not second) cyclin box of hCycT1. The hydrophobic face of cyclin box 1, normally interfaced with cyclin box 2 could possibly be protected by a well-designed linker that included important residues from cyclin box 2 that are normally

poised to interact with cyclin box 1. This linker could connect to the TRM (residues 249-280) known to be required for Tat interaction. Based on the successes and limitations of the reported EIAV chimera, the HIV transactivating chimera could include a linker and portion of HIV Tat similar in length to that described by the EIAV structure.

The success of the minimal chimera from the Peterlin studies in transactivating the transcription of the LTR in murine cells indicates that such a chimera could be successful. The smallest constructs described in the published studies contained only the 249-280 portion of hCycT1, which roughly corresponds to the TRM, and which forms a helix in the solved structures of hCycT1. EMSA studies by the Peterlin lab showed that the minimal construct bound to TAR, and was capable of binding to exogenous mCycT1 or hCycT1. A chimera containing an additional five residues of cycT1 (244-280) resulted in a drop in the level of transactivation from 150-fold over baseline levels to only 50-fold over baseline compared to a construct containing residues (249-280) of hCycT1. The hCycT1 structure indicates these additional five residues consisted of a turn and the C-terminal end of the preceding helix. Additionally, EMSA data showing that the hCycT1(249-280)-Tat construct bound both human and murine cyclinT1, indicating that chimera-directed transactivation in murine cells is accomplished through the recruiting of exogenous cyclinT1, in which the 249-280 helix is displaced. Together, these observations indicate a good potential for success in structural studies of a chimera. The purification scheme undertaken in the successful EIAV transactivating complex study will provide a good starting point for corresponding HIV chimeric study strategies.

REFERENCES

1. Olovnikov, A. M. (1973) A theory of marginotomy. The incomplete copying of template margin in enzymic synthesis of polynucleotides and biological significance of the phenomenon, *J Theor Biol 41*, 181-190.

2. Varga, T., and Aplan, P. D. (2005) Chromosomal aberrations induced by double strand DNA breaks, *DNA Repair (Amst) 4*, 1038-1046.

3. Latre, L., Tusell, L., Martin, M., Miro, R., Egozcue, J., Blasco, M. A., and Genesca, A. (2003) Shortened telomeres join to DNA breaks interfering with their correct repair, *Exp Cell Res 287*, 282-288.

4. Levis, R. W., Ganesan, R., Houtchens, K., Tolar, L. A., and Sheen, F. M. (1993) Transposons in place of telomeric repeats at a Drosophila telomere, *Cell 75*, 1083-1093.

5. Pardue, M. L., and DeBaryshe, P. G. (1999) Drosophila telomeres: two transposable elements with important roles in chromosomes, *Genetica 107*, 189-196.

6. Louis, E. J. (2002) Are Drosophila telomeres an exception or the rule?, *Genome Biol 3*, REVIEWS0007.

7. Collins, K. (1996) Structure and function of telomerase, *Curr Opin Cell Biol 8*, 374-380.

8. Preston, R. J. (1997) Telomeres, telomerase and chromosome stability, *Radiat Res 147*, 529-534.

9. McEachern, M. J., Krauskopf, A., and Blackburn, E. H. (2000) Telomeres and their control, *Annu Rev Genet 34*, 331-358.

10. Casjens, S. (1998) The diverse and dynamic structure of bacterial genomes, *Annu Rev Genet 32*, 339-377.

11. Lin, Y. S., Kieser, H. M., Hopwood, D. A., and Chen, C. W. (1993) The chromosomal DNA of Streptomyces lividans 66 is linear, *Mol Microbiol 10*, 923-933.

12. Reeves, A. R., Post, D. A., and Vanden Boom, T. J. (1998) Physical-genetic map of the erythromycin-producing organism Saccharopolyspora erythraea, *Microbiology 144 (Pt 8)*, 2151-2159.

13. Rekosh, D. M., Russell, W. C., Bellet, A. J., and Robinson, A. J. (1977) Identification of a protein linked to the ends of adenovirus DNA, *Cell 11*, 283-295.

14. Mellado, R. P., Penalva, M. A., Inciarte, M. R., and Salas, M. (1980) The protein covalently linked to the 5' termini of the DNA of Bacillus subtilis phage phi 29 is involved in the initiation of DNA replication, *Virology 104*, 84-96.

15. Lin, Y. R., Hahn, M. Y., Roe, J. H., Huang, T. W., Tsai, H. H., Lin, Y. F., Su, T. S., Chan, Y. J., and Chen, C. W. (2009) Streptomyces telomeres contain a promoter, *J Bacteriol 191*, 773-781.

16. Hopwood, D. A. (2006) Soil to genomics: the Streptomyces chromosome, *Annu Rev Genet 40*, 1-23.

17. Casjens, S. (1999) Evolution of the linear DNA replicons of the Borrelia spirochetes, *Curr Opin Microbiol 2*, 529-534.

18. Ooi, Y. S., Warburton, P. E., Ravin, N. V., and Narayanan, K. (2007) Recombineering linear DNA that replicate stably in E. coli, *Plasmid*.

19. Barbour, A. G., and Garon, C. F. (1987) Linear plasmids of the bacterium Borrelia burgdorferi have covalently closed ends, *Science 237*, 409-411.

20. Hinnebusch, J., and Barbour, A. G. (1991) Linear plasmids of Borrelia burgdorferi have a telomeric structure and sequence similar to those of a eukaryotic virus, *J Bacteriol 173*, 7233-7239.

21. Casjens, S., Murphy, M., DeLange, M., Sampson, L., van Vugt, R., and Huang, W. M. (1997) Telomeres of the linear chromosomes of Lyme disease spirochaetes: nucleotide sequence and possible exchange with linear plasmid telomeres, *Mol Microbiol 26*, 581-596.

22. Rybchin, V. N., and Svarchevsky, A. N. (1999) The plasmid prophage N15: a linear DNA with covalently closed ends, *Mol Microbiol 33*, 895-903.

23. Picardeau, M., Lobry, J. R., and Hinnebusch, B. J. (1999) Physical mapping of an origin of bidirectional replication at the centre of the Borrelia burgdorferi linear chromosome, *Mol Microbiol 32*, 437-445.

24. Ravin, N. V., Kuprianov, V. V., Gilcrease, E. B., and Casjens, S. R. (2003) Bidirectional replication from an internal ori site of the linear N15 plasmid prophage, *Nucleic Acids Res 31*, 6552-6560.

25. Kobryn, K., and Chaconas, G. (2001) The circle is broken: telomere resolution in linear replicons, *Curr Opin Microbiol 4*, 558-564.

26. Ravin, N. V., Strakhova, T. S., and Kuprianov, V. V. (2001) The protelomerase of the phage-plasmid N15 is responsible for its maintenance in linear form, *J Mol Biol 312*, 899-906.

27. Kobryn, K., and Chaconas, G. (2002) ResT, a telomere resolvase encoded by the Lyme disease spirochete, *Mol Cell 9*, 195-201.

28. Deneke, J., Ziegelin, G., Lurz, R., and Lanka, E. (2000) The protelomerase of temperate Escherichia coli phage N15 has cleaving-joining activity, *Proc Natl Acad Sci U S A 97*, 7721-7726.

29. Huang, W. M., Joss, L., Hsieh, T., and Casjens, S. (2004) Protelomerase uses a topoisomerase IB/Y-recombinase type mechanism to generate DNA hairpin ends, *J Mol Biol 337*, 77-92.

30. Steere, A. C., Coburn, J., and Glickstein, L. (2004) The emergence of Lyme disease, *J Clin Invest 113*, 1093-1101.

31. Guidoboni, M., Ferreri, A. J., Ponzoni, M., Doglioni, C., and Dolcetti, R. (2006) Infectious agents in mucosa-associated lymphoid tissue-type lymphomas: pathogenic role and therapeutic perspectives, *Clin Lymphoma Myeloma 6*, 289-300.

32. Fraser, C. M., Casjens, S., Huang, W. M., Sutton, G. G., Clayton, R., Lathigra, R., White, O., Ketchum, K. A., Dodson, R., Hickey, E. K., Gwinn, M., Dougherty, B., Tomb, J. F., Fleischmann, R. D., Richardson, D., Peterson, J., Kerlavage, A. R., Quackenbush, J., Salzberg, S., Hanson, M., van Vugt, R., Palmer, N., Adams, M. D., Gocayne, J., Weidman, J., Utterback, T., Watthey, L., McDonald, L., Artiach, P., Bowman, C., Garland, S., Fuji, C., Cotton, M. D., Horst, K., Roberts, K., Hatch, B., Smith, H. O., and Venter, J. C. (1997) Genomic sequence of a Lyme disease spirochaete, Borrelia burgdorferi, *Nature 390*, 580-586.

33. Casjens, S., Palmer, N., van Vugt, R., Huang, W. M., Stevenson, B., Rosa, P., Lathigra, R., Sutton, G., Peterson, J., Dodson, R. J., Haft, D., Hickey, E., Gwinn, M., White, O., and Fraser, C. M. (2000) A bacterial genome in flux: the twelve linear and nine circular extrachromosomal DNAs in an infectious isolate of the Lyme disease spirochete Borrelia burgdorferi, *Mol Microbiol 35*, 490-516.

34. Baril, C., Richaud, C., Baranton, G., and Saint Girons, I. S. (1989) Linear chromosome of Borrelia burgdorferi, *Res Microbiol 140*, 507-516.

35. Byram, R., Stewart, P. E., and Rosa, P. (2004) The essential nature of the ubiquitous 26-kilobase circular replicon of Borrelia burgdorferi, *J Bacteriol 186*, 3561-3569.

36. Tourand, Y., Bankhead, T., Wilson, S. L., Putteet-Driver, A. D., Barbour, A. G., Byram, R., Rosa, P. A., and Chaconas, G. (2006) Differential telomere processing by Borrelia telomere resolvases in vitro but not in vivo, *J Bacteriol 188*, 7378-7386.

37. Dworkin, M. S., Schwan, T. G., and Anderson, D. E., Jr. (2002) Tick-borne relapsing fever in North America, *Med Clin North Am 86*, 417-433, viii-ix.

38. Schwan, T. G., Policastro, P. F., Miller, Z., Thompson, R. L., Damrow, T., and Keirans, J. E. (2003) Tick-borne relapsing fever caused by Borrelia hermsii, Montana, *Emerg Infect Dis 9*, 1151-1154.

39. Smith, E. F., and Townsend, C. O. (1907) A Plant-Tumor of Bacterial Origin, *Science 25*, 671-673.

40. Goodner, B., Hinkle, G., Gattung, S., Miller, N., Blanchard, M., Qurollo, B., Goldman, B. S., Cao, Y., Askenazi, M., Halling, C., Mullin, L., Houmiel, K., Gordon, J., Vaudin, M., Iartchouk, O., Epp, A., Liu, F., Wollam, C., Allinger, M., Doughty, D., Scott, C., Lappas, C., Markelz, B., Flanagan, C., Crowell, C., Gurson, J., Lomo, C., Sear, C., Strub, G., Cielo, C., and Slater, S. (2001) Genome sequence of the plant pathogen and biotechnology agent Agrobacterium tumefaciens C58, *Science 294*, 2323-2328.

41. Moore, L. W., Chilton, W. S., and Canfield, M. L. (1997) Diversity of Opines and Opine-Catabolizing Bacteria Isolated from Naturally Occurring Crown Gall Tumors, *Appl Environ Microbiol 63*, 201-207.

42. Ravin, N. V. (2003) Mechanisms of replication and telomere resolution of the linear plasmid prophage N15, *FEMS Microbiol Lett 221*, 1-6.

43. Mobberley, J. M., Authement, R. N., Segall, A. M., and Paul, J. H. (2008) The temperate marine phage PhiHAP-1 of Halomonas aquamarina possesses a linear plasmid-like prophage genome, *J Virol 82*, 6618-6630.

44. Casjens, S. R., Gilcrease, E. B., Huang, W. M., Bunny, K. L., Pedulla, M. L., Ford, M. E., Houtz, J. M., Hatfull, G. F., and Hendrix, R. W. (2004) The pKO2 linear plasmid prophage of Klebsiella oxytoca, *J Bacteriol 186*, 1818-1832.

45. Munro, J., Oakey, J., Bromage, E., and Owens, L. (2003) Experimental bacteriophage-mediated virulence in strains of Vibrio harveyi, *Dis Aquat Organ 54*, 187-194.

46. Karunasagar, I., Pai, R., Malathi, G. R., and Karunasagar, I. (1994) Mass mortality of Penaeus monodon larvae due to antibiotic-resistant Vibrio harveyi infection, *Aquaculture 128*, 203-209.

47. Austin, B., and Zhang, X. H. (2006) Vibrio harveyi: a significant pathogen of marine vertebrates and invertebrates, *Lett Appl Microbiol 43*, 119-124.

48. Zabala, B., Hammerl, J. A., Espejo, R. T., and Hertwig, S. (2009) The linear plasmid prophage Vp58.5 of Vibrio parahaemolyticus is closely related to the integrating phage VHML and constitutes a new incompatibility group of telomere phages, *J Virol*.

49. Lan, S. F., Huang, C. H., Chang, C. H., Liao, W. C., Lin, I. H., Jian, W. N., Wu, Y. G., Chen, S. Y., and Wong, H. C. (2009) Characterization of a new plasmid-like prophage in a pandemic Vibrio parahaemolyticus O3:K6 strain, *Appl Environ Microbiol 75*, 2659-2667.

50. Penland, R. L., Boniuk, M., and Wilhelmus, K. R. (2000) Vibrio ocular infections on the U.S. Gulf Coast, *Cornea 19*, 26-29.

51. Hertwig, S., Klein, I., Lurz, R., Lanka, E., and Appel, B. (2003) PY54, a linear plasmid prophage of Yersinia enterocolitica with covalently closed ends, *Mol Microbiol 48*, 989-1003.

52. Hertwig, S., Klein, I., Schmidt, V., Beck, S., Hammerl, J. A., and Appel, B. (2003) Sequence analysis of the genome of the temperate Yersinia enterocolitica phage PY54, *J Mol Biol 331*, 605-622.

53. Hammerl, J. A., Klein, I., Appel, B., and Hertwig, S. (2007) Interplay between the temperate phages PY54 and N15, linear plasmid prophages with covalently closed ends, *J Bacteriol*.

54. Deneke, J., Ziegelin, G., Lurz, R., and Lanka, E. (2002) Phage N15 telomere resolution. Target requirements for recognition and processing by the protelomerase, *J Biol Chem 277*, 10410-10419.

55. Kobryn, K., Burgin, A. B., and Chaconas, G. (2005) Uncoupling the chemical steps of telomere resolution by ResT, *J Biol Chem 280*, 26788-26795.

56. Craig, N. L., and Nash, H. A. (1983) The mechanism of phage lambda site-specific recombination: site-specific breakage of DNA by Int topoisomerase, *Cell 35*, 795-803.

57. Pargellis, C. A., Nunes-Duby, S. E., de Vargas, L. M., and Landy, A. (1988) Suicide recombination substrates yield covalent lambda integrase-DNA complexes and lead to identification of the active site tyrosine, *J Biol Chem 263*, 7678-7685.

58. Champoux, J. J. (1981) DNA is linked to the rat liver DNA nicking-closing enzyme by a phosphodiester bond to tyrosine, *J Biol Chem 256*, 4805-4809.

59. Sherratt, D. J., and Wigley, D. B. (1998) Conserved themes but novel activities in recombinases and topoisomerases, *Cell 93*, 149-152.

60. Biswas, T., Aihara, H., Radman-Livaja, M., Filman, D., Landy, A., and Ellenberger, T. (2005) A structural basis for allosteric control of DNA recombination by lambda integrase, *Nature 435*, 1059-1066.

61. Radman-Livaja, M., Biswas, T., Ellenberger, T., Landy, A., and Aihara, H. (2006) DNA arms do the legwork to ensure the directionality of lambda site-specific recombination, *Curr Opin Struct Biol 16*, 42-50.

62. Patel, A., Shuman, S., and Mondragon, A. (2006) Crystal structure of a bacterial type IB DNA topoisomerase reveals a preassembled active site in the absence of DNA, *J Biol Chem 281*, 6030-6037.

63. Champoux, J. J. (2001) DNA topoisomerases: structure, function, and mechanism, *Annu Rev Biochem 70*, 369-413.

64. Aihara, H., Huang, W. M., and Ellenberger, T. (2007) An interlocked dimer of the protelomerase TelK distorts DNA structure for the formation of hairpin telomeres, *Mol Cell 27*, 901-913.

65. Holm, L., and Park, J. (2000) DaliLite workbench for protein structure comparison, *Bioinformatics 16*, 566-567.

66. Sattler, M., Schleucher, J., and Griesinger, C. (1999) Heteronuclear multidimensional NMR experiments for the structure determination of proteins in solution employing pulsed field gradients, *Progress in Nuclear Magnetic Resonance Spectroscopy 34*, 93-158.

67. Cornilescu, G., Delaglio, F., and Bax, A. (1999) Protein backbone angle restraints from searching a database for chemical shift and sequence homology, *J Biomol NMR 13*, 289-302.

68. Guntert, P. (2004) Automated NMR structure calculation with CYANA, *Methods Mol Biol 278*, 353-378.

69. Laskowski, R. A., Rullmannn, J. A., MacArthur, M. W., Kaptein, R., and Thornton, J. M. (1996) AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR, *J Biomol NMR 8*, 477-486.

70. DeLano, W. L. (2002) *The PyMOL User's Manual*, DeLano Scientific, Palo Alto, CA, USA.

71. Koradi, R., Billeter, M., and Wuthrich, K. (1996) MOLMOL: a program for display and analysis of macromolecular structures, *J Mol Graph 14*, 51-55, 29-32.

72. Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T. J., Higgins, D. G., and Thompson, J. D. (2003) Multiple sequence alignment with the Clustal series of programs, *Nucleic Acids Res 31*, 3497-3500.

73. Higgins, D. G., Thompson, J. D., and Gibson, T. J. (1996) Using CLUSTAL for multiple sequence alignments, *Methods Enzymol 266*, 383-402.

74. Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J., and Higgins, D. G. (2007) Clustal W and Clustal X version 2.0, *Bioinformatics 23*, 2947-2948.

75. Thompson, J. D., Higgins, D. G., and Gibson, T. J. (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*, 4673-4680.

76. Wojciak, J. M., Connolly, K. M., and Clubb, R. T. (1999) NMR structure of the Tn916 integrase-DNA complex, *Nat Struct Biol 6*, 366-373.

77. Flick, K. E., Jurica, M. S., Monnat, R. J., Jr., and Stoddard, B. L. (1998) DNA binding and cleavage by the nuclear intron-encoded homing endonuclease I-PpoI, *Nature 394*, 96-101.

78. Allen, M. D., Yamasaki, K., Ohme-Takagi, M., Tateno, M., and Suzuki, M. (1998) A novel mode of DNA recognition by a beta-sheet revealed by the solution structure of the GCC-box binding domain in complex with DNA, *Embo J 17*, 5484-5496.

79. Ryter, J. M., and Schultz, S. C. (1998) Molecular basis of double-stranded RNA-protein interactions: structure of a dsRNA-binding domain complexed with dsRNA, *Embo J 17*, 7505-7513.

80. Ramos, A., Grunert, S., Adams, J., Micklem, D. R., Proctor, M. R., Freund, S., Bycroft, M., St Johnston, D., and Varani, G. (2000) RNA recognition by a Staufen double-stranded RNA-binding domain, *Embo J 19*, 997-1009.

81. Blaszczyk, J., Gan, J., Tropea, J. E., Court, D. L., Waugh, D. S., and Ji, X. (2004) Noncatalytic assembly of ribonuclease III with double-stranded RNA, *Structure 12*, 457-466.

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

83. McGuffin, L. J., Bryson, K., and Jones, D. T. (2000) The PSIPRED protein structure prediction server, *Bioinformatics 16*, 404-405.

84. Farrow, N. A., Zhang, O., Forman-Kay, J. D., and Kay, L. E. (1995) Comparison of the backbone dynamics of a folded and an unfolded SH3 domain existing in equilibrium in aqueous buffer, *Biochemistry 34*, 868-878.

85. Cho, H. S., Liu, C. W., Damberger, F. F., Pelton, J. G., Nelson, H. C., and Wemmer, D. E. (1996) Yeast heat shock transcription factor N-terminal activation domains are unstructured as probed by heteronuclear NMR spectroscopy, *Protein Sci 5*, 262-269.

86. Kabsch, W., and Sander, C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, *Biopolymers 22*, 2577-2637.

87. Wishart, D. S., and Sykes, B. D. (1994) The 13C chemical-shift index: a simple method for the identification of protein secondary structure using 13C chemical-shift data, *J Biomol NMR 4*, 171-180.

88. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) Basic local alignment search tool, *J Mol Biol 215*, 403-410.

89. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res 25*, 3389-3402.

90. Baker, N. A., Sept, D., Joseph, S., Holst, M. J., and McCammon, J. A. (2001) Electrostatics of nanosystems: application to microtubules and the ribosome, *Proc Natl Acad Sci U S A 98*, 10037-10041.

91. Holm, L., and Sander, C. (1993) Protein structure comparison by alignment of distance matrices, *J Mol Biol 233*, 123-138.

92. Stefl, R., Xu, M., Skrisovska, L., Emeson, R. B., and Allain, F. H. (2006) Structure and specific RNA binding of ADAR2 double-stranded RNA binding motifs, *Structure 14*, 345-355.

93. St Johnston, D., Brown, N. H., Gall, J. G., and Jantsch, M. (1992) A conserved double-stranded RNA-binding domain, *Proc Natl Acad Sci U S A 89*, 10979-10983.

94. Tian, B., Bevilacqua, P. C., Diegelman-Parente, A., and Mathews, M. B. (2004) The double-stranded-RNA-binding motif: interference and much more, *Nat Rev Mol Cell Biol 5*, 1013-1023.

95. Fierro-Monti, I., and Mathews, M. B. (2000) Proteins binding to duplexed RNA: one motif, multiple functions, *Trends Biochem Sci 25*, 241-246.

96. Tian, B., and Mathews, M. B. (2001) Functional characterization of and cooperation between the double-stranded RNA-binding motifs of the protein kinase PKR, *J Biol Chem 276*, 9936-9944.

97. Chang, K. Y., and Ramos, A. (2005) The double-stranded RNA-binding motif, a versatile macromolecular docking platform, *Febs J 272*, 2109-2117.

98. Poulsen, H., Jorgensen, R., Heding, A., Nielsen, F. C., Bonven, B., and Egebjerg, J. (2006) Dimerization of ADAR2 is mediated by the double-stranded RNA binding domain, *Rna 12*, 1350-1360.

99. Patel, R. C., Stanton, P., McMillan, N. M., Williams, B. R., and Sen, G. C. (1995) The interferon-inducible double-stranded RNA-activated protein kinase self-associates in vitro and in vivo, *Proc Natl Acad Sci U S A 92*, 8283-8287.

100. Doyle, M., and Jantsch, M. F. (2002) New and old roles of the double-stranded RNA-binding domain, *J Struct Biol 140*, 147-153.

101. Daher, A., Longuet, M., Dorin, D., Bois, F., Segeral, E., Bannwarth, S., Battisti, P. L., Purcell, D. F., Benarous, R., Vaquero, C., Meurs, E. F., and Gatignol, A. (2001) Two dimerization domains in the trans-activation response RNA-binding protein (TRBP) individually reverse the protein kinase R inhibition of HIV-1 long terminal repeat expression, *J Biol Chem 276*, 33899-33905.

102. Cho, D. S., Yang, W., Lee, J. T., Shiekhattar, R., Murray, J. M., and Nishikura, K. (2003) Requirement of dimerization for RNA editing activity of adenosine deaminases acting on RNA, *J Biol Chem 278*, 17093-17102.

103. Valente, L., and Nishikura, K. (2007) RNA binding-independent dimerization of adenosine deaminases acting on RNA and dominant negative effects of nonfunctional subunits on dimer functions, *J Biol Chem 282*, 16054-16061.

104. Chatel-Chaix, L., Abrahamyan, L., Frechina, C., Mouland, A. J., and DesGroseillers, L. (2007) The host protein Staufen1 participates in human immunodeficiency virus type 1 assembly in live cells by influencing pr55Gag multimerization, *J Virol 81*, 6216-6230.

105. Chatel-Chaix, L., Clement, J. F., Martel, C., Beriault, V., Gatignol, A., DesGroseillers, L., and Mouland, A. J. (2004) Identification of Staufen in the human immunodeficiency virus type 1 Gag ribonucleoprotein complex and a role in generating infectious viral particles, *Mol Cell Biol 24*, 2637-2648.

106. Krovat, B. C., and Jantsch, M. F. (1996) Comparative mutational analysis of the double-stranded RNA binding domains of Xenopus laevis RNA-binding protein A, *J Biol Chem 271*, 28112-28119.

107. Patel, R. C., and Sen, G. C. (1998) Requirement of PKR dimerization mediated by specific hydrophobic residues for its activation by double-stranded RNA and its antigrowth effects in yeast, *Mol Cell Biol 18*, 7009-7019.

108. Nanduri, S., Carpick, B. W., Yang, Y., Williams, B. R., and Qin, J. (1998) Structure of the double-stranded RNA-binding domain of the protein kinase PKR reveals the molecular basis of its dsRNA-mediated activation, *Embo J 17*, 5458-5465.

109. Sohn, S. Y., Bae, W. J., Kim, J. J., Yeom, K. H., Kim, V. N., and Cho, Y. (2007) Crystal structure of human DGCR8 core, *Nat Struct Mol Biol 14*, 847-853.

110. Akey, D. L., and Berger, J. M. (2005) Structure of the nuclease domain of ribonuclease III from M. tuberculosis at 2.1 A, *Protein Sci 14*, 2744-2750.

111. Gan, J., Tropea, J. E., Austin, B. P., Court, D. L., Waugh, D. S., and Ji, X. (2006) Structural insight into the mechanism of double-stranded RNA processing by ribonuclease III, *Cell 124*, 355-366.

112. Connolly, K. M., Ilangovan, U., Wojciak, J. M., Iwahara, M., and Clubb, R. T. (2000) Major groove recognition by three-stranded beta-sheets: affinity determinants and conserved structural features, *J Mol Biol 300*, 841-856.

113. Fadeev, E. A., Sam, M. D., and Clubb, R. T. (2009) NMR structure of the amino-terminal domain of the lambda integrase protein in complex with DNA: immobilization of a flexible tail facilitates beta-sheet recognition of the major groove, *J Mol Biol 388*, 682-690.

114. Nash, H. A., and Robertson, C. A. (1989) Heteroduplex substrates for bacteriophage lambda site-specific recombination: cleavage and strand transfer products, *Embo J 8*, 3523-3533.

115. Rudy, C., Taylor, K. L., Hinerfeld, D., Scott, J. R., and Churchward, G. (1997) Excision of a conjugative transposon in vitro by the Int and Xis proteins of Tn916, *Nucleic Acids Res 25*, 4061-4066.

116. Lu, F., and Churchward, G. (1994) Conjugative transposition: Tn916 integrase contains two independent DNA binding domains that recognize different DNA sequences, *Embo J 13*, 1541-1548.

117. Poyart-Salmeron, C., Trieu-Cuot, P., Carlier, C., and Courvalin, P. (1989) Molecular characterization of two proteins involved in the excision of the conjugative transposon Tn1545: homologies with other site-specific recombinases, *Embo J 8*, 2425-2433.

118. Jaworski, D. D., Flannagan, S. E., and Clewell, D. B. (1996) Analyses of traA, int-Tn, and xis-Tn mutations in the conjugative transposon Tn916 in Enterococcus faecalis, *Plasmid 36*, 201-208.

119. Oakey, H. J., Cullen, B. R., and Owens, L. (2002) The complete nucleotide sequence of the Vibrio harveyi bacteriophage VHML, *J Appl Microbiol 93*, 1089-1098.

120. Wojciak, J. M., Sarkar, D., Landy, A., and Clubb, R. T. (2002) Arm-site binding by lambda -integrase: solution structure and functional characterization of its amino-terminal domain, *Proc Natl Acad Sci U S A 99*, 3434-3439.

121. Tirumalai, R. S., Kwon, H. J., Cardente, E. H., Ellenberger, T., and Landy, A. (1998) Recognition of core-type DNA sites by lambda integrase, *J Mol Biol 279*, 513-527.

122. Mardanov, A. V., and Ravin, N. V. (2009) Conversion of Linear DNA with Hairpin Telomeres into a Circular Molecule in the Course of Phage N15 Lytic Replication, *J Mol Biol*.

123. Ghosh, M., Kumar, N. V., Varshney, U., and Chary, K. V. (1999) Structural characterisation of a uracil containing hairpin DNA by NMR and molecular dynamics, *Nucleic Acids Res 27*, 3938-3944.

124. Mumm, J. P., Landy, A., and Gelles, J. (2006) Viewing single lambda site-specific recombination events from start to finish, *Embo J 25*, 4586-4595.

125. Kobryn, K., and Chaconas, G. (2005) Fusion of hairpin telomeres by the B. burgdorferi telomere resolvase ResT implications for shaping a genome in flux, *Mol Cell 17*, 783-791.

126. Chaconas, G. (2005) Hairpin telomeres and genome plasticity in Borrelia: all mixed up in the end, *Mol Microbiol 58*, 625-635.

127. Huang, W. M., Robertson, M., Aron, J., and Casjens, S. (2004) Telomere exchange between linear replicons of Borrelia burgdorferi, *J Bacteriol 186*, 4134-4141.

128. Sarkar, D., Radman-Livaja, M., and Landy, A. (2001) The small DNA binding domain of lambda integrase is a context-sensitive modulator of recombinase functions, *Embo J 20*, 1203-1212.

129. Ruby, E. G., and Nealson, K. H. (1976) Symbiotic association of Photobacterium fischeri with the marine luminous fish Monocentris japonica; a model of symbiosis based on bacterial studies, *Biol Bull 151*, 574-586.

130. Ruby, E. G., and McFall-Ngai, M. J. (1999) Oxygen-utilizing reactions and symbiotic colonization of the squid light organ by Vibrio fischeri, *Trends Microbiol 7*, 414-420.

131. Thompson, F. L., Iida, T., and Swings, J. (2004) Biodiversity of vibrios, *Microbiol Mol Biol Rev 68*, 403-431, table of contents.

132. Johnson, F. H., and Shunk, I. V. (1936) An Interesting New Species of Luminous Bacteria, *J Bacteriol 31*, 585-593.

133. Svitil, A. L., Chadhain, S., Moore, J. A., and Kirchman, D. L. (1997) Chitin Degradation Proteins Produced by the Marine Bacterium Vibrio harveyi Growing on Different Forms of Chitin, *Appl Environ Microbiol 63*, 408-413.

134. Verne, J. (1870) *Vingt mille lieues sous les mers*, Pierre-Jules Hetzel.

135. Miller, S. D., Haddock, S. H., Elvidge, C. D., and Lee, T. F. (2005) Detection of a bioluminescent milky sea from space, *Proc Natl Acad Sci U S A 102*, 14181-14184.

136. Mok, K. C., Wingreen, N. S., and Bassler, B. L. (2003) Vibrio harveyi quorum sensing: a coincidence detector for two autoinducers controls gene expression, *Embo J 22*, 870-881.

137. Nealson, K. H., and Hastings, J. W. (2006) Quorum sensing on a global scale: massive numbers of bioluminescent bacteria make milky seas, *Appl Environ Microbiol 72*, 2295-2297.

138. Oakey, H. J., and Owens, L. (2000) A new bacteriophage, VHML, isolated from a toxin-producing strain of Vibrio harveyi in tropical Australia, *J Appl Microbiol 89*, 702-709.

139. Defoirdt, T., Boon, N., Sorgeloos, P., Verstraete, W., and Bossier, P. (2007) Alternatives to antibiotics to control bacterial infections: luminescent vibriosis in aquaculture as an example, *Trends Biotechnol 25*, 472-479.

140. Luna, G. M., Biavasco, F., and Danovaro, R. (2007) Bacteria associated with the rapid tissue necrosis of stony corals, *Environ Microbiol 9*, 1851-1857.

141. Sithigorngul, P., Rukpratanporn, S., Pecharaburanin, N., Suksawat, P., Longyant, S., Chaivisuthangkura, P., and Sithigorngul, W. (2007) A simple and rapid immunochromatographic test strip for detection of pathogenic isolates of Vibrio harveyi, *J Microbiol Methods*.

142. Shivu, M. M., Rajeeva, B. C., Girisha, S. K., Karunasagar, I., Krohne, G., and Karunasagar, I. (2007) Molecular characterization of Vibrio harveyi bacteriophages isolated from aquaculture environments along the coast of India, *Environ Microbiol 9*, 322-331.

143. You, J., Xue, X., Cao, L., Lu, X., Wang, J., Zhang, L., and Zhou, S. (2007) Inhibition of Vibrio biofilm formation by a marine actinomycete strain A66, *Appl Microbiol Biotechnol 76*, 1137-1144.

144. Zhang, C., Yu, L., and Qian, R. (2007) Characterization of OmpK, GAPDH and their fusion OmpK-GAPDH derived from Vibrio harveyi outer membrane proteins: their immunoprotective ability against vibriosis in large yellow croaker (Pseudosciaena crocea), *J Appl Microbiol 103*, 1587-1599.

145. Hao, G. J., Shen, J. Y., Cao, Z., and Pan, X. Y. (2007) [Preparation and characterization of the mAb against Vibrio harveyi.], *Xi Bao Yu Fen Zi Mian Yi Xue Za Zhi 23*, 838-840.

146. Amparyup, P., Kondo, H., Hirono, I., Aoki, T., and Tassanakajon, A. (2007) Molecular cloning, genomic organization and recombinant expression of a crustin-

like antimicrobial peptide from black tiger shrimp Penaeus monodon, *Mol Immunol*.

147. Defoirdt, T., Miyamoto, C. M., Wood, T. K., Meighen, E. A., Sorgeloos, P., Verstraete, W., and Bossier, P. (2007) The natural furanone (5Z)-4-bromo-5-(bromomethylene)-3-butyl-2(5H)-furanone disrupts quorum sensing-regulated gene expression in Vibrio harveyi by decreasing the DNA-binding activity of the transcriptional regulator protein luxR, *Environ Microbiol 9*, 2486-2495.

148. Ravi, A. V., Musthafa, K. S., Jegathammbal, G., Kathiresan, K., and Pandian, S. K. (2007) Screening and evaluation of probiotics as a biocontrol agent against pathogenic Vibrios in marine aquaculture, *Lett Appl Microbiol 45*, 219-223.

149. Balcazar, J. L., and Rojas-Luna, T. (2007) Inhibitory Activity of Probiotic Bacillus subtilis UTM 126 Against Vibrio Species Confers Protection Against Vibriosis in Juvenile Shrimp (Litopenaeus vannamei), *Curr Microbiol 55*, 409-412.

150. Nakayama, T., Nomura, N., and Matsumura, M. (2006) Study on the relationship of protease production and luminescence in Vibrio harveyi, *J Appl Microbiol 101*, 200-205.

151. Nakayama, T., Ito, E., Nomura, N., Nomura, N., and Matsumura, M. (2006) Comparison of Vibrio harveyi strains isolated from shrimp farms and from culture collection in terms of toxicity and antibiotic resistance, *FEMS Microbiol Lett 258*, 194-199.

152. Vidgen, M., Carson, J., Higgins, M., and Owens, L. (2006) Changes to the phenotypic profile of Vibrio harveyi when infected with the Vibrio harveyi myovirus-like (VHML) bacteriophage, *J Appl Microbiol 100*, 481-487.

153. Faruque, S. M., Rahman, M. M., Asadulghani, Nasirul Islam, K. M., and Mekalanos, J. J. (1999) Lysogenic conversion of environmental Vibrio mimicus strains by CTXPhi, *Infect Immun 67*, 5723-5729.

154. Waldor, M. K., and Mekalanos, J. J. (1996) Lysogenic conversion by a filamentous phage encoding cholera toxin, *Science 272*, 1910-1914.

155. Wilkins, S., Millar, M., Hemsworth, S., Johnson, G., Warwick, S., and Pizer, B. (2007) Vibrio harveyi sepsis in a child with cancer, *Pediatr Blood Cancer*.

156. Masini, L., De Grandis, G., Principi, F., Mengarelli, C., and Ottaviani, D. (2007) Research and characterization of pathogenic vibrios from bathing water along the Conero Riviera (Central Italy), *Water Res 41*, 4031-4040.

157. Zabala, B., Garcia, K., and Espejo, R. T. (2009) Enhancement of UV light sensitivity of a Vibrio parahaemolyticus O3:K6 pandemic strain due to natural lysogenization by a telomeric phage, *Appl Environ Microbiol 75*, 1697-1702.

158. Cabello, F. C., Espejo, R. T., Hernandez, M. C., Rioseco, M. L., Ulloa, J., and Vergara, J. A. (2007) Vibrio parahaemolyticus O3:K6 epidemic diarrhea, Chile, 2005, *Emerg Infect Dis 13*, 655-656.

159. Fuenzalida, L., Armijo, L., Zabala, B., Hernandez, C., Rioseco, M. L., Riquelme, C., and Espejo, R. T. (2007) Vibrio parahaemolyticus strains isolated during investigation of the summer 2006 seafood related diarrhea outbreaks in two regions of Chile, *Int J Food Microbiol 117*, 270-275.

160. Fuenzalida, L., Hernandez, C., Toro, J., Rioseco, M. L., Romero, J., and Espejo, R. T. (2006) Vibrio parahaemolyticus in shellfish and clinical samples during two large epidemics of diarrhoea in southern Chile, *Environ Microbiol 8*, 675-683.

161. Gonzalez-Escalona, N., Cachicas, V., Acevedo, C., Rioseco, M. L., Vergara, J. A., Cabello, F., Romero, J., and Espejo, R. T. (2005) Vibrio parahaemolyticus diarrhea, Chile, 1998 and 2004, *Emerg Infect Dis 11*, 129-131.

162. Matsumoto, C., Okuda, J., Ishibashi, M., Iwanaga, M., Garg, P., Rammamurthy, T., Wong, H. C., Depaola, A., Kim, Y. B., Albert, M. J., and Nishibuchi, M. (2000) Pandemic spread of an O3:K6 clone of Vibrio parahaemolyticus and emergence of related strains evidenced by arbitrarily primed PCR and toxRS sequence analyses, *J Clin Microbiol 38*, 578-585.

163. Pertzev, A. V., and Nicholson, A. W. (2006) Characterization of RNA sequence determinants and antideterminants of processing reactivity for a minimal substrate of Escherichia coli ribonuclease III, *Nucleic Acids Res 34*, 3708-3721.

164. Lamontagne, B., and Elela, S. A. (2004) Evaluation of the RNA determinants for bacterial and yeast RNase III binding and cleavage, *J Biol Chem 279*, 2231-2241.

165. Spanggord, R. J., and Beal, P. A. (2001) Selective binding by the RNA binding domain of PKR revealed by affinity cleavage, *Biochemistry 40*, 4272-4280.

166. Ucci, J. W., Kobayashi, Y., Choi, G., Alexandrescu, A. T., and Cole, J. L. (2007) Mechanism of interaction of the double-stranded RNA (dsRNA) binding domain of protein kinase R with short dsRNA sequences, *Biochemistry 46*, 55-65.

167. Wu, H., Henras, A., Chanfreau, G., and Feigon, J. (2004) Structural basis for recognition of the AGNN tetraloop RNA fold by the double-stranded RNA-binding domain of Rnt1p RNase III, *Proc Natl Acad Sci U S A 101*, 8307-8312.

168. Leulliot, N., Quevillon-Cheruel, S., Graille, M., van Tilbeurgh, H., Leeper, T. C., Godin, K. S., Edwards, T. E., Sigurdsson, S. T., Rozenkrants, N., Nagel, R. J., Ares, M., and Varani, G. (2004) A new alpha-helical extension promotes RNA binding by the dsRBD of Rnt1p RNAse III, *Embo J 23*, 2468-2477.

169. Barouch, D. H. (2008) Challenges in the development of an HIV-1 vaccine, *Nature 455*, 613-619.

170. Moore, J. P., and Stevenson, M. (2000) New targets for inhibitors of HIV-1 replication, *Nat Rev Mol Cell Biol 1*, 40-49.

171. De Clercq, E. (2007) The design of drugs for HIV and HCV, *Nat Rev Drug Discov 6*, 1001-1018.

172. Bhattacharya, S., and Osman, H. (2009) Novel targets for anti-retroviral therapy, *J Infect 59*, 377-386.

173. Yuan, D., He, M., Pang, R., Lin, S. S., Li, Z., and Yang, M. (2007) The design, synthesis, and biological evaluation of novel substituted purines as HIV-1 Tat-TAR inhibitors, *Bioorg Med Chem 15*, 265-272.

174. Yang, M. (2005) Discoveries of Tat-TAR interaction inhibitors for HIV-1, *Curr Drug Targets Infect Disord 5*, 433-444.

175. Richter, S., Parolin, C., Gatto, B., Del Vecchio, C., Brocca-Cofano, E., Fravolini, A., Palu, G., and Palumbo, M. (2004) Inhibition of human immunodeficiency virus type 1 tat-trans-activation-responsive region interaction by an antiviral quinolone derivative, *Antimicrob Agents Chemother 48*, 1895-1899.

176. Richter, S. N., and Palu, G. (2006) Inhibitors of HIV-1 Tat-mediated transactivation, *Curr Med Chem 13*, 1305-1315.

177. Gatto, B., Tabarrini, O., Massari, S., Giaretta, G., Sabatini, S., Del Vecchio, C., Parolin, C., Fravolini, A., Palumbo, M., and Cecchetti, V. (2009) 2-Phenylquinolones as inhibitors of the HIV-1 Tat-TAR interaction, *ChemMedChem 4*, 935-938.

178. Narita, T., Yamaguchi, Y., Yano, K., Sugimoto, S., Chanarat, S., Wada, T., Kim, D. K., Hasegawa, J., Omori, M., Inukai, N., Endoh, M., Yamada, T., and Handa, H. (2003) Human transcription elongation factor NELF: identification of novel subunits and reconstitution of the functionally active complex, *Mol Cell Biol 23*, 1863-1873.

179. Marshall, N. F., Peng, J., Xie, Z., and Price, D. H. (1996) Control of RNA polymerase II elongation potential by a novel carboxyl-terminal domain kinase, *J Biol Chem 271*, 27176-27183.

180. Yamada, T., Yamaguchi, Y., Inukai, N., Okamoto, S., Mura, T., and Handa, H. (2006) P-TEFb-mediated phosphorylation of hSpt5 C-terminal repeats is critical for processive transcription elongation, *Mol Cell 21*, 227-237.

181. Fujinaga, K., Irwin, D., Huang, Y., Taube, R., Kurosu, T., and Peterlin, B. M. (2004) Dynamics of human immunodeficiency virus transcription: P-TEFb phosphorylates RD and dissociates negative effectors from the transactivation response element, *Mol Cell Biol 24*, 787-795.

182. Rice, A. P., and Herrmann, C. H. (2003) Regulation of TAK/P-TEFb in CD4+ T lymphocytes and macrophages, *Curr HIV Res 1*, 395-404.

183. Saunders, A., Core, L. J., and Lis, J. T. (2006) Breaking barriers to transcription elongation, *Nat Rev Mol Cell Biol 7*, 557-567.

184. Li, C. J., Friedman, D. J., Wang, C., Metelev, V., and Pardee, A. B. (1995) Induction of apoptosis in uninfected lymphocytes by HIV-1 Tat protein, *Science 268*, 429-431.

185. Westendorp, M. O., Frank, R., Ochsenbauer, C., Stricker, K., Dhein, J., Walczak, H., Debatin, K. M., and Krammer, P. H. (1995) Sensitization of T cells to CD95-mediated apoptosis by HIV-1 Tat and gp120, *Nature 375*, 497-500.

186. Pugliese, A., Vidotto, V., Beltramo, T., Petrini, S., and Torre, D. (2005) A review of HIV-1 Tat protein biological effects, *Cell Biochem Funct 23*, 223-227.

187. Sodroski, J., Patarca, R., Rosen, C., Wong-Staal, F., and Haseltine, W. (1985) Location of the trans-activating region on the genome of human T-cell lymphotropic virus type III, *Science 229*, 74-77.

188. Kuppuswamy, M., Subramanian, T., Srinivasan, A., and Chinnadurai, G. (1989) Multiple functional domains of Tat, the trans-activator of HIV-1, defined by mutational analysis, *Nucleic Acids Res 17*, 3551-3561.

189. Karn, J. (1999) Tackling Tat, *J Mol Biol 293*, 235-254.

190. Shojania, S., and O'Neil, J. D. (2006) HIV-1 Tat is a natively unfolded protein: the solution conformation and dynamics of reduced HIV-1 Tat-(1-72) by NMR spectroscopy, *J Biol Chem 281*, 8347-8356.

191. Aboul-ela, F., Karn, J., and Varani, G. (1996) Structure of HIV-1 TAR RNA in the absence of ligands reveals a novel conformation of the trinucleotide bulge, *Nucleic Acids Res 24*, 3974-3981.

192. Davidson, A., Leeper, T. C., Athanassiou, Z., Patora-Komisarska, K., Karn, J., Robinson, J. A., and Varani, G. (2009) Simultaneous recognition of HIV-1 TAR RNA bulge and loop sequences by cyclic peptide mimics of Tat protein, *Proc Natl Acad Sci U S A 106*, 11931-11936.

193. Aboul-ela, F., Karn, J., and Varani, G. (1995) The structure of the human immunodeficiency virus type-1 TAR RNA reveals principles of RNA recognition by Tat protein, *J Mol Biol 253*, 313-332.

194. Brodsky, A. S., and Williamson, J. R. (1997) Solution structure of the HIV-2 TAR-argininamide complex, *J Mol Biol 267*, 624-639.

195. Puglisi, J. D., Tan, R., Calnan, B. J., Frankel, A. D., and Williamson, J. R. (1992) Conformation of the TAR RNA-arginine complex by NMR spectroscopy, *Science 257*, 76-80.

196. Wei, P., Garber, M. E., Fang, S. M., Fischer, W. H., and Jones, K. A. (1998) A novel CDK9-associated C-type cyclin interacts directly with HIV-1 Tat and mediates its high-affinity, loop-specific binding to TAR RNA, *Cell 92*, 451-462.

197. Peng, J., Zhu, Y., Milton, J. T., and Price, D. H. (1998) Identification of multiple cyclin subunits of human P-TEFb, *Genes Dev 12*, 755-762.

198. Bieniasz, P. D., Grdina, T. A., Bogerd, H. P., and Cullen, B. R. (1999) Recruitment of cyclin T1/P-TEFb to an HIV type 1 long terminal repeat promoter proximal RNA target is both necessary and sufficient for full activation of transcription, *Proc Natl Acad Sci U S A 96*, 7791-7796.

199. Fujinaga, K., Taube, R., Wimmer, J., Cujec, T. P., and Peterlin, B. M. (1999) Interactions between human cyclin T, Tat, and the transactivation response element (TAR) are disrupted by a cysteine to tyrosine substitution found in mouse cyclin T, *Proc Natl Acad Sci U S A 96*, 1285-1290.

200. Anand, K., Schulte, A., Fujinaga, K., Scheffzek, K., and Geyer, M. (2007) Cyclin box structure of the P-TEFb subunit cyclin T1 derived from a fusion complex with EIAV tat, *J Mol Biol 370*, 826-836.

201. Baumli, S., Lolli, G., Lowe, E. D., Troiani, S., Rusconi, L., Bullock, A. N., Debreczeni, J. E., Knapp, S., and Johnson, L. N. (2008) The structure of P-TEFb (CDK9/cyclin T1), its complex with flavopiridol and regulation by phosphorylation, *EMBO J 27*, 1907-1918.

202. Garber, M. E., Wei, P., KewalRamani, V. N., Mayall, T. P., Herrmann, C. H., Rice, A. P., Littman, D. R., and Jones, K. A. (1998) The interaction between HIV-1 Tat and human cyclin T1 requires zinc and a critical cysteine residue that is not conserved in the murine CycT1 protein, *Genes Dev 12*, 3512-3527.

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

204. Chen, D., Fong, Y., and Zhou, Q. (1999) Specific interaction of Tat with the human but not rodent P-TEFb complex mediates the species-specific Tat activation of HIV-1 transcription, *Proc Natl Acad Sci U S A 96*, 2728-2733.

205. Fujinaga, K., Irwin, D., Taube, R., Zhang, F., Geyer, M., and Peterlin, B. M. (2002) A minimal chimera of human cyclin T1 and tat binds TAR and activates human immunodeficiency virus transcription in murine cells, *J Virol 76*, 12934-12939.

206. Bi, Y., Tang, Y., Raleigh, D. P., and Cho, J. H. (2006) Efficient high level expression of peptides and proteins as fusion proteins with the N-terminal domain of L9: application to the villin headpiece helical subdomain, *Protein Expr Purif 47*, 234-240.

207. Anand, K., Schulte, A., Vogel-Bachmayr, K., Scheffzek, K., and Geyer, M. (2008) Structural insights into the cyclin T1-Tat-TAR RNA transcription activation complex from EIAV, *Nat Struct Mol Biol 15*, 1287-1292.

208. Leroux, C., Cadore, J. L., and Montelaro, R. C. (2004) Equine Infectious Anemia Virus (EIAV): what has HIV's country cousin got to tell us?, *Vet Res 35*, 485-512.

209. Richter, S., Cao, H., and Rana, T. M. (2002) Specific HIV-1 TAR RNA loop sequence and functional groups are required for human cyclin T1-Tat-TAR ternary complex formation, *Biochemistry 41*, 6391-6397.