ABSTRACT

Title of Dissertation:	BIOCHEMICAL AND STRUCTURAL CHARACTERIZATION OF NUSG PARALOG LOAP
	Amr Tarek Ahmed Ragab Elghondakly Doctor of Philosophy, 2021
Dissertation directed by:	Professor Wade C. Winkler, Department of Cell Biology and Molecular Genetics

The NusG family of transcription factors is the only universally conserved family of transcription elongation regulators in all three domains of life. NusG proteins exert ubiquitous genetic regulatory effects by reversibly binding RNA-polymerase (RNAP) during transcription elongation and modulate its function. A phylogenetic analysis of the NusG family of proteins identified several distinct subfamilies of NusG paralogs that are widespread amongst bacterial species. These different NusG paralogs are likely to exert regulatory control over distinct subsets of genes. Yet, despite the importance of the genes they regulate, most of the subfamilies of NusG paralogs (*e.g.*, UpxY, TaA, ActX and LoaP) have not been investigated in depth. Additionally, the regulatory mechanisms that these transcription elongation factors employ are likely to differ between one another to allow for specific recruitment to target operons and prevent competition with the housekeeping NusG factor. The LoaP subfamily of NusG proteins is primarily encoded by Actinobacteria,

Firmicutes and Spirochaetes. While regulons for the LoaP subfamily have only been identified in a few organisms, the *loaP* gene is oftentimes found adjacent to long operons encoding for biosynthesis of secondary metabolites suggesting a regulatory relationship with these pathways. In *Bacillus velezensis*, LoaP promotes transcription antitermination of two long biosynthetic operons which encode for two different polyketide antibiotics: difficidin and macrolactin. Intriguingly, the cis-determinants for LoaP antitermination include a small RNA hairpin (~26 nts) located within the 5' leader region of target operons. LoaP associates with the RNA hairpin *in vitro* with nanomolar affinity and high specificity via basic residues that are highly conserved within the C-terminal KOW domain, in contrast to other well-characterized bacterial NusG proteins which do not exhibit RNA-binding activity. These data indicate that LoaP employs a distinct regulatory mechanism to achieve targeted regulation of large biosynthetic operons in bacteria. Furthermore, this discovery expands the repertoire of macromolecular interactions exhibited by bacterial NusG proteins during transcription elongation to include an RNA ligand. Crystallographic studies of LoaP-RNA complex are in progress, and recent results will be discussed.

BIOCHEMICAL AND STRUCTURAL CHARACTERIZATION OF NUSG PARALOG LOAP

by

Amr Tarek Ahmed Ragab Elghondakly

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2021

Advisory Committee:
Professor Wade Winkler, Chair
Professor Theodore Kwaku Dayie
Professor Nicole LaRonde
Professor David Fushman
Professor Myles Poulin
Professor Daniel Nelson

Dedication

I dedicate this dissertation among five: to *teta*, my mother, and Hadeir; to Dawlat and Sylvienne. The women whose sacrifices buoyed me across treacherous waters, and the women without whom *arriving* would not have been possible.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

DNA: Deoxyribonucleic acid

DTT: 1,4-Dithiothrietol

IDT: Integrated DNA technologies

IPTG: Isopropyl ß-D-1-thiogalactopyranoside

K_d: Equilibrium dissociation constant

LB: Lysogeny broth

MBP: Maltose-binding protein

mM: Millimolar

NCBI: National Center for Biotechnology Information

nM: Nanomolar

Nus: N-utilization substance

OD_{600nm}: Optical density at 600 nm

PAGE: Polyacrylamide gel electrophoresis

PCR: Polymerase chain reaction

PEG: Polyethylene Glycol

PMSF: Phenylmethylsulfonyl fluoride

RNA: Ribonucleic acid

rNTP: Ribonucleotide triphosphate

SDS: Sodium dodecyl sulfate

TOF LC/MS: Time of flight liquid chromatography-tandem mass spectrometry

μM: micromolar

Chapter 1: Introduction

Information processing pathways in bacteria

A key challenge of modern microbiologists is to discover the range of genetic regulatory mechanisms that can be employed by bacteria. By uncovering the molecular mechanisms that bacteria use for regulation of their genes, better strategies can be developed for synthetic biology applications and for targeting of novel drugs to biomedically relevant microbes. Since the discovery of penicillin, secondary metabolites have had a major impact on human health and are a key target of modern drug discovery. In fact, microbes continue to be an important source of new bioactive compounds for the pharmaceutical industry. Microbial natural products account for half of all commercially available pharmaceuticals (1), such as anticancer drugs (*e.g.*, bleomycin, dactinomycin, doxorubicin, and staurosproin) (2), immunosuppressants (*e.g.*, cyclosporine, rapamycin) (3), and antimycotic agents, (*e.g.*, anidulafungin) (4). Currently, with less than 1% of the microbial world having been explored, there is a growing interest in searching for microbial sources that may produce new classes of antibiotics and bioactive secondary metabolites (5).

While most bacteria produce a few bioactive secondary metabolites, only a few phylogenetic branches include bacteria that are truly prolific producers. Yet, it is still difficult to acquire these compounds, as their direct extraction from native hosts is usually a cumbersome task. This is because microorganisms, which are adept at synthesizing secondary metabolites, are usually difficult to culture in the laboratory.

Furthermore, perhaps due to the high energetic cost required for the synthesis of these structurally-complex compounds, secondary metabolites are not constitutively synthesized in bacteria (6). Instead, they are often produced from cryptic genetic clusters that are transcriptionally inactive under standard laboratory conditions.

The quest for producing biomedically important secondary metabolites from within heterologous host cells might be a viable approach to overcome the limitations associated with microorganisms that are difficult to culture under standard laboratory conditions. A major advantage of this approach is that it relies on heterologous expression of secondary metabolites using widely studied bacteria. A second advantage is that the explosion of genome sequencing in the past decade (7-10) has provided a vast library of potential secondary metabolite gene clusters. However, while this approach is logical, it has proved to be largely unsuccessful. Simply put, the heterologous expression of secondary metabolite synthesis gene clusters usually fails. This might be due to a lack of understanding of the genetic regulatory mechanisms that oversee the production of secondary metabolites.

Bacteria have evolved a multitude of mechanisms to regulate gene expression in response to their continuously changing environment. Controlled regulation of gene expression means prompting appropriate genetic responses to environmental and metabolic signals within a specified time frame – a process that is essential to the overall fitness of bacterial cells throughout their life cycle. While all stages of bacterial gene expression can be subject to regulatory control (11-14), the transcription apparatus is a frequent target. The process of transcription is driven by RNA-polymerase (RNAP), which is a highly conserved enzyme among bacteria.

RNAP typically consists of four subunits (α, α, β, β') with $\alpha_2\beta\beta'$ stoichiometry (15) constituting the core enzyme *in vitro*.

The transcription cycle is a multistep process consisting of initiation, elongation, and termination steps (14, 16-20). Although the core enzyme is catalytically active, it is unable to initiate transcription at promoter sites without the assistance of an additional set of proteins known as sigma factors (11). This is because DNA-binding proteins like RNAP bind genomic DNA in a non-specific manner (21), and therefore, in that state, they are not available for promoter-directed transcription. Sigma factors are highly dynamic proteins comprising approximately four independent domains connected together via flexible linkers, with each domain playing a different role (22). The modular structure of these domains is essential in the formation of the competent initiation complex. Sigma domains bind a conserved location on the enzyme's surface and simultaneously associate with characteristic sequences in promoter DNA, thereby positioning RNAP at the correct start site (17, 22). An additional function of sigma factors is to help trigger unwinding of the double-stranded DNA template at transcription start sites. This critical role ensures that template DNA enters RNAP active sites only when it's bound to correct promoter sequences (23-25).

The most prevalent mechanism by which transcription is regulated in bacteria occurs at the initiation step (26) through the association of sigma factors with RNAP and through association of DNA-binding proteins that act as activators or repressors (27). All bacteria encode for a predominant housekeeping sigma factor (22) that is required for expression of the majority of genes. Additionally, nearly all bacteria

encode for one or more alternative sigma factors (11) that – when expressed – redirect a subset of cellular RNAP to alternative promoter sites, otherwise not recognized by the house keeping sigma factor. As a consequence, bacterial gene expression can be modulated by changing the set of promoters to which RNAP can bind, leading to subsequent changes in the transcriptional output.

The ubiquity of transcription regulation at the initiation step by transcription factors, however, might have detracted attention from additional post-initiation regulatory mechanisms. This is perhaps, in part, due to a lack of knowledge of the mechanistic diversity of post-initiation regulatory mechanisms – especially as investigations of transcription regulatory mechanisms in non-model bacterial organisms are presently lacking. Yet, many examples of such mechanisms have been described previously (28-30), and they have been shown to exert stringent regulatory control over gene expression. Moreover, post-initiation regulatory mechanisms are comprised of a diverse set of strategies that are mediated via auxiliary protein factors, *cis*-acting signal-responsive RNA elements, non-coding RNA, and trans-encoded small RNAs. Of particular importance, several examples of elongation-based regulatory mechanisms have been shown to regulate polyketide synthase pathways (31, 32), polysaccharide production (33, 34), and phage gene expression (35).

Post-initiation transcription regulation

Termination signals can serve as regulatory checkpoints

Transcription is the first step in the information processing pathway in all living organisms; therefore, each step in the transcription process is subject to stringent genetic regulatory mechanisms (36). After RNAP escapes the promoter site,

it undergoes conformational changes to transition into the transcription elongation complex (TEC) (37). The transition into the elongation state is demarcated by structural rearrangement of RNAP subunits leading to the formation of an exit channel for nascent RNA (38, 39) and subsequent dissociation of sigma factors (40, 41). TECs are very stable at most positions along the DNA template as they synthesize RNA one nucleotide at a time. Their stability is critical during transcription elongation since some operons in bacteria can comprise up to 10^5 nucleotides in length (36, 42). As the transcribing RNAP lacks the ability to reinitiate transcription from truncated RNA molecules longer than 2-4 nucleotides in length (43), TECs must transcribe entire operons in a single attempt.

Transcription terminates when the elongation complex encounters a termination signal that triggers complex disassembly and RNA release (44, 45). Termination signals in bacteria are broadly characterized as either Rho-dependent (46) or intrinsic terminators (45). Both processes employ distinct mechanistic strategies, which disrupt the stability of the elongation complex to terminate transcription. Rho-dependent termination requires the recruitment of a hexameric ATP-dependent RNA translocase at Rho utilization sites (*rut*), which lack a consensus sequence (47, 48), but are primarily characterized by cytosine-rich regions. Once recruited, Rho utilizes energy from ATP hydrolysis to translocate along RNA transcript in 5'-3' direction towards the elongating RNAP (49, 50). It is possible that Rho mechanically disrupts the elongation processes by "pulling" at RNA from the transcription bubble, resulting in transcription termination (51). This model of Rho-dependent termination therefore necessitates that the translocation rate of Rho needs

to exceed that of the elongation complex in order for physical interactions to be established within an operative timeframe. However, alternative proposals (52) argue that such a necessity for a fast translocation rate is not an absolute requirement, and it is more likely that the elongation complex encounters pause sites which effectively stalls RNAP along the DNA template, thereby providing enough time for Rho to "catch-up". This proposed model is further supported by biochemical evidence (53) indicating that the elongation factor NusG – which remains associated with RNAP throughout the elongation phase – exhibits a measurable binding affinity for Rho and aids in Rho recruitment (54).

In contrast, intrinsic terminators are encoded nucleic acid sequences that function independently in their RNA form (55), and do not necessarily require the participation of additional factors. However, many intrinsic terminators are further enhanced by NusA, an elongation factor that associates to RNAP near the RNA exit tunnel (56-58), and they may also be influenced by an additional transcription factor called NusG (59). Intrinsic terminators are characterized by a GC-rich RNA hairpin that forms in the emerging transcript, which is immediately followed by a downstream poly-uridine tract. It is worth noting that neither the poly-uridine tract or terminator hairpin alone is sufficient to trigger transcription termination (60, 61). Instead, the poly-uridine tract forms weak, consecutive Watson and Crick pairings with template DNA (rU/dA) inside the transcription bubble, which partially induces transcriptional pausing (62), and sometimes causes transcript slippage. Pausing of the TEC at the poly-uridine tract provides sufficient time for the formation of RNA terminator hairpin within the exit tunnel, which together trigger dissociation of TEC

(45) and effectively terminate transcription. Moreover, it is thought that the energy provided by ATP hydrolysis to destabilize the elongation complex in Rho-dependent termination is somewhat equal to the energy of folding for intrinsic terminator hairpins (51), whose formation shortens the DNA/RNA hybrid in the catalytic center to disrupt the elongation complex at its most critical site.

In addition to their critical role in the terminal stage of the transcription cycle, termination signals can also strongly impact the transcriptional outcome and subsequently alter gene expression. Typically, the location of termination signals at the end of operons demarcate gene boundaries and prevent unintended transcription of downstream genes. However, when intrinsic terminators are located in leader regions – between transcription start sites and downstream genes – or in the intergenic space, they serve as regulatory checkpoints with the efficiency of termination often modulated by specific regulators (30). As a consequence, bacterial gene expression can be fine-tuned in a non-binary mode based on the probability of the termination event. Termination regulation therefore is a frequent target of post-initiation regulatory mechanisms (61), which only recently have been found to be more prevalent and mechanistically diverse than previously appreciated.

Different modes of transcription termination

Conditional transcription termination can be regulated by two distinct types of mechanisms: transcription attenuation (63), and transcription antitermination (61). These mechanisms control the efficiency of termination by tricking the elongation complex into bypassing regulatory termination signals, thereby allowing genes downstream of terminators to be expressed. The major differences between both types

of regulatory control, however, lie in the number of regulated terminators as well as the identity of the regulatory targets. Transcription attenuation mechanisms directly regulate individual terminators by promoting the formation of mutually exclusive alternative structures. In the typical transcription attenuation mechanism (30), a signal-responsive, cis-acting regulatory RNA can attenuate activity at a specific termination signal located immediately downstream of the signal-integration site. Many classes of cis-acting regulatory RNAs (28, 29), such as riboswitches, have been discovered to control transcription attenuation. These regulatory systems couple transcription attenuation to a diverse range of signals, including but not limited to: paused ribosomes, uncharged tRNAs, RNA-binding proteins, metabolites, metal ions, and changes in temperature (reviewed in refs (64-66)).

Role of general transcription factors in processive antitermination

Transcription attenuation mechanisms fundamentally differ from antitermination mechanisms. While transcription attenuation mechanisms exert their regulatory influence over individual termination sites, processive antitermination systems occur when the elongating complex is modified to become broadly resistant to termination signals (fig. 1-1) (36, 51, 61) and is less sensitive overall to pause sites (60). In these instances, the recruitment of specialized factors to the elongation complex promote readthrough of sequential intrinsic terminators over long genomic distances as well as antagonize Rho recruitment; therefore, under these conditions, transcription elongation proceeds past Rho-dependent and intrinsic termination signals.

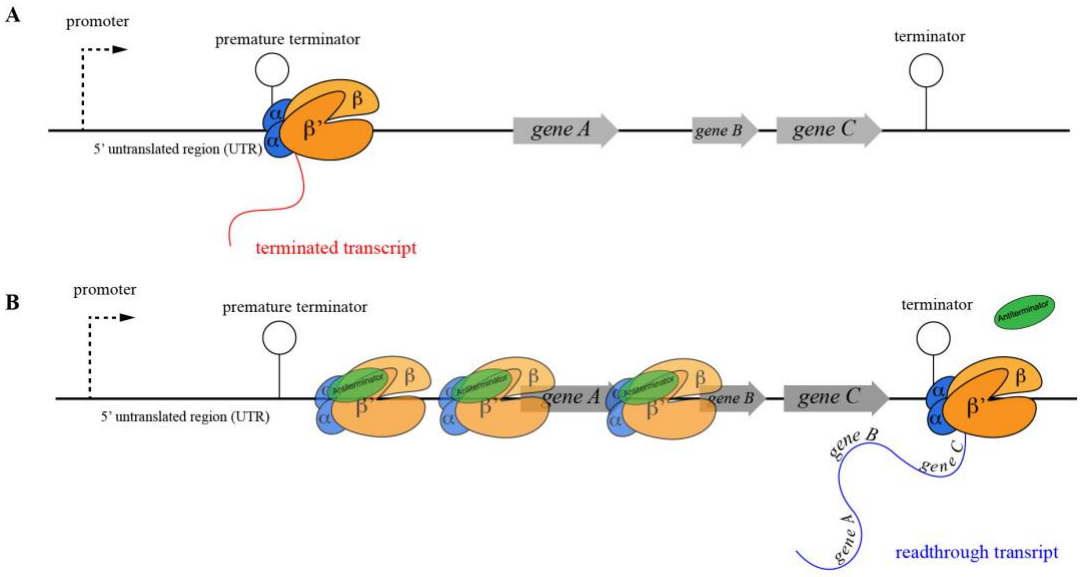


Figure 1-1. Schematic representation of a processive antitermination mechanism. (A) Premature termination occurs when the elongation complex encounters an intrinsic terminator within the 5' untranslated region (UTR), thereby silencing downstream genes. (B) Antitermination factors are recruited to the elongation complex to bypass these premature intrinsic terminator signals.

Intriguingly, processive antitermination factors induce structural modifications to transform RNAP into a termination-resistant state. These structural modifications are often orchestrated in concert with RNAP-associated general elongation factors NusA, and NusG. NusA and NusG factors are highly conserved regulators in bacteria (67, 68), and they participate in the majority of post-initiation regulatory mechanisms.

These host-encoded factors routinely associate with RNAP during the elongation phase to modulate the rate of RNA synthesis, facilitate transcription-translation coupling, or all together alter the transcriptional outcome. Moreover, these factors interact independently with the elongation complex, as indicated by their distinct binding sites on the surface of RNAP; however, they can act either concertedly or antagonistically (67) to modify the properties of the elongation complex.

E. coli NusA is a monomeric protein comprising of six functional domains: N-terminal domain, RNA-binding domains (S1, KH1, KH2 – commonly abbreviated SKK), and an autoinhibitory C-terminal domain (69-71). The most discernible role of NusA during the elongation phase is to enhance transcriptional pausing (57), and to increase the termination efficiency of intrinsic terminators (57, 72). However, NusA also participates in antitermination mechanisms, RNA-folding, and Rho-dependent termination (73, 74). NusA-NTD reversibly binds the β flap domain of RNAP at the RNA exit channel, in close proximity to the emerging RNA transcript. NusA SKK domains together form a compact RNA-binding domain that interacts with nascent RNA to aid in modulating pausing and termination events.

In *E. coli* and other γ-proteobacteria, the NusA SKK domain is autoinhibited in solution by an acidic repeat domain AR2, located in the C-terminal portion, particularly in the absence of the elongation complex. Once bound to RNAP, this autoinhibition is released via interactions between AR2 and CTD of α-subunit of RNAP (75); therefore, the interactions between the SKK domain and RNA only occur when NusA is bound to the elongation complex.

While the mechanistic details of NusA-induced pausing remain to be fully uncovered, it is commonly accepted that NusA promotes transcriptional pausing by promoting the formation of small RNA hairpins at the exit channel via its RNA-binding domain (SKK) (51), or by inducing allosteric changes to the active site of RNAP. Indeed, in the Gram-positive bacterium *Bacillus subtilis*, NusA broadly affects transcription termination at many locations in the genome (57) and particularly enhances the efficiency of intrinsic termination *in vivo* and *in vitro* at weak terminator hairpins, which either contain a short stem or an interrupted poly-uridine tract – both of which reduce the termination efficiency of intrinsic terminators.

Contrarily, the functional roles of NusG factors are generally orthogonal to that of NusA. In particular, NusG regulators often suppress pausing (76), enhance processivity (76, 77), and facilitate transcription-translation coupling (78). These regulators comprises a large family of proteins, and they are the only conserved family of elongation factors in all three domains of life (79-81). Bacterial NusG proteins typically contain two flexible domains: a unique N-terminal (NGN) domain, and a flexibly connected, C-terminal Kyrpides-Ouzounis-Woese (KOW). The NusG

NGN domain binds to the surface of the elongating RNAP near the conserved clamp helices of the β' subunit and the gate loop of β subunit, thus arranging NusG at the solvent exposed surface of the transcription bubble in close proximity to the non-template (NT) DNA (77, 82). By binding at this location, NusG generally forms a processivity clamp to confine nucleic acids to RNAP in a closed, pause-resistant state (76), and to prevent backtracking of RNAP (83). Whereas NusG NTD alone can partially promote processivity and suppress pausing (84), the flexibly-linked CTD acts as an interaction platform (85), which mediates contacts with cotranscriptional factors ultimately resulting in disparate regulatory outcomes. Surprisingly, some CTD-mediated interactions exert regulatory effects on transcription that are orthogonal to the NusG processivity enhancing role. For example, in *E. coli*, NusG silences gene expression of horizontally transferred genes and foreign DNA by directly interacting with Rho terminator to facilitate premature transcription termination (52, 86).

The modular structure of NusG regulators underpins their versatility as transcription factors and their functional importance in transcription regulation. The interaction network of NusG CTD includes ribosomal protein NusE (S10), NusA, Rho terminator, and phage lambda N antiterminator. The interaction between NusE (S10) and NusG helps bridge interactions between the leading ribosome and RNAP, as part of the expressosome, which is the large complex formed between the TEC and the 'leading ribosome', formed during transcription-transcription coupling (78, 87-89). Interestingly, neither domain alone is sufficient to promote transcription-translation coupling (88), Rho termination, or phage lambda N antitermination

suggesting a functional role embedded in the conserved structure of NusG proteins. Therefore, the functional versatility of NusG, along with its broad conservation in all three domains of life, make NusG-like proteins attractive targets for study of post-initiation transcription regulation.

Processive antitermination mechanisms

Phage λN antitermination

Only a few examples of processive antitermination complexes have been discovered thus far, which, although they share some conceptual similarities, differ in their molecular mechanisms and functional outcomes. Processive antitermination mechanisms often incorporate auxiliary transcription factors that associate with the elongation complex, particularly with host-encoded Nus factors, to suppress pausing and premature termination. The first processive antitermination mechanism was discovered in bacteriophage λ (90). During bacteriophage λ lytic phase, early phage gene expression is temporally controlled via a conditional transcription termination event where regulatory terminators (91), located downstream of two promoter sites P_L (leftward) and P_R (rightward), prevent expression of downstream phage genes. The first gene downstream of the P_L promoter precedes a regulatory termination signal in the operon and encodes for a small intrinsically disordered protein called λN. Once expressed, λN protein is recruited to the elongating *E. coli* RNAP where it interacts with host-encoded elongation factors NusA, NusB, NusE and NusG (92, 93). Nus factors typically do not promote processive antitermination on their own, but they are required for the formation of the specialized λN-mediated antitermination

complex. Immediately after λN recruitment, the bacterial transcription apparatus is kinetically reprogrammed to bypass downstream terminators, thereby activating gene expression of essential phage genes (94, 95).

The assembly of the λN antitermination complex occurs at two genetically defined *nut* recruitment sites (96) in the phage chromosome, which are functional in their RNA form (91, 94, 97), and are comprised of two components (92). The λN protein directly recognizes a 15-nucleotide component known as *boxB*, which forms a GNRA-type RNA hairpin structure (98). The second component, *boxA*, is a short RNA sequence that serves as a binding site for a heterodimer of NusE:NusB (99, 100), thereby providing additional stabilization to the complex. Both of these ribonucleoprotein (RNP) complexes remain attached to TEC as the nascent RNA loops out of the exit channel (97), and are further stabilized by host elongation factors NusG and NusA. Remarkably, the *E. coli* λN-associated antitermination complex retains antitermination activity over long genomic distances. In fact, in instances where phage λ is integrated into the bacterial chromosome, λN-mediated antitermination can proceed up to tens of kilobases downstream of the recruitment site, well past the phage genes into the bacterial chromosome (91), bypassing numerous Rho-dependent and intrinsic termination signals.

The persistent antitermination activity conferred through the association of λN with the elongation complex is a result of a structural reorganization strategy that ultimately transforms the TEC into a termination-resistant state. Though structurally disordered in its free form (101), λN adopts an elongated α-helical conformation composed of three α helices (α1-3) flexibly connected by disordered regions, upon its

recruitment to the elongation complex. In addition to binding *boxB* RNA, λN establishes specific contacts with the elongation complex at several critical sites (102). The λN central α3 helix forms a stable complex with the NusA NTD and with the RNAP β flap tip helix (FTH), which is located close to the rim of the RNA exit channel and routinely interacts with RNA hairpins during transcription elongation. Through λN-mediated interactions, the position of NusA is altered in the λN antitermination complex, as compared to the standard TEC, causing it to be moved approximately 45° away from the RNA exit tunnel (101). Moreover, the λN α3 helix remodels the conformation of RNAP β flap tip helix (FTH) as well, potentially interfering with interactions with the emerging transcript. In this arrangement, nascent RNA is redirected out of the RNA exit channel along an alternative pathway that is guided by the repositioned NusA S1 domain towards *nut* RNA elements (see fig. 1-2).

Conversely, in the standard TEC, nascent RNA is guided by the NusA NTD along a pathway flanked by the β' dock domain on one side of the exit channel and both the β' zinc finger and β flap tip domain on the other (56), favoring the formation of RNA secondary structures (103, 104), particularly intrinsic terminator hairpins. Intriguingly, the steric hinderance imposed on the emerging transcript along its alternate pathway impedes the formation of RNA secondary structures as one strand of the intrinsic terminator would be distally positioned from its complementary strand (101). The repositioning of NusA by λN away from the RNA exit channel therefore not only subverts the pause-enhancing role of NusA by disrupting canonical NusA-mediated interactions, but it astonishingly reprograms NusA from a termination factor

into an antitermination factor, all without covalently modifying the protein or altering its tertiary structure.

Additional λN-mediated interactions provide further stabilization to the antitermination complex. The NusA KH2 domain associates favorably with the λN α1:*boxB* complex, affixing *boxB* RNA to the antitermination complex for an unspecified length of time, during which the elongation complex adopts a processive termination-resistant state (97). While λN RNA-binding activity is central to the formation of the antitermination complex, the bent α helical structure also establishes multiple interactions as it meanders through the elongation complex. For example, the α2 helix rests in a crevice formed between the surfaces of NusA NTD-S1-KH1 domains and NusE (101). The latter binds specifically to *boxA* RNA element as well as NusG CTD. Indeed, in the λN antitermination complex, the position of NusG CTD is redirected from the site it typically occupies towards the binding interface of globular NusE (105), thereby providing additional structural stabilization to the λN-NusA-NusE conformation.

Consequently, this sequestration of the NusG CTD is likely to provide a structural basis for preventing Rho-dependent termination. This is because interactions of the NusG CTD with NusE and Rho are mutually exclusive (106). Alternatively, the λN antitermination complex simply offers less free nascent RNA to act as a substrate for Rho-binding. Furthermore, it is possible that the accelerated rate of the antitermination complex along the DNA template may assist with inhibition of Rho-dependent termination. Perhaps it is more challenging for Rho to "catch-up"

with the antitermination complex and establish the required physical contacts to trigger efficient termination.

In summary, the λN acts like a "molecular thread" that stitches host-encoded elongation factors to the elongating RNAP in a distinct arrangement. By doing so, the elongation complex is transiently transformed into a termination-resistant form that can bypass numerous termination signals. In this context, λN mediates a broad network of molecular interactions that extends the longevity of processive antitermination by kinetically-controlling the dissociation rate of the λN-*nut* RNA-Nus factors-RNAP complex (97), allowing processive antitermination to persist for long distances over the transcription timescale.

Phage λQ antitermination

Bacteriophage λ encodes for an additional antitermination factor known as λQ. λQ is expressed during early-middle phage growth cycle, and it orchestrates the switch from middle to late phage gene expression via a processive antitermination event (60, 107). Though functionally similar to λN, λQ-mediated antitermination is mechanistically distinct, particularly with respects to its recognition sites and the mechanism of association with the elongation complex. Whereas λN binds the elongation complex at *nut* sites following promoter escape and the subsequent synthesis of *boxB* RNA, λQ is a DNA-binding protein that is loaded to a σ_{70}-associated paused elongation complex.

It is generally accepted that the sigma factor dissociates from RNAP shortly after initiation, during the transition to the elongation complex (108). It is replaced

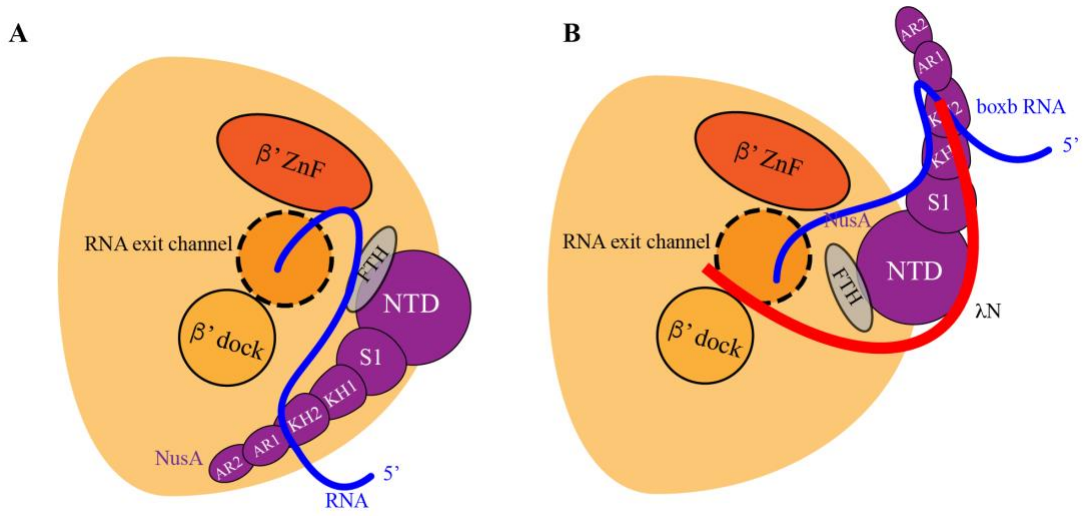


Figure 1-2. Schematic diagram showing the structural basis of λN mediated antitermination. (A) NusA protein binds the elongation complex at a conserved binding site near the RNA exit channel. The emerging RNA transcript is guided through a path that is flanked by the NusA NTD and the β' dock domain. NusA NTD facilitates secondary structure formation of intrinsic terminators by positioning one strand of the terminator in close proximity to its complementary strands, thereby enhancing termination efficiency. (B) Association of λN antiterminator repositions NusA ~40° away from the RNA exit channel, relocating the S1 domain, and displaces FTH domain thereby forging an alternative pathway for nascent RNA out of the exit channel. Along this alternative pathway, nascent RNA interacts directly with the S1 domain to impose steric hinderance on secondary structure formation, as one strand of the intrinsic terminator is sequestered by the NusA S1 domain, thereby bypassing terminator signals. *Adapted from (101).*

by the transcription elongation factor NusG, as they share overlapping binding locations on the surface of RNAP. In the absence of λQ, the elongation complex pauses at a specific σ_{70}-dependent pause site, and then terminates at an intrinsic terminator immediately downstream of the pause site. In contrast, association of λQ to the RNAP renders the elongation complex resistant to intrinsic termination over long genomic distances. However, recruitment of λQ alone is insufficient for the formation of the antitermination complex. Instead, λQ-mediated antitermination is dependent on two *cis*-acting elements: a -10-like sequence, which acts as the σ_{70}-dependent pause site immediately downstream of the promoter, and a λQ-binding element (QBE) (109), which is embedded in the promoter region. Genetic mutations in the wild-type pause site lead to complete inhibition of antitermination activity (109). Interestingly, artificially pausing λQ-associated RNAP via nucleotide deprivation fails to produce an antitermination complex (60, 109) suggesting that the specific substrate for λQ-mediated antitermination is the σ_{70}-associated, paused elongation complex (PEC).

These biochemical findings were recently bolstered by structural data obtained from Cryo-EM structures of bacteriophage 21 Q-associated antitermination complex (110, 111). It is worth noting that the structural basis for Q21-mediated antitermination might be specific to that particular bacteriophage subtype, and it is yet to be determined whether λQ adopts a similar antitermination mechanism. However, both λQ and Q21 antiterminators belong to the same protein family and perform equivalent functions. Regardless, the structure of Q21-associated antitermination complex reveals a simple yet elegant way by which processive antitermination is

achieved. While the Cryo-EM structure indicates that Q21 confers antipausing activity by preventing RNAP swiveling during transcription elongation, processive antitermination occurs primarily because Q21 forms a molecular "nozzle" that towers above – and partially extends inside – the RNA exit channel effectively narrowing its diameter (111). Narrowing and extending the exit channel imposes steric restrictions on emerging nascent RNA, and therefore inhibits formation of pause and terminator RNA hairpins. Another striking aspect of this antitermination mechanism is that the Q21 molecular nozzle forms very stable interactions with the elongation complex and remains associated over tens of thousands of base pairs.

Ribosomal RNA antitermination

Although processive antitermination systems were initially discovered in lambdoid phage operons, researchers have also identified a few other examples of similar antitermination mechanisms, which appear to regulate diverse sets of genes in bacteria. Bacterial operons that encode for long non-coding RNA, such as ribosomal RNA (rRNA) operons, are often targets of premature Rho-dependent termination. As discussed earlier, Rho-dependent termination occurs when hexameric Rho terminator binds *rut* sites, and translocates along nascent RNA until it reaches RNAP, at which point transcription terminates. Recruitment of Rho at *rut* sequences is strongly facilitated by lack of RNA secondary structures, and more importantly, by the absence of a leading ribosome (*i.e.*, the first ribosome to translate an mRNA, which has also been proposed to associate with the TEC). This is thought to antagonize Rho recruitment by blocking *rut* sites. Since rRNA operons are not translated, and

therefore lack a leading ribosome, rRNA transcripts must be protected from premature Rho-dependent termination by other mechanisms.

E. coli rRNA operons are subjected to transcription regulation by an antitermination mechanism that suppresses Rho termination to ensure complete rRNA synthesis and maturation. Current evidence suggest that the rRNA antitermination complex also influences rRNA folding, and, in conjunction with RNase III (112), is involved in rRNA processing. In *E. coli*, all seven rRNA operons, which comprises the genes coding for 16S, 23S, and 5S rRNA (113), contain a highly conserved RNA element in the leader and spacer regions that closely resembles the λN *boxA* RNA element. Moreover, the leader regions in these operons encode for *boxB*-like hairpin in addition to a linear less well characterized *boxC* sequence; however, only *boxA* appears to be required for processive antitermination.

Assembly of the rRNA antitermination complex occurs at *boxA* site in a manner that shares some similarity to the λN antitermination mechanism. Specifically, similar to the λN antitermination mechanism, the NusE:NusB heterodimer binds *boxA* RNA and interacts directly with NusG CTD (78, 114, 115) to discourage recruitment of Rho. However, these factors alone are insufficient to promote antitermination. When purified Nus factors are added to a halted RNAP, antitermination activity is still strongly stimulated by addition of cell extracts (113). This observation suggests that additional factors are required for reconstitution of rRNA antitermination complexes *in vitro*.

Recent studies (113, 116) revealed two candidates that associate with rRNA antitermination complexes: ribosomal protein S4 and inositol monophosphatase

SuhB. Intriguingly, addition of SuhB alone is sufficient to stimulate suppression of Rho-dependent termination, indicating that SuhB interacts directly with Nus factors or RNAP to promote antitermination.

The role of SuhB in rRNA antitermination was unexpected, as the *E. coli suhB* gene was first identified as a suppressor of temperature-sensitive mutations in protein export (117), heat shock stress response (118), and DNA replication (119). Yet, addition of SuhB alone – with or without S4 ribosomal protein – to transcription reactions containing RNAP, Nus factors, and Rho terminator significantly delayed or suppressed premature Rho-dependent termination. The antitermination activity conferred by addition of SuhB is unsurprisingly dependent on RNAP-associated Nus factors – particularly NusA –, however, the exact role of SuhB is still not well understood. It has been suggested that SuhB promotes antitermination in concert with NusA, given that some data show that they can interact together in solution (116). Perhaps this intermolecular interaction helps to reposition NusA away from RNA exit channel in a manner reminiscent of λN-dependent antitermination. Yet, the specific details of how this interaction between SuhB and NusA antagonizes Rho recruitment or suppresses Rho termination still remain elusive.

Antitermination by *cis*-acting regulatory RNA elements

In the aforementioned processive antitermination mechanisms, assembly of antitermination complexes largely depends on the incorporation of auxiliary transcription factors (*e.g.*, λN, λQ, SuhB) to adapt RNAP to a termination-resistant state. However, some examples of antitermination mechanisms have been discovered to be driven primarily by RNA elements that may or may not require protein

cofactors. For example, the lambdoid phage HK022 encodes for λQ protein, which is involved in late phage gene expression, but it lacks λN. Yet during early phage gene expression, the HK022 transcription apparatus still has to proceed past several premature termination signals, suggesting that antitermination is still required. Instead of λN antitermination, the early-expressed HK022 phage genes are regulated by a phage-encoded RNA element called *put* (120, 121), which is located in the leader region. Remarkably, it is capable of promoting antitermination without need for additional phage-encoded proteins.

In contrast to *boxA* and *boxB* RNA, the *put* RNA element is larger at approximately 70 nucleotides in length and consists of two hairpins that are separated by a single base (122). Mutational analysis of *put* RNA indicated that both hairpins are required for antitermination, as deletion of either hairpin significantly reduced readthrough efficiency. Furthermore, variations in stem loop lengths and some conserved residues are tolerable as long as these mutations are not disruptive to *put* RNA secondary structure nor to the orientation of hairpins relative to one another. These findings support the prediction that the secondary structure of *put* RNA is an integral component in its antitermination activity (122). Although mechanistic details of *put*-mediated antitermination are still lacking, current evidence suggests that *put* RNA interacts directly with β' subunit of RNAP (123) to somehow promote antitermination of Rho-dependent and intrinsic terminators.

Another example of RNA-mediated antitermination was found previously by our laboratory and is widespread amongst Bacillales (124). This antitermination system differs in that it involves a larger and more structurally complex RNA element

that is almost always associated with bacterial genes encoding for biosynthesis of biofilm or capsule exopolysaccharides. This regulatory RNA element was first discovered in *B. subtilis* (124), where it is located between the second and third genes in an operon encoding for biofilm exopolysaccharides (*eps*). Because of this, it was named EAR for *eps*-associated RNA. Comparative sequence analysis of EAR sequences identified ~125 nucleotides in the intergenic region that are functionally required for expression of downstream genes, which are otherwise silenced in the absence of EAR RNA. Furthermore, these comparative sequence analyses indicated that EAR is comprised of five conserved helical segments with a characteristic pseudoknot at its 5' terminus; however, no high-resolution structural data is yet available.

Intriguingly, the EAR element promotes processive antitermination of intrinsic terminators located thousands of nucleotides downstream of EAR. The EAR antitermination element is also likely to be a modular element, as it promotes readthrough of heterologous terminators that originate from unrelated sources. However, EAR-mediated antitermination activity is yet to be recapitulated *in vitro*. Because of this, it is possible that the EAR antitermination complex requires additional yet-to-be-identified factors. Although the EAR antitermination mechanism has not yet been uncovered, its discovery suggests that processive antitermination mechanisms may exhibit broader mechanistic diversity than previously appreciated, and that some of them may involve RNA elements that resemble the size and complexity of riboswitches. It also underscores that processive antitermination mechanisms could target gene expression of major cellular functions such as biofilm

formation. We speculate that more antitermination mechanisms, especially those which are mediated by structurally complex RNA elements, still await discovery.

Antitermination by NusG paralog RfaH

NusG/Spt5 proteins constitute a widespread family of transcription factors that affect transcription elongation for organisms from all three domains of life (reviewed in refs (36, 81, 125)). Strikingly, this family of transcription regulators exhibits the same degree of conservation as seen in RNAP subunits (126), suggesting that they carry out an ancient essential function in transcription. NusG is a core protein for the bacterial RNA polymerase machinery. It associates generally with transcription elongation complexes after sigma factors are ejected during the transition from initiation to elongation complexes. NusG is present within most transcription elongation complexes as they synthesize RNAs across the genome, although it can promote different regulatory effects. In general, NusG enhances transcription processivity (76), stimulates transcription-translation coupling (78), and aids in Rho-dependent termination (52), which helps NusG to silence foreign DNA. The eukaryotic NusG homolog Spt5 similarly enhances transcription processivity, but it also associates with factors mediating pre-mRNA processing (127), histone modification (128), and somatic hypermutation (129).

While virtually all bacteria encode for a core NusG protein, many bacteria also encode for one or more additional NusG paralogs. Phylogenetic analyses revealed that there are several distinct sub-classes of NusG paralogs (130, 131). These sub-families of NusG-like proteins exhibit specific but differing conservation patterns across bacterial phyla. It is presumed that they exert regulatory control over distinct

sets of genes. Yet, despite the importance of the genes they regulate, most of the subfamilies of NusG paralogs (*e.g.,* UpxY (34), TaA (132), ActX (133) and LoaP (131)) have not been investigated in depth. To this day, only one of these groups of NusG paralogs (RfaH) has been mechanistically characterized.

For many gammaproteobacteria, RfaH is thought to outcompete NusG for occupancy on the TEC as it transcribes a specific set of operons (134, 135). The targeted operons are important for expression of genes encoding for key virulence factors (136, 137), conjugation (138), and cell wall biosynthesis (139), which are otherwise silenced by Rho termination. It is estimated that core NusG outnumbers RfaH 100:1 within bacterial cells (85), yet RfaH displaces RNAP-associated NusG at its target operons to activate their gene expression. Interestingly, a characteristic DNA sequence (called the '*ops*' element), located in 5' UTR of target operons (140, 141) was discovered to act as an RfaH recruitment signal. RfaH forms specific contacts with flipped out bases in the non-template *ops* DNA strand in the transcription bubble (77), significantly enhancing its binding affinity to the elongation complex. It is proposed that RfaH binds the elongation complex at *ops* sites with ~1000-fold times higher affinity than core RNAP (134) thereby compensating for its low cellular abundance. Therefore, specific interactions with *ops* DNA ensures recruitment of RfaH to its target operons while also preventing RfaH from interfering with the global regulatory functions of core NusG.

Once recruited to the elongation complex, RfaH suppresses Rho-dependent termination (135) and stimulates coupling of RNAP to the leading ribosome (134). This is because, unlike NusG, RfaH is incapable of associating with Rho (76);

therefore, when NusG is replaced by RfaH, the TEC is rendered resistant to Rho termination (142). In addition to activating the expression of its target operons by inhibiting recruitment of Rho, RfaH exhibits an improved interaction with NusE (S10) (143). The enhanced binding of the RfaH to NusE (S10) promotes more efficient coupling between RNAP and the leading ribosome. This in turn enhances translation of open reading frames that contain alternate start codons or weak ribosome binding sites, features that are known to decorate the operons comprising the RfaH regulon (143). This improved coupling of transcription-translation apparatuses may also reduce the overall efficiency of intrinsic termination as the leading ribosome would sterically hinder the formation of terminator hairpins emerging from RNAP. These RfaH-mediated mechanisms together promote uninterrupted RNA synthesis of the targeted mRNAs while enhancing translation of their associated genes.

The discovery of RfaH and the elucidation of its regulatory mechanism have significantly expanded the overall expectations on the potential cellular roles of the NusG family of proteins. Therefore, further investigations into other sub-classes of NusG paralogs are crucial in order to explore their mechanistic diversity and to ascertain the extent of elongation-based regulatory control.

Antitermination by NusG paralog LoaP

Another subfamily of NusG specialized paralogs was previously discovered in our lab (131). Briefly, an extensive bioinformatic search was initially performed on all sequenced bacterial genomes in the NCBI genomes database to extract all available NusG sequences. NusG N-terminal NGN domain (NTD) associates with the

elongation complex at a conserved binding site on RNAP; therefore, this domain was used as a search model to broadly identify NusG homologs which are similar, but not identical, to core NusG. This is based on the prediction that mechanistic diversity within NusG family of proteins is more likely to emerge from binding interactions at the flexibly linked C-terminal domain (CTD), and less likely through NGN domain – as observed in RfaH subfamily.

We were interested to investigate whether specialized NusG paralogs could potentially exert regulatory control over biosynthetic gene clusters in bacteria. For this reason, identified NusG homologs were then filtered to only include sequences which are located within close proximity to biosynthetic operons encoding for secondary metabolites, particularly those that encode for production of several polyketides, non-ribosomal lipopeptides, and bacteriocins. This is because these operons are typically exceptionally long, and they range between a few to approximately a hundred kilobases in length (131). This observation hints at potential uncharacterized transcription-elongation based regulatory mechanisms which ensure uninterrupted and complete synthesis of exceptionally long mRNA. Therefore, we speculated that, perhaps, similar to long operons encoding for phage genes in bacteria, biosynthetic gene clusters are subjected to hitherto uncharacterized processive antitermination mechanisms.

This bioinformatic search was further restricted to bacterial organisms which encode for at least two NusG homologs to improve the likelihood of identifying specialized NusG paralogs associated with biosynthetic gene clusters and not core NusG. This bioinformatic analysis successfully identified a cohesive outgroup of

NusG paralogs to which the name LoaP was given for long operon associated protein (131). The LoaP subfamily of NusG proteins is widespread in Gram-positive bacteria, and it is particularly conserved in Actinobacteria, Firmicutes and some Spirochaetes. In these organisms the *loaP* gene is oftentimes found adjacent to biosynthetic gene clusters encoding for secondary metabolites suggesting a regulatory relationship with these pathways. Indeed, deletion of the *loaP* gene in *Bacillus velezensis* dramatically reduced the transcription abundance of two different polyketide synthase (PKS) gene clusters, which encode for two antibiotics: difficidin (*dfn*) and macrolactin (*mln*). Both compounds are synthesized from polyketide synthase pathways (PKS), and have been shown to be efficient antimicrobial agents against many bacterial pathogens (144-147). *B. velezensis* LoaP was shown to act as an antiterminator, as it promotes readthrough of intrinsic terminator sites located within the targeted operons as well as intrinsic terminators which were obtained from unrelated sources.

In an unrelated study (148), investigators investigated the utility of NusG factors in affecting expression of secondary metabolites in Gram-positive bacteria. They showed that overexpression of a particular NusG homolog in *Ruminoclostridium cellulolyticum* led to the discovery of an unprecedent class of natural products, to which the name closthioamide was given. Although this study initially intended to investigate whether housekeeping NusG could activate secondary metabolite gene clusters through its global regulatory effects on transcription, we were surprised to discover that the gene chosen to overexpress NusG in fact encodes for LoaP. Moreover, we also discovered that *R. cellulolyticum* encodes for two additional LoaP homologs located elsewhere in the chromosome, which appear to be

distinct from the homolog associated with the discovery of closthioamide. This suggests that *R. cellulolyticum* employs multiple LoaP regulons, which are likely to utilize discrete molecular mechanisms to control the correct recruitment of LoaP factors.

Surprisingly, closthioamide (CTA) is the only thioamide-containing non-ribosomal peptide (NRP) identified to date (149) exhibiting a broad-range antimicrobial activity against Gram-positive bacteria, and it is particularly potent against methicillin-resistant *Staphylococcus aureus* (MRSA) and vancomycin-resistant *Enterococcus faecium* (VRE) strains (148). These findings suggest that NusG paralog LoaP acts as a specialized transcription regulator that targets gene expression of a set of biosynthetic gene clusters, and thus exhibits functional roles that are distinct from housekeeping NusG. However, the mechanism that recruits LoaP to its target operons is unknown, as is the rest of the molecular mechanism of LoaP antitermination.

Therefore, it is essential to investigate the LoaP regulatory mechanism as it will give new insight into regulatory mechanisms that oversee production of biomedically relevant natural products (fig. 1-3) and will reveal new aspects of the mechanisms of transcription elongation. In this project, we purify LoaP proteins and examine some of its molecular interactions that are critical for antitermination. We also establish conditions for X-ray crystallographic analysis of LoaP complexes.

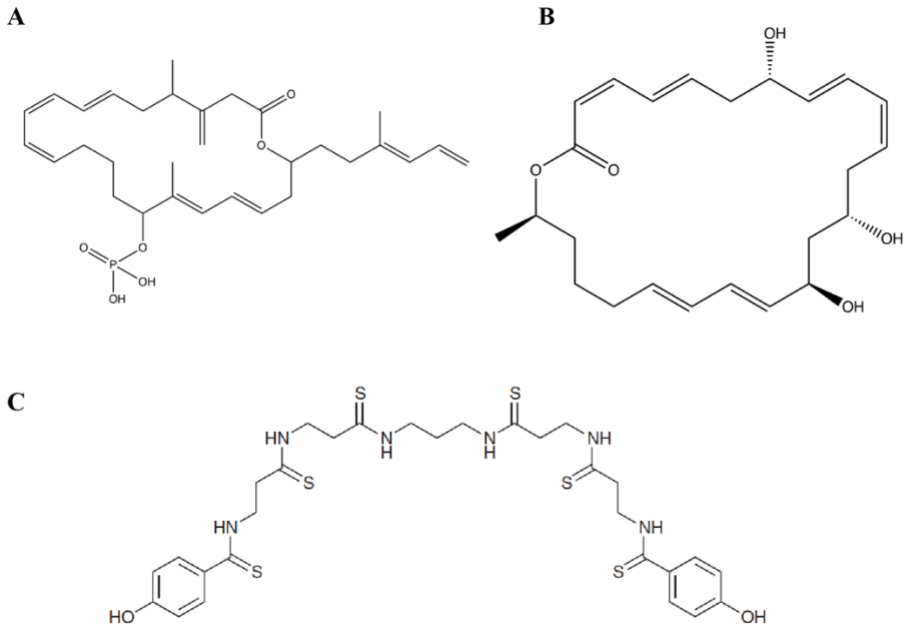


Figure 1-3. A structural representation of secondary metabolites which are synthesized from biosynthetic gene clusters regulated by LoaP. (A) Difficidin (*dfn*) (B) Macrolactin (*mln*) (C) Closthioamide (*cta*).

Chapter 2: Development of a purification protocol for NusG specialized paralog LoaP

Introduction

A phylogenetic analysis of the NusG family of transcription factors identified a distinct subfamily of NusG paralogs that was named LoaP for long operon associated protein (131). In this project, we began to explore LoaP's molecular features, including its macromolecular interactions and its structure. We reasoned that the identification of unique binding partners to LoaP proteins along with a comparative structural analysis relative to core NusG could reveal vital aspects of its antitermination mechanism. Crucial to this goal is the preparation of highly pure LoaP proteins.

The only example of a well-characterized specialized NusG paralog in bacteria is RfaH. In *E. coli*, RfaH associates with paused elongation complex (PEC) at specific DNA sequences called operon polarity suppressor (*ops*). Specific interactions with the elongation complex at *ops* sites are mediated through conserved amino acid residues in RfaH which are not present in core NusG from the same bacterium (150, 151). Intriguingly, the Cryo-EM structure of RfaH-associated TEC (77) revealed unique interactions between conserved residues in RfaH and solvent exposed non-template DNA strand encoding *ops* sequence. The latter forms a hairpin-like structure in the transcription bubble which acts as a recruitment signal for RfaH but is unrecognized by core NusG.

Housekeeping NusG proteins routinely associate with RNAP during transcription elongation to modulate its function. Binding of NusG proteins to elongation complexes affects the rate of RNA synthesis, transcription-translation coupling (152), and transcription termination at Rho-termination sites (86, 106). While the association of housekeeping NusG with elongation complexes is thought to occur non-specifically across most genes, NusG paralogs are typically operon-specific and therefore their association with the elongation complex occurs via distinct recruitment signals at target operons (141, 153). Also, it is possible that nucleic acid determinants play a role in the antitermination mechanisms used by NusG paralogs, perhaps similar to how nucleic acid components assist lambda and *rrn* antitermination complexes. Therefore, we hypothesized that nucleic acid determinants may be important for LoaP proteins.

We speculated that operons that are directly regulated by LoaP may encode recruitment signals that are analogous to *ops* DNA and/or feature nucleic acid elements that are required for antitermination. Upon searching for conserved features, our bioinformatic analysis of LoaP-associated gene clusters revealed a short DNA sequence (~26 nts) that appeared to be present in the 5' leader regions of multiple LoaP-associated gene clusters (Chapter 3). Therefore, we set out to determine if purified LoaP proteins might exhibit measurable binding affinity to this conserved nucleic acid element *in vitro*. However, our first objective was to overexpress and purify LoaP proteins that could be used as representatives of the LoaP subfamily at large.

My initial attempts at overexpressing LoaP, however, often led to low expression levels and poor purity which were determined to be unsuitable for further *in vitro* biochemical assays (fig. 2-1). To circumvent this problem, we subcloned *B. velezensis* LoaP into a plasmid encoding for a N-terminal hexahistidine-maltose binding protein (MBP) tag in order to improve total protein yield. Though the method proved successful with respect to protein yield, it did not lead to purification of untagged protein, as free LoaP routinely associated with the solubility tag following tag cleavage. Furthermore, small quantities of untagged LoaP proteins that had been purified using this method precipitated in dialysis cassettes or in storage within a few days, despite extensive efforts at improving its solubility.

It was preferred to obtain untagged LoaP in high yield and purity suitable for biochemical and structural studies; therefore, a new purification protocol was needed to be developed. In this chapter, we describe the development of a purification strategy which allowed for purification of untagged LoaP proteins in high yield and purity. Moreover, we elucidate key variations in the amino acid sequences of LoaP proteins that shed some light on their observed insolubility and biochemical interactions. The successful purification of LoaP proteins was a fundamental requisite for the research goals of this dissertation – without pure and soluble protein, the completion of this dissertation would not have been attainable.

Results

High salt concentration in the lysis buffer improves recovery of LoaP proteins

A comparative analysis was performed using Expasy ProtParam tool to evaluate the hydrophobicity index of LoaP sequences relative to their NusG

homologs. While there were no significant differences in the theoretical hydrophobicity values between homologs from the same bacterium, our analysis revealed that LoaP proteins exhibit a sequence-attributed property that appears to be distinct from NusG and RfaH proteins. LoaP proteins are broadly characterized by a preponderance of positively charged residues (fig. 2-2 A). Intriguingly, several amino acid residues that are conserved in NusG and RfaH as negatively charged residues (Glu and Asp) are similarity conserved but oppositely charged in LoaP subfamily. This phenomenon of charge swapping lends LoaP proteins an overall net positive charge at physiological pH. Indeed, the theoretical pI of LoaP from *B. velezensis* (pI ~ 9.7) is much higher than that of core NusG (pI ~ 5.3) from the same bacterium (fig. 2-2) reflecting differences in conservation patterns of charged amino acid residues.

We speculated that the recovery of highly charged LoaP proteins could be improved by increasing the ionic strength of the lysis buffers. To test this hypothesis, cells from IPTG-induced cultures expressing LoaP from *B. velezensis* (UniProtKB: A0A411A7S9) were harvested and suspended in 2x lysis buffer (50 mM NaH_2PO_4 pH 7.2, 2 mM EDTA, 2 mM DTT, 10% glycerol). Lysis buffers were then supplemented with increasing concentrations of sodium chloride to increase the ionic strength incrementally. Resuspended cell pellets were lysed using 0.5 mg/mL lysozyme in 100, 300, and 1000 mM NaCl. Following this treatment, lysates were clarified by centrifugation and analyzed on SDS-PAGE gels. The results indicated that higher salt concentration slightly increased the recovered amount of overexpressed protein in the soluble fraction (fig 2-3A), while the amount of overexpressed protein in the insoluble pellet decreased. Furthermore, lysis buffers

A **B**

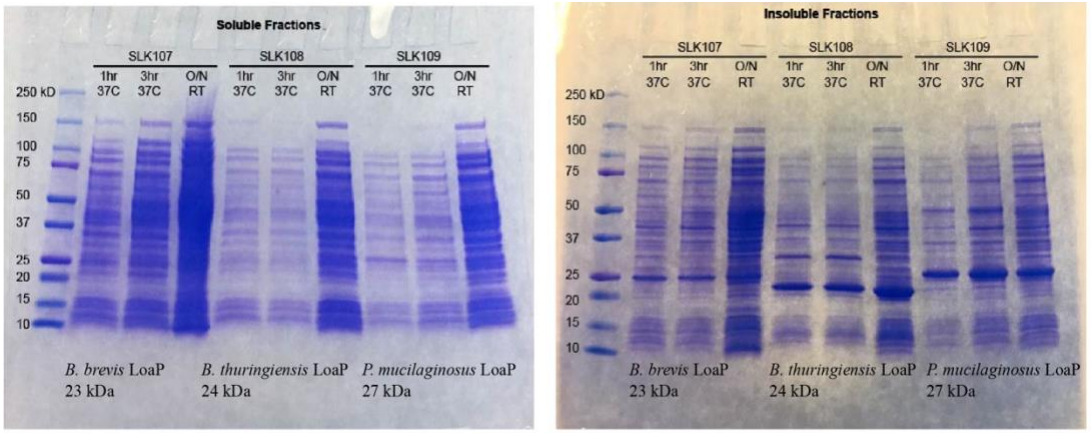


Figure 2-1: Overexpressed LoaP proteins aggregate in the insoluble pellet. Analysis of cell lysates on 4-20% SDS-PAGE gels indicated that overexpressed LoaP homologs from *Bacillus brevis*, *Bacillus thuringiensis*, and *Paenibacillus mucilaginous* are not detected in the soluble fraction (A) and that they aggregate in the insoluble pellet (B). Each strain containing an inducible copy of LoaP homolog was grown at 37°C until OD_{600}nm= 0.6-0.7, at which point protein overexpression was induced by the addition of 1 mM IPTG. Cells were harvested at different points via centrifugation and lysed upon incubation with 0.2 mg/mL lysozyme in 10 mM Tris-HCl pH 7.5, 200 mM NaCl, 0.1 mM EDTA, 1 mM DTT. Cell lysates were centrifuged at 12,000xg for 30 minutes to separate the soluble fraction from the insoluble cellular debris.

containing high salt concentration (1000 mM NaCl) show a better enrichment of overexpressed protein relative to the total protein content in cell lysates. This data together indicates that high salt concentration enhances the recovery and purity of LoaP proteins during cell lysis.

Polyethyleneimine precipitation removes bound nucleic acids

While protein recovery in cell lysates was improved by increasing salt concentration, purified LoaP often precipitated overnight during dialysis in salt concentrations below 300 mM. We attributed the precipitation of purified protein to reduction of salt concentration and we reasoned that the solubility of LoaP proteins might depend on the ionic strength of the buffer. Moreover, we observed that LoaP proteins routinely eluted from Ni-NTA columns bound to contaminating nucleic acids. This was indicated by the measured $A_{260/280nm}$ value, which was approximately 1.2. For this reason, we opted to remove contaminating nucleic acids prior to any further affinity purification steps by titrating polyethyleneimine (PEI) directly into cell lysates.

To determine the amount of PEI required to preferentially remove bound nucleic acids from cell lysates, we titrated PEI dropwise, from a 10% stock at pH 7.2, directly into lysates to 0.1, 0.2, 0.4, 0.6% (v/v) final concentrations. Lysis buffers in this experiment contained 1 M NaCl – at this concentration, PEI preferentially binds to nucleic acids leaving the majority of cellular proteins in the supernatant. The addition of PEI directly to cell lysates formed dense white precipitate, which is typically observed when nucleic acids become insoluble. We observed that precipitation occurred at all PEI concentrations tested in this experiment, therefore lysates from each

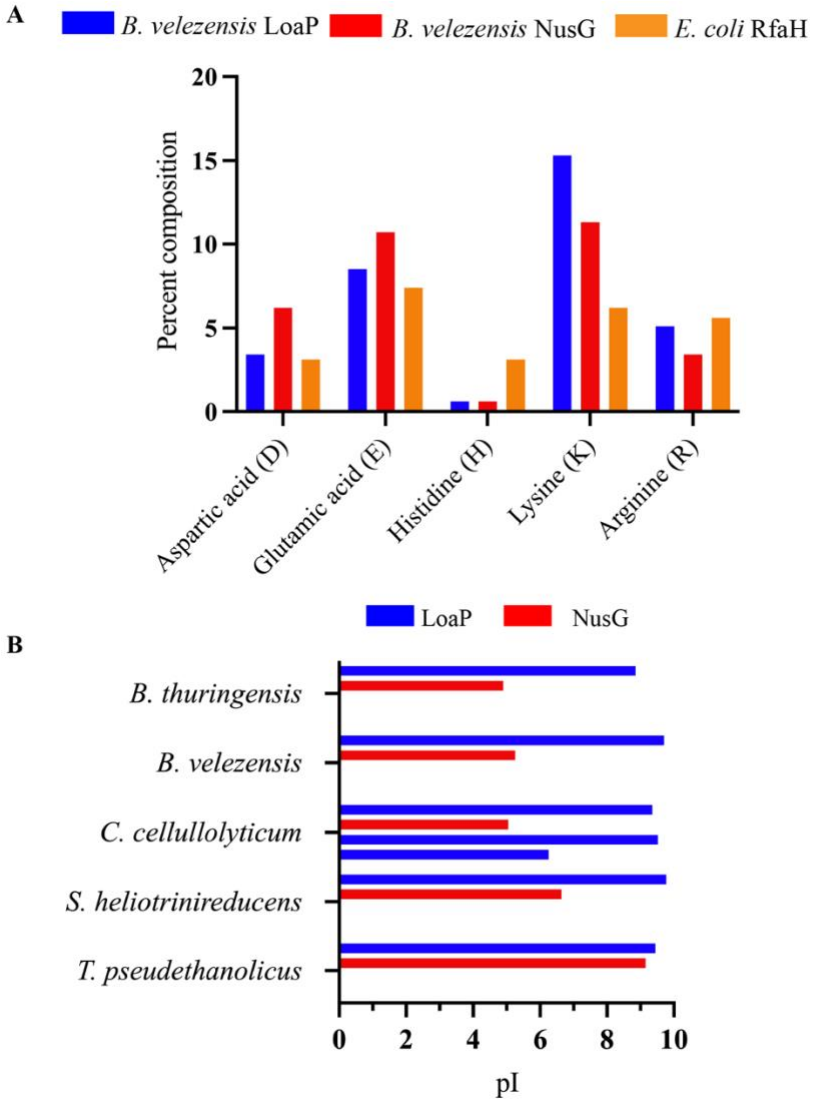


Figure 2-2: LoaP proteins are enriched in positively-charged amino acid residues relative to housekeeping NusG and *E. coli* NusG paralog RfaH. (A) Bar chart showing the distribution of charged amino acids in *B. velezensis* LoaP, *B. velezensis* NusG, and *E. coli* RfaH. Amino acid distribution frequency was calculated using Expassy ProtParam tool and plotted on y-axis as percent composition. (B) Relative to core NusG proteins from the same bacterium, NusG paralog LoaP is often enriched in positively-charged residues in several bacterial species resulting in higher pI values. Note: *Clostridium cellulolyticum* encodes for three LoaP homologs in addition to core NusG. Also, *C. celluloyticum* was recently renamed *Ruminoclostridium cellulolyticum.*

titration were analyzed by SDS-PAGE to determine their protein composition. As expected, addition of PEI at 1 M NaCl concentration had no measurable effect on total protein yield in the soluble fraction (fig 2-3B). Therefore, we decided to titrate PEI to 0.6% (v/v) final concentration to all lysates to ensure the complete removal of nucleic acids.

Ammonium sulfate precipitation removes residual polyethyleneimine

To determine the amount of ammonium sulfate required to precipitate LoaP, small aliquots of PEI-treated cell lysates were incubated with ammonium sulfate powder in 1 mL Eppendorf tubes which were then incubated in an ice bath for 30 minutes. The amount of ammonium sulfate powder added to each 1 mL lysate was determined based on its density value at 0°C (154). The concentration of ammonium sulfate at which proteins precipitate is typically determined empirically for every protein of interest. Therefore, ammonium sulfate powder was added to PEI-treated lysates to final concentrations of 25, 30, 35, 40, 45, 50% (w/v). Precipitated material from each ammonium sulfate concentration was then harvested by centrifugation, resuspended in lysis buffer, and analyzed by SDS-PAGE. The results indicated that the majority of soluble LoaP precipitated in 45-50% AS concentration range (fig 2-3C).

Ammonium sulfate pellets containing precipitated protein were then resuspended in lysis buffer and dialyzed twice against 1 L buffer containing: 20 mM NaH_2PO_4 pH 7.2, 50 mM NaCl, 1 mM EDTA, 1 mM DTT, 5 % glycerol. To verify the presence of LoaP in soluble fractions, small aliquots from the dialyzed fractions as well as the precipitated material (dissolved in 8M urea) were analyzed again by

SDS-PAGE (fig 2-5). These results indicated that purified LoaP remained soluble throughout multiple dialysis steps in low salt buffers (~100 mM NaCl); therefore, we concluded that the removal of contaminating nucleic acids proved essential in obtaining soluble protein.

Untagged LoaP is separated and purified via cation-exchange chromatography

Previously, our attempts to remove MBP solubility tags fused to LoaP proteins were largely unsuccessful – for reasons discussed earlier. To circumvent this issue, we replaced the 44-kDa MBP tag with a smaller solubility tag (~9 kDa bdSENP SUMO, (155)) fused to the N-terminus in frame with a decahistidine tag, and tested the efficiency of tag cleavage at 4°C in 60 minutes, and overnight.

LoaP proteins were purified by treating cell lysates with 0.6% PEI (v/v) followed by the addition of 50% ammonium sulfate powder (w/v). Resuspended ammonium sulfate pellets containing overexpressed LoaP proteins were applied onto a Ni-NTA column to purify tagged-LoaP, then the eluate was dialyzed twice in 1 L cleavage buffer (10 mM NaH_2PO_4 pH 7.2, 50 mM NaCl, 0.1 mM EDTA, 2 mM $MgCl_2$, 1 mM DTT, 5% glycerol) for two hours, and then again overnight.

Cleavage reactions were set up as follows: 50 μM tagged-LoaP in 1x cleavage buffer (50 μL total volume), 0.5 μM purified bdSENP SUMO protease. At indicated time points, 10 μL aliquots were mixed with 1x SDS-loading buffer, heated at 95°C for 3 minutes, and then analyzed by SDS-PAGE. The results indicated that SUMO protease cleaved the solubility tag with high efficiency within 60 minutes at 4°C (fig 2-5).

To separate the cleavage products, cleavage reactions were applied onto 1-mL

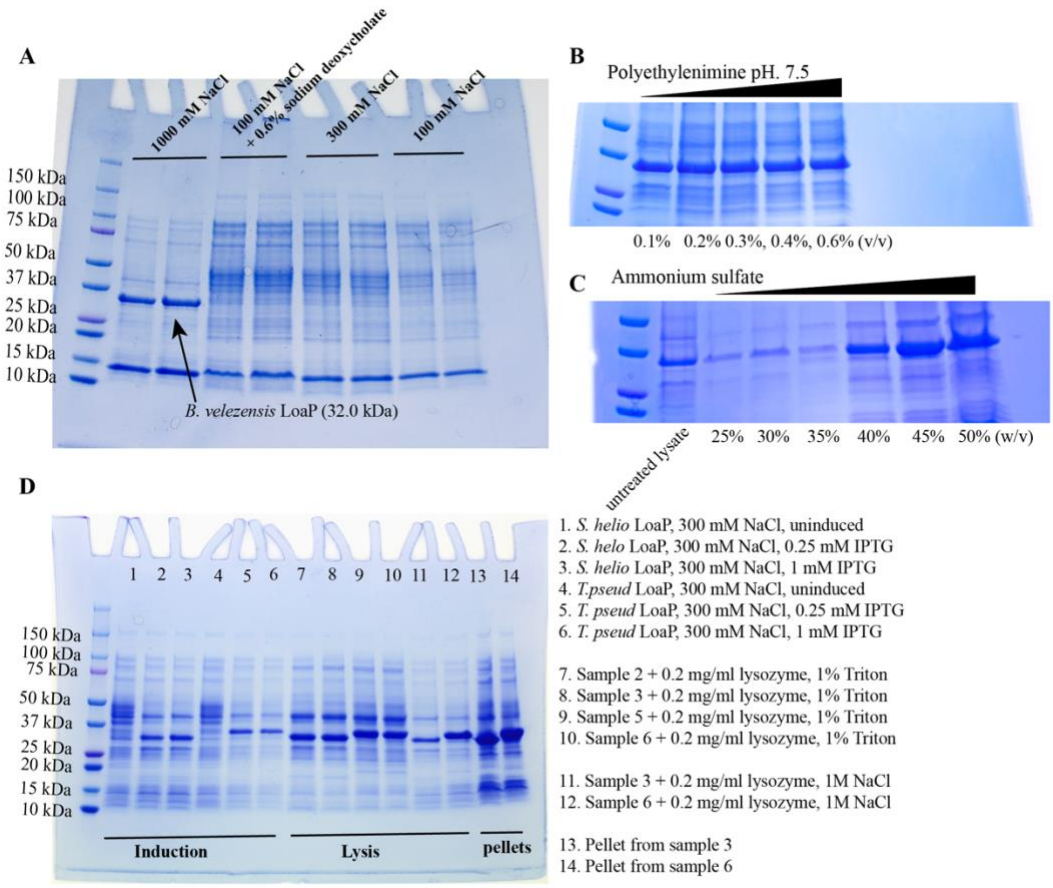


Figure 2-3: The yield of overexpressed LoaP proteins is significantly enhanced by high concentration of sodium chloride in the lysis buffer. Cell lysates were analyzed on 4-20% SDS-PAGE gels to assess protein composition. (A) *B. velezensis* LoaP homolog is poorly soluble in low-mid salt concentrations (100-300 mM) but becomes soluble at 1000 mM NaCl. Cell lysis in this experiment was carried out in duplicates and 0.6% sodium deoxycholate was used to test if the solubility of overexpressed LoaP is affected by detergents. (B) Addition of polyethyleneimine (PEI) to cell lysates containing 1000 mM NaCl preferentially precipitates nucleic acids leaving almost all cellular proteins in the lysates within the titration range used (0.1-0.6% (v/v)). (C) Small-scale ammonium sulfate precipitation test showing the composition of precipitated proteins at 25, 30, 35, 40, 45, and 50% ammonium sulfate concentrations (w/v) added to PEI-treated cell lysates. In these tests, ammonium sulfate was added in powder form to PEI-treated cell lysates, then incubated on ice for 30 minutes. (D) 4-20% SDS-PAGE gel showing the effects of high salt concentration and 1% triton on the solubility of two overexpressed LoaP homologs obtained from *S. heliotrinireducens* and *T. pseudethanolicus*, respectively. Cells were lysed using 0.2 mg/mL lysozyme at room temperature in (A-C) and via sonication at 4°C in (D)

HiTrap SP-HP cation exchange column (GE healthcare) at 0.25 mL/min flowrate. We selected a cation-exchange chromatography technique due to the distinct positively-charged character of LoaP proteins. After the cleavage reactions were loaded on HiTrap SP HP column, the column was washed with 10 column volumes cleavage buffer, and then subjected to a linear salt gradient (50-1000 mM NaCl) over 60 minutes. Untagged LoaP protein preferentially bound SP resin, and eluted at a higher salt concentration (~ 600 mM NaCl, fig. 2-4A) compared to other proteins in the reaction. Based on this data, we concluded that SP-HP columns provide a feasible and scalable method to separate untagged LoaP after cleavage of the SUMO-tag. This method also allowed for purification of untagged LoaP to a higher level of purity, which was previously unattainable.

Gel filtration chromatography yields pure untagged LoaP

Elution fractions containing pure untagged LoaP were pooled, concentrated, and applied onto Superdex 75 Increase 10/300 GL (Cytvia) column to exchange into storage buffer (10 mM NaH_2PO_4 pH 7.2, 100 mM NaCl, 0.1 mM EDTA, 1 mM DTT, 10% glycerol). Untagged LoaP protein eluted from the size-exclusion column as a single peak at a nearly identical time compared to purified core NusG protein under the same experimental conditions (fig. 2-4B). Finally, fractions containing untagged LoaP proteins were concentrated to ~ 12 mg/mL, flash frozen in liquid nitrogen, and stored at -80°C.

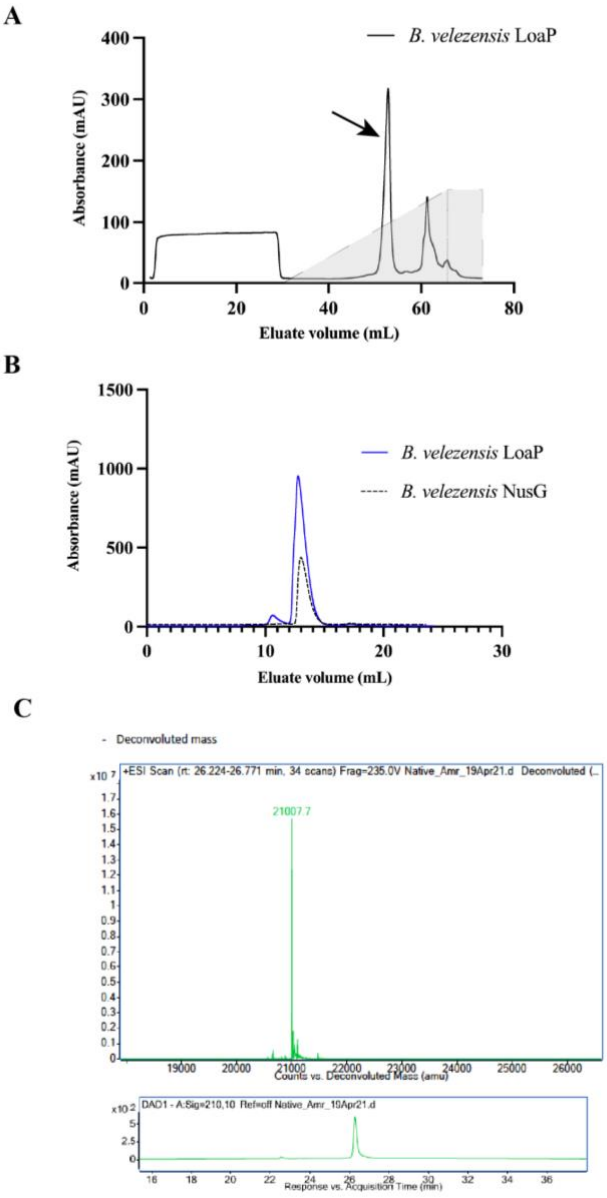


Figure 2-4: Affinity chromatography purification of *B. velezensis* LoaP. (A) FPLC chromatogram of untagged LoaP obtained by purification on SP Sepharose cation exchange column. Untagged LoaP protein was eluted with a linear salt gradient: 100-1000 mM sodium chloride over 30 minutes (Grey shaded area). Black arrow points at untagged LoaP peak eluting around 450 mM NaCl (25 mS/cm). (B) Size-exclusion chromatography of untagged LoaP relative to NusG standard. (C) *Top*: deconvoluted mass spectrum showing experimentally-determined molecular wight (21007.7 Da). *Bottom:* TOF LC/MS analysis of protein stocks using 0-70 % Acetonitrile linear gradient at 2% per minute flow rate. Mass spectrometry analysis was performed by Dr. Duck-Yeon Lee at National Heart Lung Blood Institute (NHLBI/NIH), Biochemistry core.

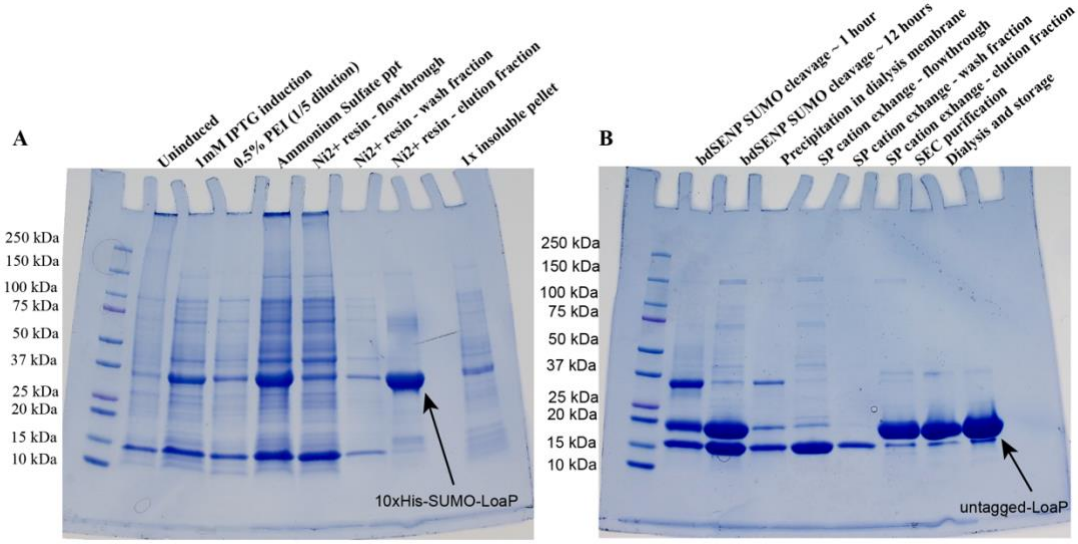


Figure 2-5: Tracking protein purification steps in the newly developed purification protocol. (A) 4-20% SDS-PAGE gel showing purification of 10xHIS-SUMO tagged *B. velezensis* LoaP. (B) 4-20% SDS-PAGE gel showing cleavage and purification of untagged *B. velezensis* LoaP.

Discussion

Recently, NusG specialized paralog LoaP was identified in a few bacterial organisms. In *Bacillus velezensis*, LoaP specifically targets gene expression of two independent biosynthetic clusters encoding for two antibiotics. This perhaps suggests a potential regulatory mechanism not yet identified within the NusG family of transcription regulators. Inspired by these findings, we set out to investigate whether LoaP antiterminators are recruited to their target operons by binding to conserved nucleic acid sequences contained within the leader regions. Furthermore, we aimed to solve the crystal structure of LoaP to gain deeper insight into the LoaP subfamily.

Our experimental strategy relied entirely on obtaining LoaP proteins with sufficient purify suitable for biochemical and biophysical experiments. However, overexpression and purification of LoaP proteins presented major hurdles in the quest to obtain purified LoaP. In order to proceed with the aims of this dissertation research, we decided to develop a new purification strategy aimed at optimizing protein yield as well as achieving a high level of purity. No biochemical data was available in the literature at the time outlining purification strategies for LoaP proteins. In contrast, core NusG from *E. coli* and *B. subtilis*, in addition to NusG paralog RfaH, have been routinely overexpressed and purified using standard purification protocols.

To gain a better insight into the physical properties of LoaP proteins, we performed a comparative sequence analysis on LoaP sequences against core NusG from the same bacterium. This comparative analysis revealed an intriguing physical property that appears to be diagnostic of LoaP proteins. Unlike the well-characterized

NusG proteins, LoaP proteins are enriched in basic amino acid residues (Lys and Arg) resulting in a significant increase in the theoretical pI values (fig. 2-2). As such, LoaP proteins bear a positively-charged character at physiological pH, in contrast to core NusG and RfaH – both of which are mildly acidic proteins.

We speculated that the accumulation of overexpressed protein in the insoluble pellet (fig. 2-1 B) could arise from non-specific ionic interactions with the negatively charged cell membrane and/or *E. coli* host-encoded proteins. We tested this hypothesis by incrementally increasing the ionic strength of the lysis buffers to disrupt any non-specific ionic interactions thereby releasing membrane-associated LoaP proteins. As expected, the addition of 1 M NaCl during cell lysis significantly increased the amount of overexpressed protein in the soluble fraction (fig. 2-3A). Analysis by SDS-PAGE indicated that lysis buffers with >300 mM sodium chloride concentration contained higher yield of overexpressed protein, and more importantly, reduced the amount of contaminating cellular proteins (fig 2-3A). Moreover, lysis buffers containing 1 M NaCl also reduced the amount of overexpressed protein in the insoluble pellet (fig 2-5) indicating that high salt concentration is indeed required for the solubilization of overexpressed protein.

Similar results were obtained when LoaP homologs from *Slackia heliotrinireducens* and *Thermoanaerobacter pseudethanolicus* were subcloned into the same parent vector and overexpressed under the same conditions. In these experiments, the addition of 1 M NaCl to the lysis buffer improved protein recovery – albeit to a lesser extent, and reduced the level of contaminating cellular proteins in the soluble fraction (fig. 2-3D).

Consequently, lysis buffers were supplemented with 1 M sodium chloride – to ensure optimal protein recovery – and cell lysates were clarified and incubated with pre-equilibrated Ni-NTA resin for one hour. His-tagged LoaP proteins were obtained by elution from Ni-NTA using stepwise imidazole gradient. Elution fractions were concentrated and dialyzed in a storage buffer containing 200 mM sodium chloride to test the stability of purified protein in low salt conditions. This purification method yielded a reasonable amount of purified LoaP protein, however, protein precipitation occurred during dialysis in storage buffers containing 200 mM NaCl.

When we measured the absorbance values of elution fractions at A_{280nm} to assess protein purity, we discovered – serendipitously – that in nearly all elution fractions, the $A_{260/280nm}$ ratio (~ 1.2) was higher than what is typically reported for pure proteins. It is generally accepted that $A_{260/280nm}$ value for pure proteins is around 0.6, while the value for pure DNA and RNA is 1.8 and 2.0 (156), respectively. This observation indicated that purified LoaP proteins co-elute with bound nucleic acids from Ni-NTA columns. This is perhaps consistent with the observed protein precipitation in dialysis cassettes as the negatively charged contaminating nucleic acids could potentially reduce LoaP solubility in aqueous solutions via charge neutralization. Therefore, we hypothesized that the removal of contaminating nucleic acids from protein samples could significantly improve the purity of the recovered LoaP proteins.

We searched the literature for biochemical techniques to effectively remove contaminating nucleic acids during the purification of proteins. It was preferred to avoid the addition of DNases and RNases as they are difficult to remove and often

interfere with downstream biochemical assays. Instead, we decided to try a precipitation technique commonly used in the removal of nucleic acids from cell lysates. This technique relies on the addition of a cationic polymer called polyethyleneimine (PEI) directly to cell lysates. Polyethyleneimine (PEI) was initially selected as it is a rapid and relatively inexpensive technique. It has been used traditionally in the removal of genomic DNA from cell lysates as well as purification of DNA-binding proteins, and RNA-binding viral proteins.

PEI is a cationic polymer which exhibits a stronger binding affinity toward nucleic acids than proteins. Binding to nucleic acids results in their immediate precipitation in aqueous solutions at salt concentrations below ~1.6 M sodium chloride (154), which can then be removed by centrifugation. We titrated PEI directly to cell lysates to 0.6% (v/v) final concentration to ensure complete removal of nucleic acids. The addition of PEI at this concentration preferentially precipitated nucleic acids and did not reduce the amount of overexpressed protein in the soluble fraction (fig. 2-3B).

Typically, it is necessary to remove residual PEI from the soluble fraction prior to purifying proteins by affinity chromatography. This is because PEI could disrupt binding interactions between proteins of interests and the chromatography column. PEI is a branched cationic polymer with molecular weights ranging from 30-90 kDa (157, 158). As such, complete removal of PEI molecules from protein solutions cannot be achieved by dialysis. We chose to use an ammonium sulfate based precipitation technique in lieu of dialysis to precipitate LoaP proteins from PEI-treated cell lysates. This method has two advantages: (1) selective precipitation of

LoaP from cell lysates improves protein purity (2) complete removal of soluble PEI which does not co-precipitate with proteins and can be discarded by decanting the supernatant.

Ammonium sulfate powder was added directly to PEI-treated cell lysates at increasing concentrations. Analysis of the precipitated material from each ammonium sulfate concentration by SDS-PAGE revealed that purified LoaP proteins precipitated at 45% ammonium sulfate concentration (w/v) (fig. 2-3C). Ammonium sulfate pellets were then gently resuspended in a low salt buffer solution (10 mM NaH2PO4 pH 7.2, 50 mM NaCl, 0.1 mM EDTA, 1 mM DTT, 5% glycerol). The resuspension was dialyzed in the same buffer overnight at 4°C to test protein stability. To our surprise, minimal protein precipitation was observed in the dialysis cassette as salt concentration was gradually reduced to ~50 mM NaCl. This observation was unexpected because LoaP proteins routinely precipitated during dialysis steps prior to treatment with PEI and ammonium sulfate. To assess nucleic acid contamination, $A_{260/280nm}$ values were measured after treatment of cell lysates with PEI and ammonium sulfate. UV-absorbance spectra revealed that $A_{260/280nm}$ values decreased from 1.2 to 0.75 indicating the removal of nucleic acids from cell lysates.

Following the removal of contaminating nucleic acids, LoaP proteins were purified on Ni-NTA column and then incubated with bdSENP SUMO protease for 60 minutes at 4°C to remove SUMO solubility tag. Cleavage products were then separated on SP HP cation exchange column at 0.25 mL/min rate. Untagged LoaP eluted from SP column with a linear salt gradient (50-1000 mM NaCl) over 60

minutes and was effectively separated from SUMO-tagged LoaP (fig. 2-4A and fig. 2-5B).

Fractions containing purified LoaP were concentrated and applied on Superdex 75 increase size-exclusion column to exchange into the storage buffer (10 mM NaH2PO4 pH 7.2, 100 mM NaCl, 0.1 mM EDTA, 1 mM DTT, 10% glycerol). Moreover, core NusG protein (20.1 kDa) was used as a size marker to estimate the size of purified LoaP protein. LoaP (21.0 kDa) eluted as a single peak and at approximately the same time as core NusG protein suggesting that purified LoaP exists in a monomeric form in solution. Finally, the purity of purified LoaP obtained from the herein described purification strategy was determined by electrospray ionization mass spectrometry (ESI-MS) to be approximately 97% pure (see fig. 2-4C). Protein yield was calculated to be approximately 6.5 mg purified protein per 1 L culture and the purification procedure lasted for 4 days (see fig. 2-6 for a detailed timeline of the purification procedure).

Based on these results, we concluded that the removal of contaminating nucleic acids from cell lysates improved protein purity and, more importantly, significantly enhanced protein recovery. It has not escaped our notice that these observations all together hinted at a capability of LoaP proteins to exhibit nucleic acid binding interactions warranting further biochemical investigation, which will be explored in the following chapter.

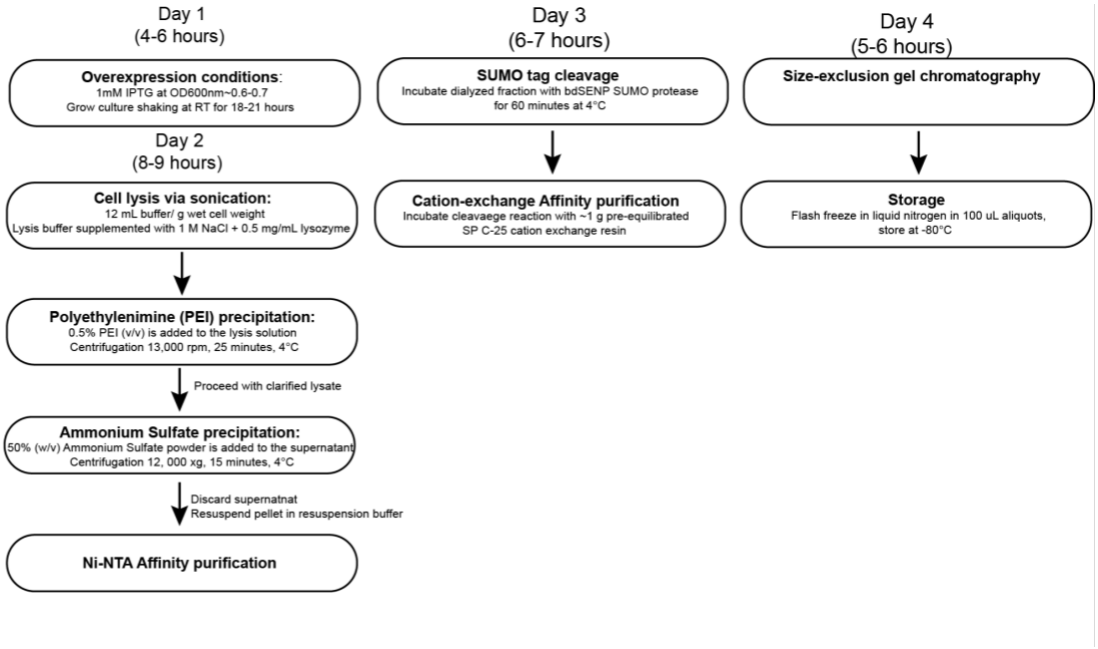


Figure 2-6: A schematic diagram showing the timeline of the purification protocol.

Materials and methods

Strain construction

LoaP DNA sequences were obtained from National Center for Biotechnology Information (NCBI) genomes database and purchased from Integrated DNA technologies (IDT) as gblocks. *LoaP* gblocks were subcloned into a plasmid (pAmr30) containing an IPTG-inducible N-terminal decahistidine tag fused to a bdSENP SUMO tag. Plasmids were assembled using Gibson assembly technique and transformed into both *E.coli* XL10-Gold (Agilent) for plasmid replication and *E.coli* T7 express (New England BioLabs) for protein overexpression. All plasmids were verified using Sanger sequencing.

LoaP Purification protocol

Cultures were grown in 2xYT media, shaking at 37°C, until reaching an OD_{600} of ~0.6. 1 mM isopropyl-β-*D*-thiogalactoside (IPTG) was added and cultures were incubated, with shaking, at room temperature for 16-18 hours. Cells were harvested by centrifugation, resuspended at 12 mL/g wet cell weight in lysis buffer (50 mM Na_2HPO_4 pH 7.2, 1 M NaCl, 10 mM EDTA, 1mM dithiothreitol (DTT), 0.5 mg/mL lysozyme, 1 mM PMSF, 5% glycerol) and incubated, with gentle rocking, at room temperature for 20 minutes and then on ice for an additional 10 minutes. The suspensions were sonicated 4 times for 20 seconds at 4°C, incubating in an ice bath for 2 minutes between each sonication cycle. Polyethyleneimine (PEI) was titrated dropwise with gentle stirring to a final concentration of 0.6 % (v/v). The lysate was then centrifuged at 13,000 x g at 4°C for 25 minutes to remove cell debris and precipitated nucleic acids. Clarified lysate were collected and centrifugated for

additional 15 minutes at 13,000 x g. Ammonium sulfate was added gradually in powder form, with stirring, to a final concentration of 50% (w/v) and the lysate was incubated on ice for 20 minutes. The lysate was centrifuged for 15 minutes at 10,000 x g and the pellet was resuspended in 20 mL resuspension buffer (50 mM Na_2HPO_4 pH 7.2, 100 mM NaCl, 0.1 mM EDTA, 5% glycerol). The supernatant was then filtered through 0.4µm filter to get rid of any precipitated material and incubated with cOmplete™ His-Tag purification resin (Sigma-Aldrich) at 4°C for one hour, with gentle rocking, before transferring to a gravity column, draining (<0.5 mL/min), and washing with 10 column volumes of resuspension buffer, and 10 column volumes of 25 mM imidazole in resuspension buffer. LoaP was eluted with 5 column volumes of 400 mM imidazole in resuspension buffer and dialyzed overnight in S-loading buffer (10 mM Na_2HPO_4 pH 7.2, 50 mM NaCl, 0.1 mM EDTA, 1mM DTT, 0.5% glycerol). Dialyzed fractions were transferred to a sterile tube and $MgCl_2$ and DTT were each added to a final concentration of 1 mM. This was then incubated with bdSENP1 protease on ice for 1 hour before being loaded onto a HiTrap™ SP HP column (GE Healthcare) at 0.25 mL/min flow rate. The column was then washed with 10 column volumes of S-loading buffer and processed LoaP was eluted from the column with a linear 50-1000 mM NaCl in S-loading buffer at 0.25 mL/min over 1 hour 4°C. Fractions containing cleaved LoaP were pooled and concentrated to 1-2 mL and then loaded onto a Superdex 75 10/300 GL size exclusion column which was pre-equilibrated in storage buffer (20 mM Na_2HPO_4, pH 7.2, 200 mM NaCl, 1mM DTT, 0.1 mM EDTA, 10% glycerol) at 0.35 mL/min at 4°C. Fractions containing LoaP

were concentrated to 12 mg/mL and flash frozen in liquid nitrogen. Samples were stored in 0.2 mL aliquots at -80°C or at -20°C and 50% glycerol after thawing.

Chapter 3: LoaP proteins exhibit high affinity, specific, RNA-binding activity

Copyright Notice

Introduction

NusG paralogs are broadly distributed in bacteria; however, most of the sub-classes of specialized NusG paralogs remain uncharacterized (131). Their regulatory targets remain unidentified, and their molecular mechanisms remain undiscovered. While the discoveries on RfaH may provide an important preview into the genetic regulatory mechanisms used by other NusG paralogs, several recent observations serve to temper these expectations and suggest there may be fundamental differences between NusG-like proteins. For example, while the primary role of RfaH is to improve transcription-translation coupling, a recent study argued that some bacteria may not even incorporate routine interactions between RNAP and the ribosome (159). This study revealed that in *Bacillus subtilis*, elongating RNAPs outpace the leading ribosome resulting in a 'runaway transcription' where RNAP remains uncoupled to the translation machinery and thus the rates of transcription and translation are different. This suggests that there may be fundamental differences between bacteria in the

regulatory mechanisms they employ for control of transcription elongation. In particular, it remains to be determined how NusG paralogs might exert an influence on the TEC for Gram-positive bacteria, such as *B. subtilis*.

The LoaP sub-family of NusG proteins is primarily encoded by Actinobacteria, Firmicutes and Spirochaetes (131). In these organisms, the gene encoding for LoaP is typically located near biosynthetic gene clusters of polysaccharides or secondary metabolites. *Bacillus velezensis* LoaP activates expression of two different antibiotic synthesis gene clusters. It acts as an antiterminator protein in this organism, promoting readthrough of intrinsic terminators located within the targeted operons. *B. subtilis* and *B. velezensis* are very closely related organisms (131, 160); therefore, analysis of LoaP antitermination is likely to demonstrate how NusG paralogs can regulate transcription elongation in bacteria that do not couple transcription and translation. In particular, the absence of a coupled ribosome during transcription elongation raises the possibility that Rho terminator could be dispensable in some termination/antitermination complexes. Instead, transcription attenuation in these organisms is achieved through interactions with nascent RNA and specific RNA-binding regulators.

Herein, we report that *B. velezensis* LoaP exhibits the unexpected ability to bind an RNA hairpin with high affinity and high selectivity. This RNA hairpin can be found in LoaP-associated operons and is required for antitermination activity *in vivo*. Analyses of LoaP proteins from other species revealed that they too bind the RNA element, while NusG and RfaH exhibited no measurable RNA-binding activity. Finally, our data show that the RNA-binding activity is mediated by C-terminal KOW subdomain, adding a new macromolecular interaction to the already impressive list of

NusG CTD partners (*e.g.*, S10, Rho, and NusA). Together, these data significantly expand the mechanistic diversity of NusG-like proteins and suggest that sub-classes of NusG specialized paralogs are likely to employ unique molecular strategies.

Results

A characteristic DNA sequence is conserved in some LoaP-associated operons

B. velezensis LoaP promotes readthrough of intrinsic terminators in operons that encode for synthesis of difficidin (*dfn*) and macrolactin (*mln*) antibiotics (131). LoaP antiterminators presumably rely on nucleic acid determinants for recruitment to these operons and for antitermination activity. Yet, manual inspection of the *dfn* and *mln* operons did not reveal sequences resembling *boxA*, *boxB*, or *ops*. Both operons are preceded by an unusually long 5' leader region (131, 160), containing an intrinsic terminator that is bypassed upon expression of LoaP. Manual inspection of this region, however, revealed the presence of a small, inverted repeat that is found in similar locations in both antibiotic operons in *B. velezensis* (fig. 3-2). It was not known if this sequence was functional in its DNA or RNA form. However, since the inverted repeat is located downstream of the promoter but upstream of the intrinsic terminator, we speculated that it serves an important role in LoaP-mediated antitermination. As such, it stands to reason that similar sequences should be present in leader regions of other LoaP-regulated operons. Given its small size and sequence degeneracy, we were unable to use traditional search algorithms for RNA

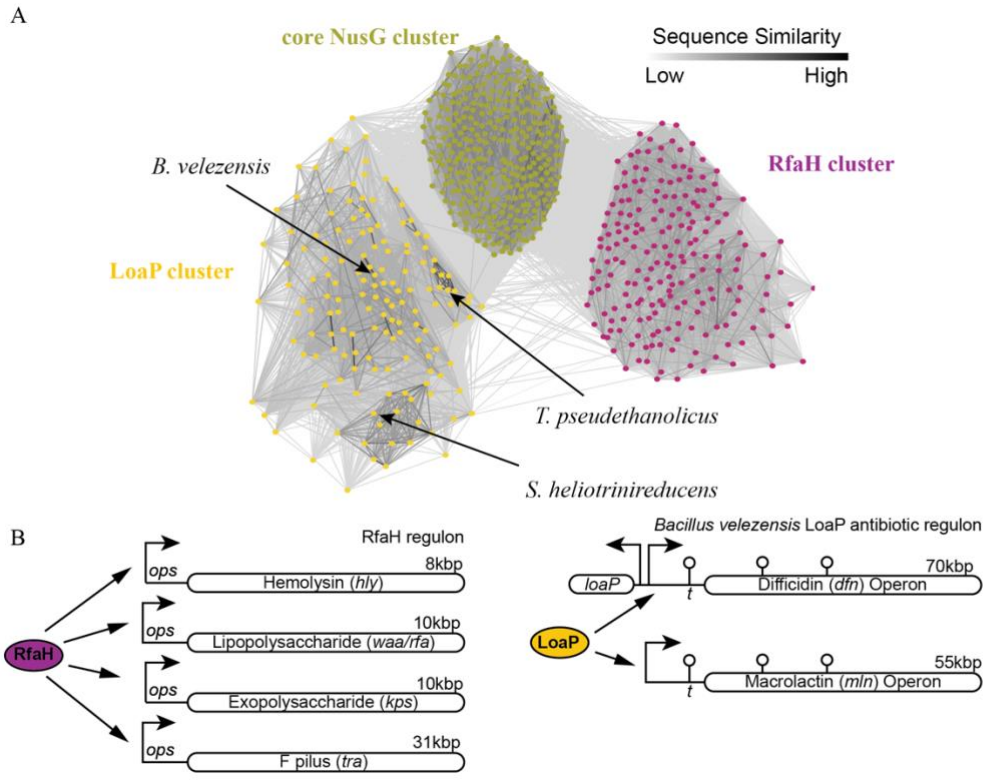


Figure 3-1. LoaP proteins form a coherent outgroup within NusG family of transcription factors. (A) CLANS clustering analysis of 671 NusG family member proteins. Sequences are represented by vertices arranged in the 2-dimensional space. Edges connecting the vertices indicate the pairwise sequence similarity calculated from iterative all-against-all BLAST/PSIBLAST comparisons; edges are shaded on a greyscale according to the P-value of high-scoring segment pairs (black, P-value ≤ 10-324; light grey, P-value ≤ 10-15). The three identified clusters are labeled LoaP (yellow, n = 152), NusG (green, n = 322), and RfaH (magenta, n = 197) based on clustered sequences exhibiting shared-greatest similarity to a reference sequence of the corresponding NusG paralog (LoaP: A7Z6E4; NusG: P0AFG0; RfaH: P0AFW0). Arrows are pointing at LoaP sequences which were chosen to represent LoaP family at large in biochemical experiments. (B) Schematic diagram of a RfaH regulon (left), and the B. velezensis LoaP regulon (right). RfaH recruitment sequence is denoted *ops* in 5' UTR, and intrinsic terminators are denoted with *t*. CLANS clustering analysis was performed by Chih Hao Wu.

motifs (161). Instead, a manual search of sequences within *loaP*-associated operons from other organisms revealed several instances where remarkably similar inverted repeats could be identified within putative 5' leader regions (fig. 3-2). Together, these data suggest that the identified sequence is likely to be functionally involved in LoaP antitermination.

LoaP binds characteristic leader sequence in its RNA form

The λN ATC features a small GNRA RNA hairpin (*boxB*) just downstream of the promoter region, which plays an important role in assembly of the ATC (101, 102). λN binds this RNA element with high affinity and specificity. Secondary structure prediction tools indicated that in its RNA form, the identified inverted repeat folds into a UNCG-type hairpin (162) with two single-nucleotide bulges in the base-paired portion (fig. 3-2). Since *boxB* GNRA hairpin is similar to the size and complexity of UNCG-type hairpin, we speculated that LoaP might perform a role similar to λN, by associating to the inverted repeat in its RNA form.

We purified LoaP proteins from *Bacillus velezensis*, *Thermoanaerobacter pseudethanolicus*, and *Slackia heliotrinireducens* (fig. 3-1) to represent the LoaP subfamily at large. DNA and RNA molecules encoding the inverted repeat sequence from the difficidin (*dfn*) operon in *B. velezensis* were purchased from IDT, radiolabeled and incubated with varying concentrations of LoaP proteins. Nucleic acid sequences used in the equilibrium binding experiments were modified to include additional GGAAA sequence appended to 5' terminus to improve radiolabeling efficiency. Total

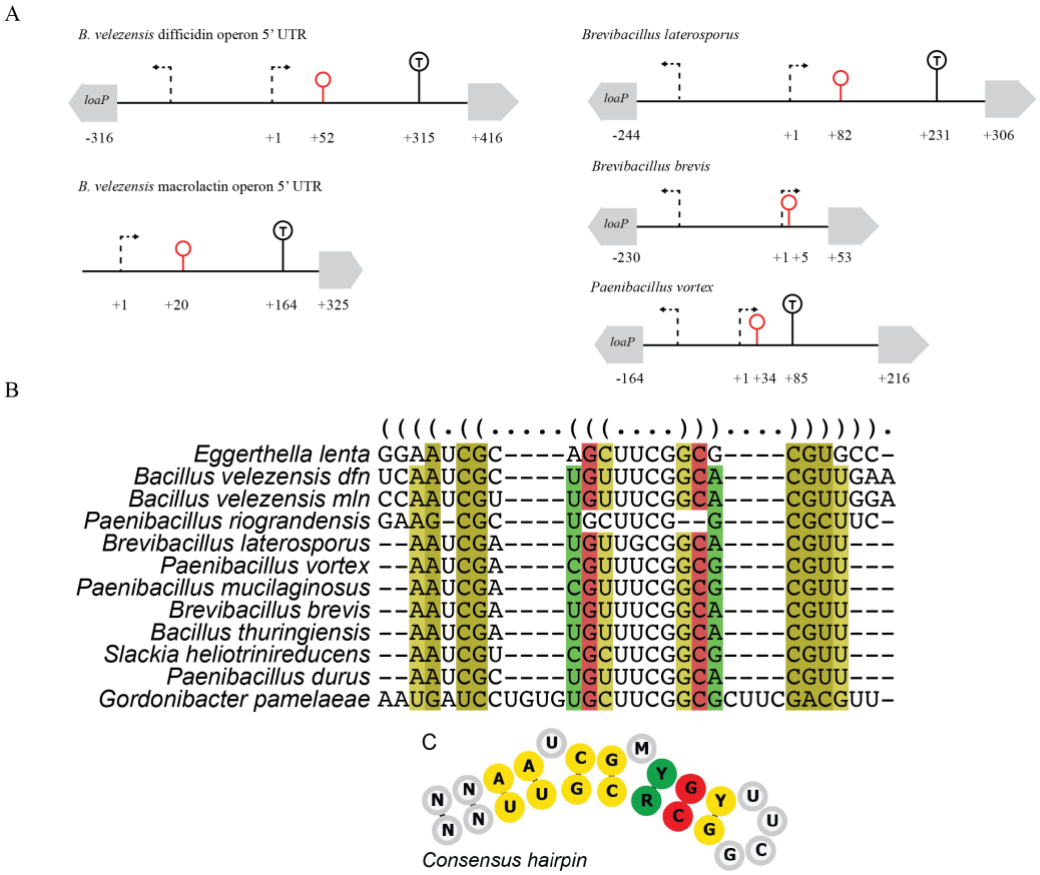


Figure 3-2. Identification of a short, conserved sequence in the 5' leader region of LoaP-associated gene clusters. (A) Schematic diagram depicting location of identified conserved leader sequence (shown in red) in difficidin and macrolactin operons in *B. velezensis*. This sequence is also found in similar locations within 5' UTR in LoaP-associated operons in *B. laterosporus*, *B. brevis*, and *P. vortex*. Intrinsic terminators are denoted with T. (B) *Left*: sequence alignment of conserved leader sequence from multiple species. *Right*: representation of predicted secondary structure, wherein M is either Adenine or Cytosine, Y is pyrimidine, N is any base, and R is purine. Base-paired residues in putative secondary structure are highlighted with identical colors, while unpaired residues in hairpin loop and stem are not highlighted.

yeast RNA (100:1 molar excess) was added as competitor to reduce detection of nonspecific ionic interactions.

These protein-RNA mixtures were spotted onto nitrocellulose membranes and investigated by the differential radial capillary ligand diffusion assay (DRaCALA) (163, 164), a method used previously for quantifying RNA-protein complexes (165). All three LoaP proteins associated with *dfn* sequence only in its RNA form, with equilibrium dissociation constants (K_D) of 160 ± 9 nM, 57 ± 5 nM, and 69 ± 4 nM, respectively (fig. 3-3), while no measurable binding activity was detected for the same sequence in the DNA form. To corroborate these results, *B. vel* LoaP was titrated to reactions containing *dfn* RNA sequence which was labeled with Cy3 fluorophore at 5' terminus and binding interactions were quantified by fluorescence anisotropy (fig. 3-3C). The equilibrium dissociation constant (K_d) obtained using this method was approximately 3-times lower (58 ± 9 nM) than that obtained from DRaCALA (160 ± 9 nM) for the same LoaP protein. The discrepancy in the measured values likely arises from the type and concentration of the competitors added to the binding reactions. Instead of total yeast RNA (100:1 molar excess) used in DRaCALA-based binding experiments, non-template DNA *dfn* sequence was added to fluorescence anisotropy reactions (10:1 molar excess). It is likely that total yeast RNA acts as a better competitor than single-stranded DNA, and therefore reduces non-specific LoaP-RNA interactions resulting in a higher K_d value. However, both values clearly indicate that LoaP proteins exhibit high affinity RNA-binding activity *in vitro*.

LoaP proteins exhibit specific RNA-binding activity

To further investigate the specificity of the LoaP-RNA complex, we incubated a radiolabeled, wildtype hairpin (1 nM) with LoaP (250 nM) and then added increasing amounts of competitor RNAs, each containing different mutational alterations of the *dfn* hairpin (fig. 3-5). When the terminal loop was altered to GNRA sequences (M2, M10), the resulting RNAs could not compete for binding. Similarly, deletion of both side bulges (M4) resulted in RNAs that could not act as competitors. A reduced ability to compete for binding was observed when one of the side bulges was deleted (M7, M8) and when the base-pairing residues were swapped within the hairpin helix (M5, M6). However, a mutant RNA containing a single nucleotide change of the second position of the UNCG hairpin (M9) competed similar to a wild-type RNA. Together, these data suggest that LoaP proteins bind with high affinity and specificity to the *dfn* RNA hairpin and that LoaP is likely to recognize determinants located in the bulged residues and terminal loop.

To corroborate these results, LoaP proteins were incubated with an unrelated RNA: *boxB* hairpin (fig. 3-4). Almost no RNA-binding activity was detectable, except at the highest concentrations (mid-micromolar range). Furthermore, deletion of the bulged residues (M4) in the *dfn* hairpin resulted in a 10-fold decrease in binding affinity (fig. 3-4). This is expected since M4 hairpin did not compete with wildtype hairpin for binding LoaP further confirming that the two bulged nucleotides are required in this binding interaction. A moderate decrease in binding affinity was observed when the terminal loop was swapped with a GAAA sequence (M10), although the extent of this decrease varied among the three LoaP proteins. For example, *S. heliotrinireducens*

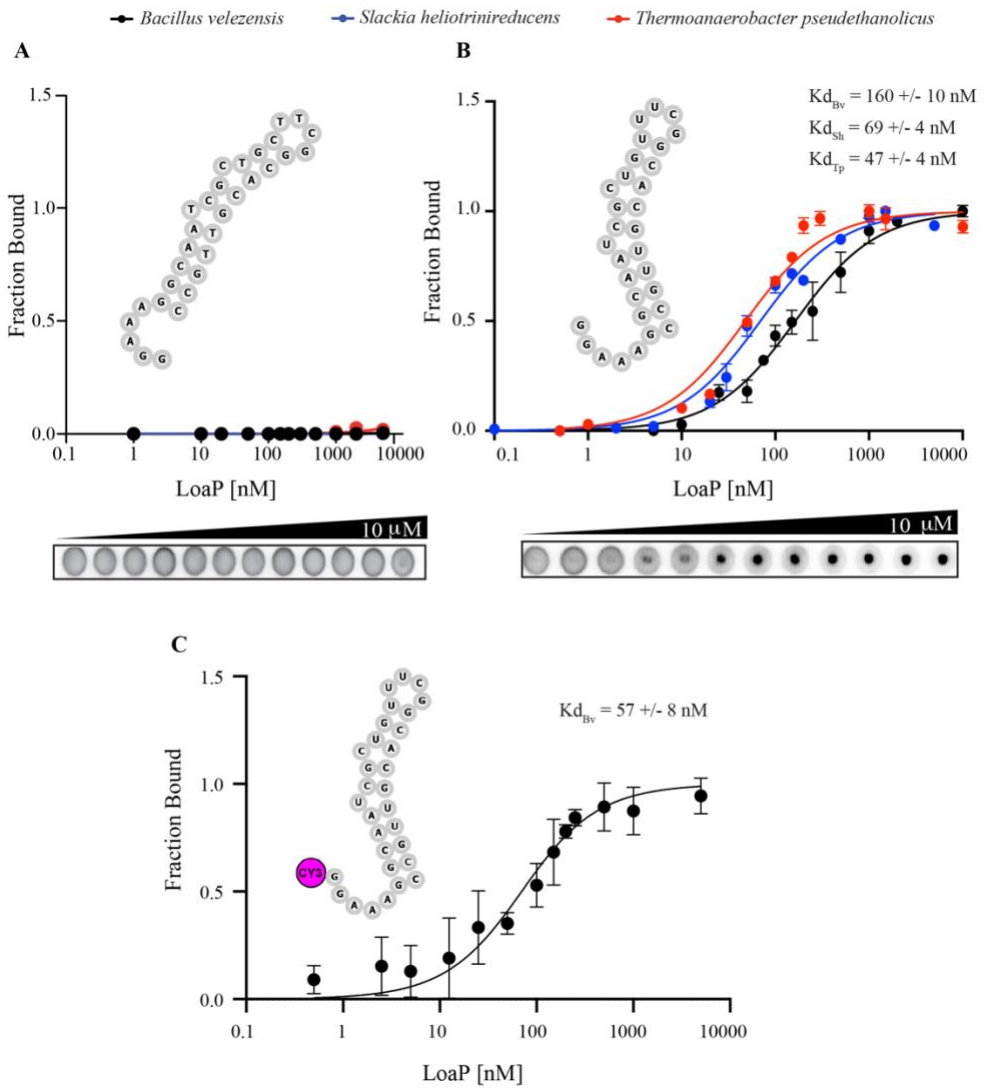


Figure 3-3. LoaP proteins bind conserved leader sequence in its RNA form. Equilibrium binding curves of *B. velezensis* (Bv), *S. heliotrinireducens* (Sh), and *T. pseudethanolicus* (Tp) LoaP proteins incubated with conserved leader sequence in DNA form (A) and RNA form (B) and spotted on nitrocellulose paper (*Bottom:* raw binding data showing diffusion of radiolabeled receptors on nitrocellulose paper). (C) Fluorescence anisotropy-based binding curve of *B. velezensis* LoaP incubated with *dfn* RNA which was labeled with Cy3 fluorophore at 5' terminus. Binding measurements are derived from at least three experimental replicates and shown as the normalized fraction bound, with error bars reflecting standard deviation from the mean.

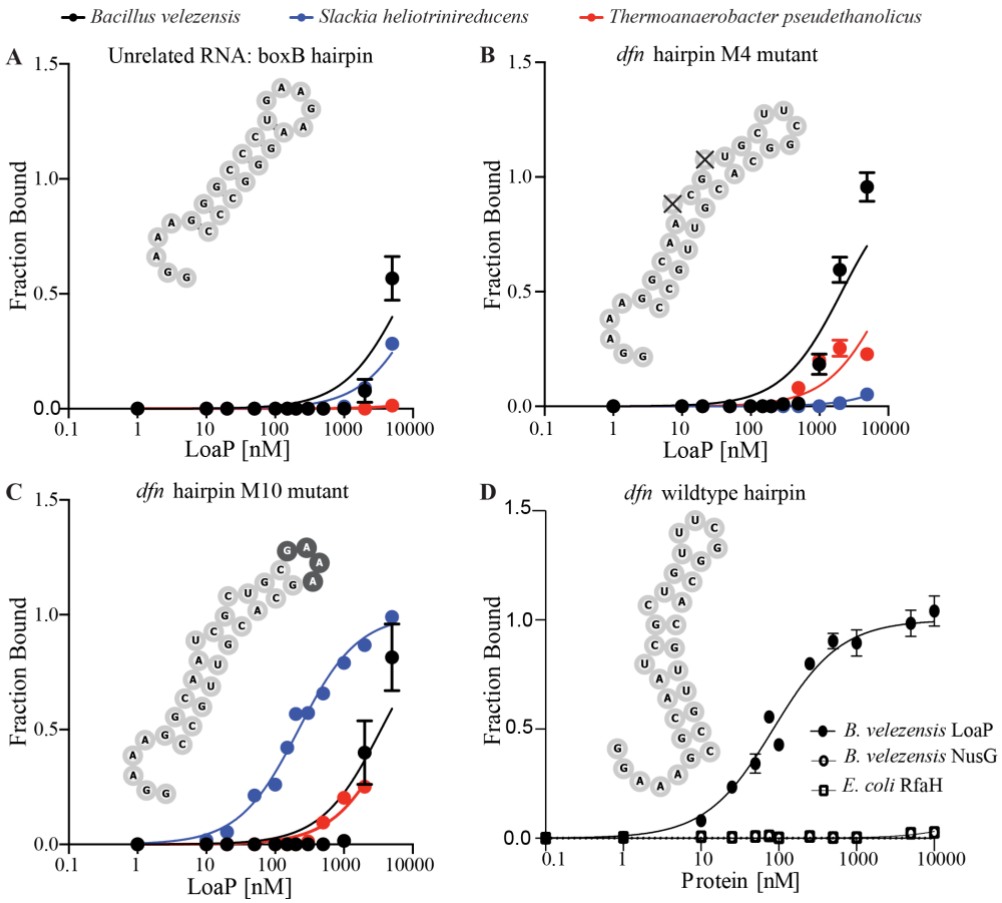


Figure 3-4. LoaP proteins exhibit specific RNA-binding activity. Equilibrium binding curves of *B. velezensis* (Bv), *S. heliotrinireducens* (Sh), and *T. pseudethanolicus* (Tp) LoaP proteins incubated with *boxB* RNA hairpin (A), *dfn* mutant M4 (B), and *dfn* mutant M10 (C). Mutations are represented with X to denote residue deletion, or shaded in gray to denote change in residue identity relative to wildtype *dfn* RNA hairpin. (D) Binding affinity of *B. velezensis* LoaP, NusG, and *E.coli* RfaH to *dfn* RNA hairpin. Binding measurements are derived from at least three experimental replicates and shown as the normalized fraction bound with error bars corresponding to standard deviation from the mean.

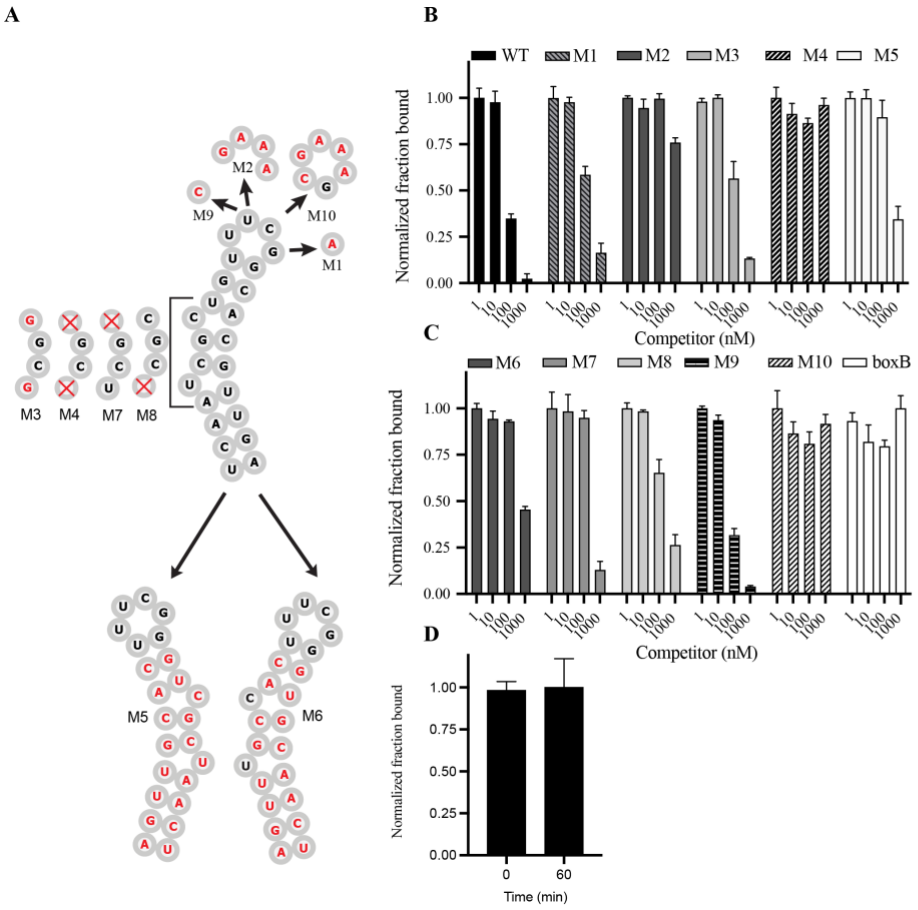


Figure 3-5. Conserved RNA residues mediate interactions with LoaP proteins. (A) Schematic representation of *dfn* RNA mutants (M1-M10) used as competitors. Site-directed mutations are shown in red. (B-C) Competition assays of mutant RNAs against wildtype *dfn* hairpin. Wildtype *dfn* hairpin (1 nM) was 5'-radiolabeled and incubated with *B. velezensis* LoaP (250 nM). Increasing amounts of unlabeled competitor RNAs (1, 10, 100, and 1,000 nM) were then added to the LoaP-RNA mixture and aliquots were assessed by DRaCALA. (D) As a negative control, wildtype LoaP-RNA complex was analyzed using DRaCALA after incubation at room temperature for 60 minutes to ensure complex stability within the time range of competition assays. Binding affinity measurements are derived from at least three experimental replicates and shown as the normalized fraction bound.

LoaP exhibited a 3-fold decrease for the tetraloop replacement, while *B. velezensis* LoaP shifted to the low-micromolar range. This may suggest that different LoaP family sequences may exhibit differences in the type of contacts involved in the interaction with the RNA hairpin. Intriguingly, the observation that LoaP proteins exhibit high affinity and specific RNA-binding activity was unexpected, as there is no precedence for high affinity RNA-binding activity by bacterial NusG proteins. But to directly test this possibility, we incubated *B. velezensis* NusG and *E. coli* RfaH with *dfn* RNA hairpin and observed no evidence of binding activity.

Analysis of LoaP sequences reveal unique conservation pattern

B. velezensis LoaP exhibits approximately 23% sequence similarity to core NusG from the same bacterium. Features that are unique to the LoaP subfamily of NusG proteins should stem from amino acid preferences that may be diagnostic for LoaP proteins. Presumably, a portion of these amino acid preferences confer unique molecular functions of the respective NusG family proteins, such as LoaP's RNA-binding ability.

To investigate this hypothesis further, we searched for variations in the amino acid composition of LoaP sequences, which appeared to be distinct from that of core NusG in both the NTD and CTD. This comparative sequence analysis revealed that many residues in both domains exhibit conservation patterns unique to LoaP sequences, as compared to NusG or RfaH. We assumed that the CTD constitutes a more logical site than NTD for RNA-binding activity, for several reasons. First, during transcription elongation, NusG NTD is affixed to RNAP at a conserved binding site, while the flexibly linked CTD is free to associate a multitude of auxiliary factors, and it could in

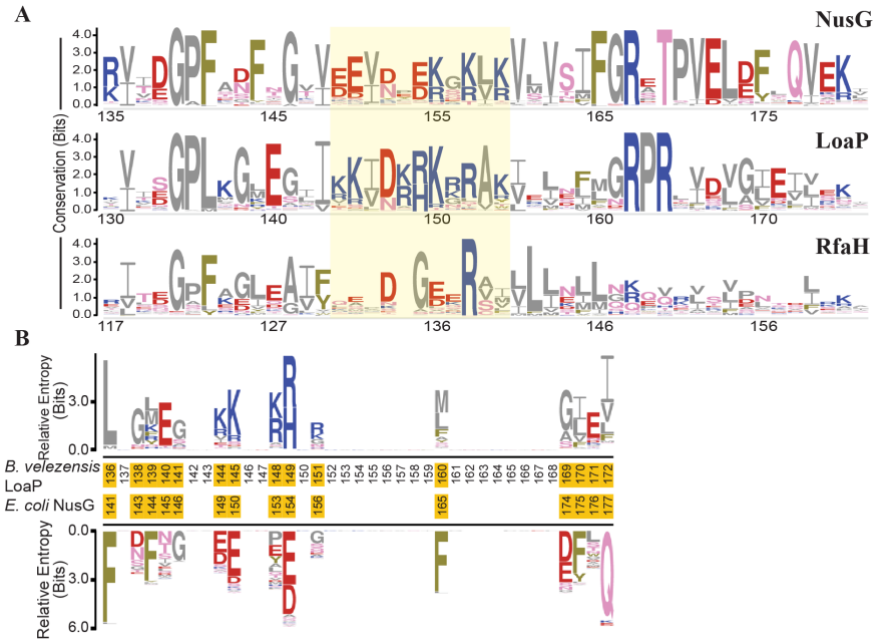


Figure 3-6. Identification of amino acid residues that are conserved for RfaH, NusG and LoaP C-terminal domains. (A) Weblogo depiction of sequence conservation for a portion of the CTD for proteins identified by the clustering analysis shown in fig. 3-1; stack widths are scaled to the relative propensities of the amino acids. Residues are colored by their chemical properties (basic, blue; acidic, red; aromatic, green; non-polar, grey; polar-uncharged; purple). The region shown corresponds to the ungapped sequences of E. coli NusG (P0AFG0), starting from position 135. Top: *E. coli* NusG, middle: *B. velezensis* LoaP, bottom: *E. coli* RfaH. Yellow shaded area corresponds to a small region within LoaP CTD which is enriched in basic amino acids in contrast to core NusG, and RfaH. (B) Comparative sequence logos identify positions which are most enriched within LoaP sequences, as compared to NusG sequences. *Top*: Across the positive y-axis are residues which are enriched in LoaP sequences but infrequent in the corresponding positions of NusG sequences. *Bottom*: Across the negative y-axis conversely show residues enriched in NusG sequences, but infrequent in LoaP sequences. Stack heights are scaled to the position-specific relative entropy calculated between LoaP and NusG clustered sequences; individual letter heights indicate the amino acid frequencies. Only those positions above the 75th percentile of relative entropy are shown.

theory encounter nascent RNA emerging from the RNA exit tunnel. Second, NusG CTD contains a KOW subdomain, which aside from NusG proteins can also be found in several classes of RNA-binding proteins, such as ribosomal proteins (166), RNA helicases, and rRNA processing factor Mtr4 (167). Therefore, precedence has been established for some KOW domain-containing proteins to exhibit RNA-binding activity.

We compared the distribution of amino acids within CTD at every position between LoaP and NusG protein sequences using relative entropy or Kullback-Leibler divergence (168, 169), which quantifies the difference in amino acid frequency between the two groups to identify substantial differences even in the absence of near-invariant positions. Higher entropy scores (fig. 3-6) correspond to amino acid residues which are selectively enriched in either LoaP and NusG and do not share similar chemical properties (i.e polarity, hydrophobicity, and charge). Contrarily, low entropy scores correspond to residues which are conserved in both proteins but share similar chemical properties. This analysis identified a small number of residues which exhibit distinct enrichment patterns within the CTD of LoaP or NusG sequences.

In particular, LoaP positions 144, 145, 148 and 149 show enrichment for positively charged arginine or lysine residues relative to NusG, which instead contains negatively charged residues at the corresponding positions, thus higher entropy scores. We speculate that this conserved pattern of basic residues might resemble arginine-rich motifs (ARM), which are present in several classes of RNA-

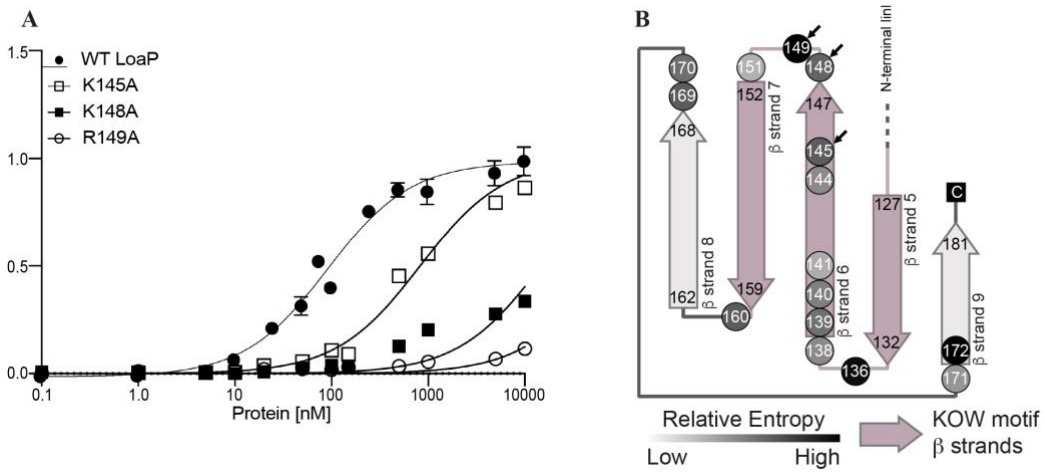


Figure 3-7. Site-directed mutation of CTD residues impairs RNA-binding activity of LoaP. (A) Equilibrium binding curve of three *B. velezensis* LoaP containing alanine substitutions at positions 145, 148, 149 to wildtype *dfn* RNA hairpin. Binding affinity measurements are derived from at least three experimental replicates and shown as the normalized fraction bound, with error bars corresponding to standard deviation from the mean. (B)Visualization of high relative entropy positions, corresponding to those highlighted in figure 3-6, shown schematically on a topology diagram of the predicted secondary structure for LoaP CTD. Arrows indicate residues that were altered by site-directed mutagenesis to alanine.

binding proteins, such as λN and HIV Tat (170, 171). Moreover, the basic residues that are preferentially exhibited by LoaP sequences are all located within the KOW subdomain. Therefore, we speculated that specific RNA-binding activity is most likely to occur at the KOW subdomain within LoaP CTD.

LoaP CTD residues are involved in RNA binding

Based on the pattern of amino acid preferences displayed by LoaP sequences, we mutated three CTD residues to assess their importance for RNA-binding activity. K145, K148, and R149 were individually altered to alanine and the mutant proteins were assayed for their ability to bind RNA (fig. 3-7). This revealed that K145A bound the hairpin RNA with 4-fold reduced affinity compared to wild-type LoaP (K_D ~ 600 nM). Furthermore, K148A and R149A resulted in complete loss of any detectable RNA-binding activity. These findings together indicate that LoaP's RNA-binding activity is localized in its CTD, most likely involving residues in the region between β strands 6 and 7 (fig 3-7).

The *dfn* hairpin is required for LoaP-mediated antitermination activity *in vivo*

We subcloned the *B. velezensis dfn* leader sequence upstream of a *yfp* reporter, which was then expressed in *B. subtilis* (fig. 3-8). This reporter was used to measure antitermination activity by quantifying *yfp* fluorescence signal. Briefly, when the elongation complex encounters the intrinsic terminator located within 5' UTR, transcription terminates prematurely therefore *yfp* is not expressed. Alternatively, if the elongation complex bypasses the intrinsic terminator via LoaP-mediated

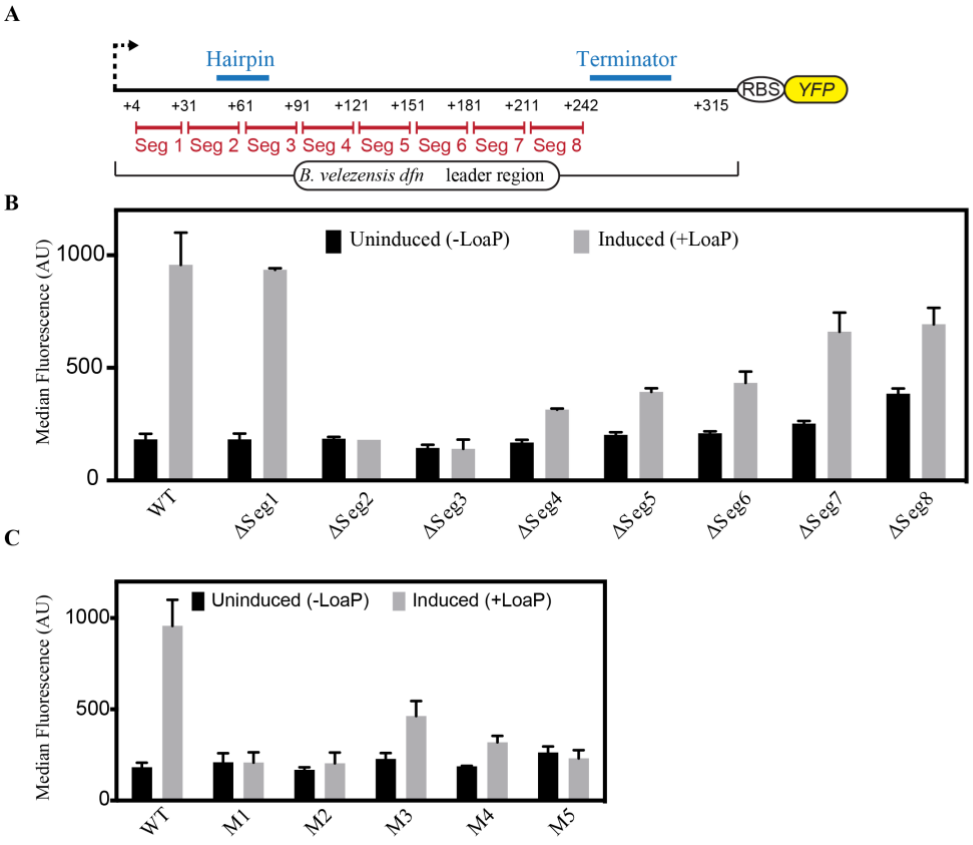


Figure 3-8. The *dfn* hairpin is necessary for LoaP-mediated antitermination *in vivo* (A) Schematic depiction of *dfn* leader-*yfp* reporter construct, which was integrated into the *B. subtilis* nonessential *amyE* gene. A xylose-inducible copy of the *B. velezensis loaP* gene was then integrated into the *thrC* locus. Locations of the UNCG hairpin and intrinsic terminator are denoted in blue and sites of 30 base pair deletions are indicated in red. (B) Median fluorescence of flow cytometry analyses for strains containing deletions within the *dfn* leader-*yfp* reporter, and single site mutations (C). The standard deviation for three biological replicates is shown.

antitermination, *yfp* is expressed and fluorescence signal is measured. The results indicated that fluorescence is dependent upon LoaP expression. Thus, LoaP determinants should be fully contained within the *dfn* leader region, upstream of the intrinsic terminator and upstream of the *dfnA* coding region.

To assess the importance of *dfn* RNA in LoaP-mediated antitermination, YFP fluorescence was assessed for *dfn* leader-*yfp* constructs containing deletions of 30 base pair segments (fig 3-8). Deletion of segments which encode for LoaP determinants should hypothetically abrogate antitermination activity. While several deletions resulted in reduced YFP fluorescence, two segments, each encoding for a portion of *dfn* RNA, resulted in a complete loss of antitermination activity (between +31 and +91, relative to the transcription start site). This indicates that *dfn* RNA plays a critical role in LoaP-mediated antitermination.

To corroborate these results, several of the hairpin mutations used in equilibrium binding assays (fig 3-8C) were then introduced into the *dfn* leader-*yfp* reporter fusion and assessed for antitermination activity *in vivo*. Most mutations that resulted in a significant loss in RNA-binding activity (M2, M4, M5) also exhibited a complete loss in antitermination activity. However, the M3 mutant, which contains a change in sequence for both nucleotide bulges, still competes for binding to LoaP *in vitro* but exhibited a reduction in antitermination activity *in vivo*. Similarly, the M1 mutant, which alters a single residue of the terminal loop, binds LoaP *in vitro*, but caused a complete loss in antitermination activity *in vivo*. These data further confirm the importance of the hairpin and suggest that its functional requirements may be greater *in vivo* than for equilibrium binding to LoaP *in vitro.*

Discussion

The *E. coli* NusG (79-81) NGN domain binds to the surface of the TEC near the conserved clamp helices of the β' subunit and the gate loop of β subunit, arranging NusG at the solvent exposed surface of the transcription bubble in close proximity to the non-template (NT) DNA (77, 82). Once bound to RNAP, *E. coli* NusG NGN domain forms a processivity clamp to maintain RNAP in a closed, pause-resistant state, while the CTD orchestrates interactions with other cellular factors to affect transcription and translation. However, there may be some mechanistic differences exhibited by NusG-associated RNAP between different bacteria.

Structural analyses of transcription elongation complexes (TEC) *in vitro* (172) including structural resolution of the TEC by X-ray crystallography and single-particle cryogenic electron microscopy (77, 141), have revealed some specific interactions between NusG proteins and NT DNA. For example, the *Bacillus subtilis* RNAP-associated NusG has been shown to induce pausing at characteristic NT DNA sequences (173-175). Core NusG stimulates pausing at two sites in the leader region of the tryptophan biosynthesis operon (*trp*) by specifically interacting with conserved residues protruding from the transcription bubble. These NusG-stimulated pauses are directly involved in transcription attenuation and translation repression mechanisms involved in tryptophan biosynthesis. The specific interaction between *B. subtilis* NusG and NT DNA in *trp* operons is thought to prevent forward translocation of the elongating RNAP (172) providing sufficient time for *trp* RNA-binding attenuation protein (TRAP) to bind nascent mRNA and prevent the formation of an antiterminator element.

NusG specialized paralog RfaH also exhibits sequence-specific contacts with the NT DNA which are required for its recruitment to the elongation complex and its activation (77, 140, 141). RfaH associates with RNAP through its NTD at the same site normally occupied by core NusG and therefore competes with NusG for occupancy on RNAP. However, RfaH NTD binds specifically to 12-nt DNA sequence called the operon polarity suppressor, or *ops*. It is generally thought that interactions between NusG proteins and nucleic acids, therefore, occur at RNAP-associated NTD because the processivity enhancing functions of NusG is conferred through the association of NGN domain with RNAP. Yet the regulatory diversity of NusG proteins arise predominantly from the network of interactions mediated via flexibly linked CTD to orchestrate the coupling of transcription to other cellular events.

When compared to NusG and RfaH sequences, LoaP NTD exhibits its own pattern of uniquely conserved residues. We speculate that these residues are likely to be involved in associating with the TEC and in recognizing yet-to-be-identified determinants in the NT DNA. Furthermore, our data demonstrate that LoaP proteins exhibits a unique capability to specifically bind a small UNCG-type RNA hairpin which is commonly located in 5' UTR of LoaP-associated gene clusters. This specific RNA-binding activity appear to emanate from conserved residues within CTD, particularly in KOW subdomain.

Although RNA-binding activity has not been previously observed in NusG proteins for bacteria, eukaryotic Spt5 was found to mediate sequence-specific interactions with single-stranded RNA (176, 177). Specifically, Spt5 in yeast (176) associates with RNA sequences bearing multiple AA repeats, and lacking distinct

secondary structure. It is worth noting that Spt5 factor is much larger (~110 kDa), and more complex than bacterial NusG. While bacterial NusG contains an N-terminal NGN domain connected by a flexible linker to a single KOW domain, eukaryotic Spt5 contains an N-terminal acidic domain, an NGN domain, five KOW domains, and a C-terminal repeat domain (CTR). The Spt5 NGN domain is structurally similar to its bacterial counterpart (178), but forms a heterodimer with Spt4, which has no known homologue in bacteria. This Spt4/5 complex directly binds RNAP near the central cleft (179), effectively locking the cleft in a closed conformation. Isolated Spt5 KOW domains from *Saccharomyces cerevisiae* lack specific RNA-binding activity (176, 178). Instead, association with specific RNA ligands is mediated via the heterodimer of Spt4 with the Spt5 NGN domain; neither monomer alone exhibits the same RNA-binding activity *in vitro.* These observations indicate that while some Spt5 KOW domains show non-specific and weak binding affinity to nucleic acids, the Spt4/5 NGN complex alone is sufficient to account for the Spt5's RNA-binding activity.

Based on the precedence established by NusG and RfaH, we do not anticipate that NGN residues are the source of LoaP's RNA-binding activity. Instead, we hypothesize LoaP's RNA-binding activity is derived from its CTD. LoaP CTD contains a KOW domain, which aside from NusG proteins can also be found in three families of bacterial and eukaryotic ribosomal proteins (180), where it is employed for binding ribosomal RNA (181, 182). Therefore, while the KOW domain is not universally used for binding RNA and is inessential for binding RNA in Spt5, it has proved to be evolutionarily amenable to this molecular function.

Our LoaP sequence analyses revealed an enrichment of several specific residues in the CTD. The variation in residue identity was quantified as a relative entropy score between multiple sequences of LoaP and NusG. Interestingly, there are only a few residues in LoaP CTD which are highly conserved, but are oppositely charged and equally conserved in NusG and RfaH indicating a potential functional significance that is unique to LoaP CTD. In particular, residues K145, K148, and R149 correspond to residues E150, Y153, and E154 in E.coli NusG CTD (PDB: 2JVV) which are located in a flexible loop between β2: F144-D152, and β3: R157-I164 and are located within KOW subdomain. In addition to their high relative entropy scores, these residues are conserved as positively-charged amino acids and thus were selected as potential candidates for RNA-binding. Other residues which also exhibit high relative entropy scores (i.e. L146 and E140) but were otherwise hydrophobic or acidic were not selected for single-alanine mutagenesis. Indeed, single-alanine mutagenesis of the highest scoring basic residues (K145, K148, and R149) weakened the RNA-binding affinity of LoaP. While one mutant K145A weakened the binding affinity approximately10-fold, the two other mutants K148A and R149 shifted the equilibrium dissociation constant beyond the titration range of these binding assays (>10 μM). From these aggregate observations, we speculate that LoaP's RNA-binding activity emanates from KOW residues located between β strands 6 and 7.

The association of LoaP with the *dfnA* RNA hairpin is functionally important *in vivo* as disruption of this interaction results in premature transcription termination. Yet, the role of this RNA-protein complex during antitermination is unclear. In the λ and *rrn* ATC, the *boxB* RNA hairpin helps recruit cellular factors, which among other

outcomes, repositions NusA moved away from the RNA exit tunnel, causing a decrease in termination efficiency. It is yet to be determined if LoaP-mediated antitermination requires participation of other Nus elongation factors like NusA, NusE, and NusB. Therefore, while the mechanistic details of LoaP antitermination still remains elusive, our data suggest that the interaction between LoaP and cognate RNA hairpin play a critical role in its processive antitermination mechanism.

Materials and Methods

Clustering analysis

We extracted 43,836 protein sequences containing a single NusG NGN domain (PFAM: PF02357). Most of these sequences also contained a KOW domain (PFAM: PF00467) in tandem with the NGN domain. A BLASTp search performed against this dataset using LoaP (ID: A7Z6E4), RfaH (ID: P0AFW0), and NusG (ID: P0AFG0) reference sequences produced three sets of sequences that were further filtered to remove redundant (above 60% sequence identity) or truncated sequences, resulting in 671 sequences that clustered into groups corresponding to LoaP, RfaH, and core NusG. After filtering, a clustering analysis was performed on the resultant 671 sequences using CLANS in 2-dimensional space; clusters larger than 100 sequences were manually selected to represent the different NusG paralogues described herein.

Strain construction

Plasmids were assembled using standard molecular cloning strategies and transformed into *E. coli* XL10-Gold. Deletions of the *dfnA* leader region were accomplished using NEB's Q5 Site-Directed Mutagenesis Kit. All plasmids were

verified by Sanger sequencing. *B. subtilis* 168 strains were transformed using a protocol described previously (124). For mutants containing deletions of *nusB*, *nusG*, or *rho*, the initial knockout strains were acquired from the Bacillus Genetic Stock Center (Columbus, Ohio) which contained a marker replacement of the target gene for an erythromycin resistance cassette. The latter was subsequently removed using Cre recombinase to create markerless strains. A plasmid containing a reporter fusion of the *dfn* leader region fused to the *yfp* gene was then integrated into the nonessential *amyE* locus. A plasmid containing a xylose-inducible copy of the *B. velezensis loaP* gene was then integrated into the *thrC* locus. To create the *nusA* deletion strain, a tetracycline-inducible copy of the *nusA* gene was first integrated into the nonessential *ganA* locus. The chromosomal copy of *nusA* was then deleted by marker replacement using a nonreplicable plasmid. Growth of this strain is highly dependent on the presence of the tetracycline inducer (not shown), as NusA is an essential gene.

Flow cytometry

Aliquots of 5 mL LB were inoculated with 100 mL of overnight culture. Cultures either received 100 mL 25% xylose (for a final concentration of 0.5%) or 100 mL sterile water. The cultures were incubated shaking at 37°C for 3h, then pelleted and washed in phosphate-buffered saline (137 mM NaCl, 2.7 mM KCl, 10 mM Na_2HPO_4, 2 mM KH_2PO_4) before being resuspended in phosphate-buffered saline to an OD_{600} of 0.1 and subjected to flow cytometry. During analysis, a gate was applied to exclude events that appeared to be outliers in a plot of FSC versus SSC. The same gate was applied to all samples.

Differential radial capillary action of ligand assay (DRaCALA)

RNA and DNA molecules used in the *in vitro* binding assays were purchased from IDT (Appendix) and 5'-radiolabeled with [γ-^{32}P] ATP using T4 Polynucleotide Kinase (T4 PNK; New England Biolabs). RNA was purified using the Zymo RNA Clean & Concentrator Kit™ and stored in 1x folding buffer (50 mM NaH_2PO_4 pH 7.2, 100 mM NaCl, 0.1 mM EDTA). RNA and DNA folding was performed by heating the samples in 1x folding buffer at 85°C for 2 min and quick-cooling in an ice-bath for 10 min. Equilibrium binding reactions were set up in 96-well plates by incubating (~1 nM) radiolabeled RNA/DNA with increasing amounts of LoaP in binding buffer (50 mM NaH_2PO_4 pH 7.2, 100 mM NaCl, 0.1 mM EDTA, 1 mM $MgCl_2$, 100 mg/mL BSA, 25 mg/mL total yeast RNA) for 45 min at room temperature. 2 μL aliquots were spotted onto nitrocellulose membrane using a fixed replicator pin tool, and allowed to air dry for 20 min. Nitrocellulose membranes were exposed to a phosphor screen for 20 min and visualized using a phosphorimager and quantified using ImageQuant and Graphpad Prism software.

The fraction bound was determined by measuring the signal intensity sequestered in the inner circle (I_{Inner}) divided by the total signal intensity of the spot (I_{Total}). However, since the inner circle overlaps with the total area of the spot, the intensity of the inner circle represents the sum of bound ligand and unbound ligand. To accurately determine the intensity of the bound ligand, the intensity of the background ($I_{background}$) was subtracted from the intensity of the inner circle (I_{Inner}). Assuming uniform distribution of unbound ligand in the spot, the background intensity ($I_{background}$) was calculated by multiplying the intensity of the outer circler by the area of the inner

circle (see equation 3.1). The total ligand bound by the protein was then calculated by subtracting $I_{background}$ from I_{Inner}, which was then divided by the total intensity to calculate the fraction bound (equation 3.2). A more rigorous description of this analysis is outlined in (165). Values representing the fraction bound at each protein concentration were then normalized in GraphPad Prism 9, and the data was analyzed by fitting a steady-state quadratic curve using nonlinear regression (equation 3.3).

Equation 3.1:

$$I_{Background} = Area_{Inner} \times \frac{(I_{Total} - I_{Inner})}{(Area_{Total} - Area_{Inner})}$$

Equation 3.2:

$$Fraction\ Bound = \frac{I_{Inner} - I_{Background}}{I_{Total}}$$

Equation 3.3:

$$Fraction\ Saturation = \frac{(R + x + K) - \sqrt{(-R - x - K)^2 - (4 \times R \times K)}}{2R}$$

*R is receptor concentration
**x is ligand concentration
***K is equilibrium dissociation constant

Fluorescence Anisotropy binding assay

RNA molecules were purchased from IDT labeled with Cy3 fluorophore at 5' terminus. RNA folding was performed as previously described. Cy3-labeled *dfn* RNA (~ 5 nM) was incubated with LoaP protein at various concentrations in a 96-well plate in binding buffer (50 mM NaH_2PO_4, 100 mM NaCl, 0.1 mM EDTA, 1 mM DTT, 50 nM *dfn* NT-DNA). The reactions were then incubated in the dark at room temperature for 30 minutes to reach equilibrium. The fluorescence polarization was measured in a

Molecular Devices Spectramax M5 plate reader at 535 nm excitation and 580 nm emission. Data was fit using GraphPad Prism 9 using equation 3.3.

Chapter 4: Crystallographic studies of LoaP-RNA complex

Introduction

NusG factors are typically comprised of two independently-folded domains (84). The NusG N-terminal domain (NTD) binds the TEC surface near the non-template (NT) strand of the transcription bubble (77, 82) to maintain RNAP in a pause-resistant state (76) or, conversely, to induce pausing at characteristic NT DNA sequences (173-175). An unstructured linker connects the NTD to a C-terminal (KOW) domain. This domain is distally positioned (fig. 4-1), such that it can transiently associate with several cellular factors, including Rho, NusA, and NusE (S10). The complex formed between S10 and NusG helps bridge interactions between the leading ribosome and RNAP (78, 87-89, 105). When not bound to NusE (S10), a mutual exclusive interaction can be made between the NusG CTD and Rho, which aids Rho-mediated termination (46, 106).

The modular structure of NusG proteins underpins their diverse regulatory role during transcription elongation which ultimately determines the transcriptional outcome. The NusG NTD exhibits a mixed α/β topology (85) that mediates direct contacts with the two largest RNAP subunits. The crystal structure (181) of NusG from *Aquifex aeolicus* revealed that the NTD is comprised of a four-stranded antiparallel β-sheet which is flanked by three α-helices: two on one side, and a single helix on the other. Interestingly, there is a striking congruence between the structure of this domain and a well-characterized RNA-binding motif, which is commonly found in

ribonucleoproteins (RNPs) (183), suggesting a potential functional role for NusG-nucleic acids interactions. This is perhaps not entirely surprising given that the conserved RNAP-binding site for NusG spans the non-template DNA interface; therefore, NusG is located in close proximity to the solvent-exposed non-template DNA strand. In line with these observations, the binding of *B. subtilis* NusG to conserved DNA residues protruding from the transcription bubble (73) in the leader region of the tryptophan biosynthesis operon (*trp*) indicate that interactions with nucleic acids could be functionally important for NusG proteins, although the specificity of these interactions is likely to vary between species.

In addition to the NTD, NusG proteins contain a smaller, flexibly-linked CTD. Despite its small size (<9 kD), the NusG CTD can associate with NusA, NusB, S10 and sometimes Rho. In contrast to the NTD, the NusG CTD folds into a five-stranded β-barrel, and it encompasses a KOW subdomain, which is found in ribosomal protein L24 (182) where it mediates direct interactions with rRNA. However, core NusG and NusG-paralog RfaH have not been shown to directly interact with RNA in solution. It is possible that these interactions could be occluded in the isolated proteins and become uncovered only in functional complexes.

NusG-paralog RfaH exhibits structural features which are distinct from core NusG. Free RfaH exists in an autoinhibited state where the CTD adopts an inactive α-helical conformation instead of the exclusively β-sheet fold typical of the NusG CTD (85) (fig. 4-1C and 4-1D). Moreover, the α-helical form of the RfaH CTD interacts directly with the NTD via conserved residues, partially masking the RNAP-binding site and thereby preventing its recruitment to the elongation complex. Transient contacts to

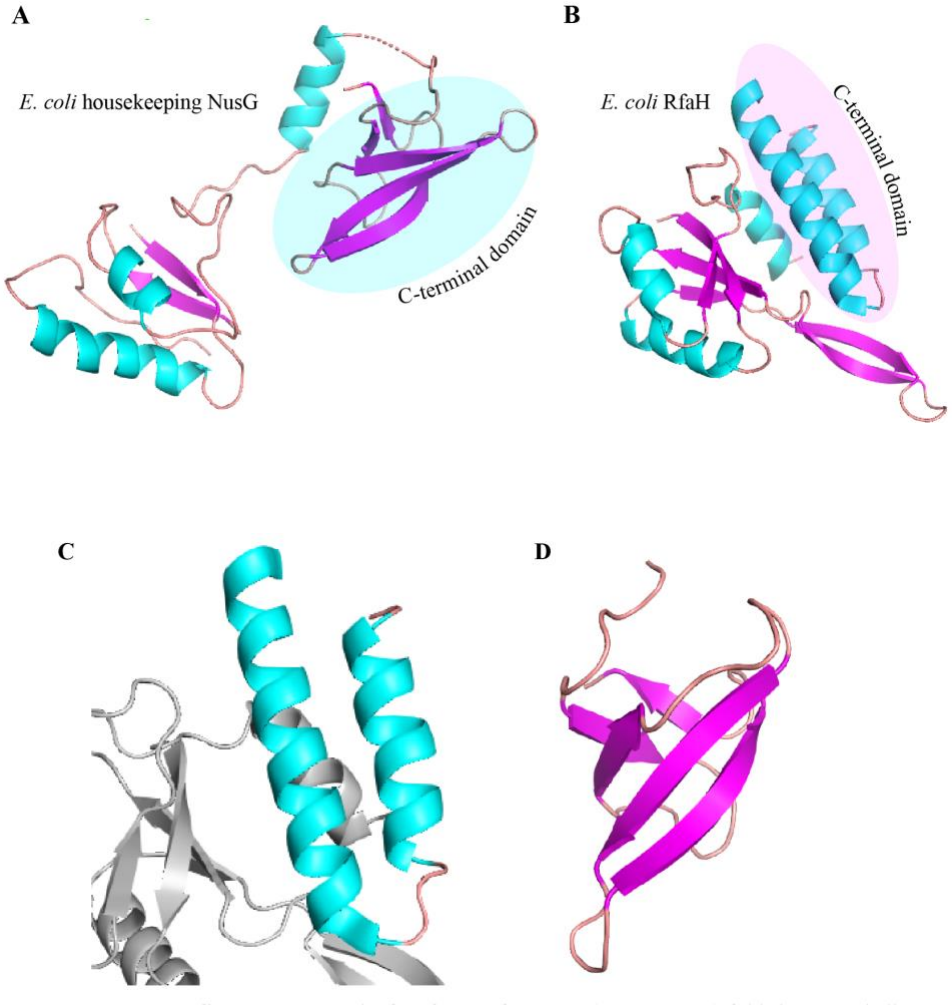


Figure 4-1. Crystal structures of full-length *E. coli* NusG and *E. coli* NusG-paralog RfaH. (A) *E. coli* housekeeping NusG structure was obtained from crystal structure in complex with elongating RNAP (PDB: 5TBZ). NusG proteins typically contain two independent domains: N-terminal domain which adopts a mixed α/β topology and a smaller C-terminal domain folds into a 5-stranded β-barrel. (B) Crystal structure of *E.coli* NusG-paralog RfaH (PDB: 2OUG). In this structure, RfaH adopts an autoinhibited state where CTD forms two α-helical strands which interacts with NTD masking conserved RNAP binding site and therefore prevents recruitment to the elongation complex. In both structures, α-helices are colored in turquoise, while β-strands are colored in magenta. In its free form, RfaH CTD (PDB: 2OUG) folds into an α-helical conformation (C) which autoinhibits RfaH by interacting directly with NTD and preventing association with elongation complex. Additionally, α-helical CTD lacks the ability to interact with the ribosome. Once activated, RfaH CTD (PDB: 2LCL) refolds into 5-stranded β-barrel (D) that resembles NusG CTD, recruits ribosome, and activates translation.

the *ops* sequence are thought to release the CTD from the autoinhibited state, and trigger the re-folding of the entire domain to an active NusG-like β-barrel (85) (fig. 4-1). This refolding event facilitates the interaction between the NTD and the elongation complex but, more importantly, enable the RfaH CTD to interact with ribosomal protein S10 promoting transcription-translation coupling.

It is likely that other NusG specialized paralogs associate with TEC via unique mechanisms and interact specifically with nucleic acid determinant which are encoded by their target operons. These interactions could be essential for the recruitment of specialized NusG paralogs for the same reasons they are for RfaH: to act as recruitment signals encoded in target operons, and to prevent the off-target recruitment of specialized paralogs to non-target operons. Surprisingly, our data show that NusG-paralog LoaP associates selectively and with high affinity to a characteristic RNA hairpin. This RNA-binding activity is mediated by conserved residues within the CTD adding a new macromolecular interaction to the already impressive list of NusG CTD partners. These data significantly expand the mechanistic diversity of NusG-like proteins and confirm that sub-classes of NusG specialized paralogs are likely to employ different molecular strategies.

Herein, we discuss current progress in our crystallographic studies of LoaP-RNA complex. We obtained crystals of L-selenomethionine labeled LoaP-RNA complex which yielded diffraction data with 2.45 Å resolution limit. Furthermore, we show that the electron density maps contain well-defined regions corresponding to RNA and protein components. Currently, model building efforts are underway to solve the crystal structure of LoaP-RNA complex, and recent results will be discussed.

Results

LoaP forms a stable complex with *dfn* RNA

LoaP-RNA complexes were prepared by incubating 100 μM purified *T. pseudethanolicus* LoaP with 100 and 150 μM purified *dfn* RNA hairpin in 500 μL crystallization buffer (10 mM Tris.HCl pH 7.2, 100 mM NaCl, 0.1 mM EDTA, 1 mM DTT) at room temperature for 45 minutes. The complexes were then purified by loading onto a Superdex 75 increase 10/300 GL (Cytvia) size-exclusion column at 0.25 mL/min at room temperature, which was pre-equilibrated in crystallization buffer. In these experiments, *T. pseudethanolicus* LoaP protein (21.0 kDa) was used as a size marker (fig. 4-2) and eluted at 12.0 mL. Samples containing LoaP-RNA complexes at 1:1 and 1:1.5 (protein:RNA) molar ratios were then analyzed on size-exclusion column to test for complex formation and stability. UV-spectra obtained from these runs indicate that LoaP-RNA complexes were stable under these conditions throughout the entire period of the experiment (~2 hours). Moreover, addition of purified *dfn* RNA (8.3 kDa) at 1:1 molar ratio is sufficient to saturate LoaP proteins (fig. 4-2). This was further confirmed by analyzing complexes formed at 1:1.5 molar ratio which revealed excess unbound RNA eluting at 13.4 mL (fig. 4-2). We concluded that adding RNA at 1:1 molar ratio is therefore sufficient to prepare LoaP-RNA complexes for crystal screening.

Analysis of crystal composition indicate protein and RNA components

To determine the molecular composition of diffracting crystals and confirm the

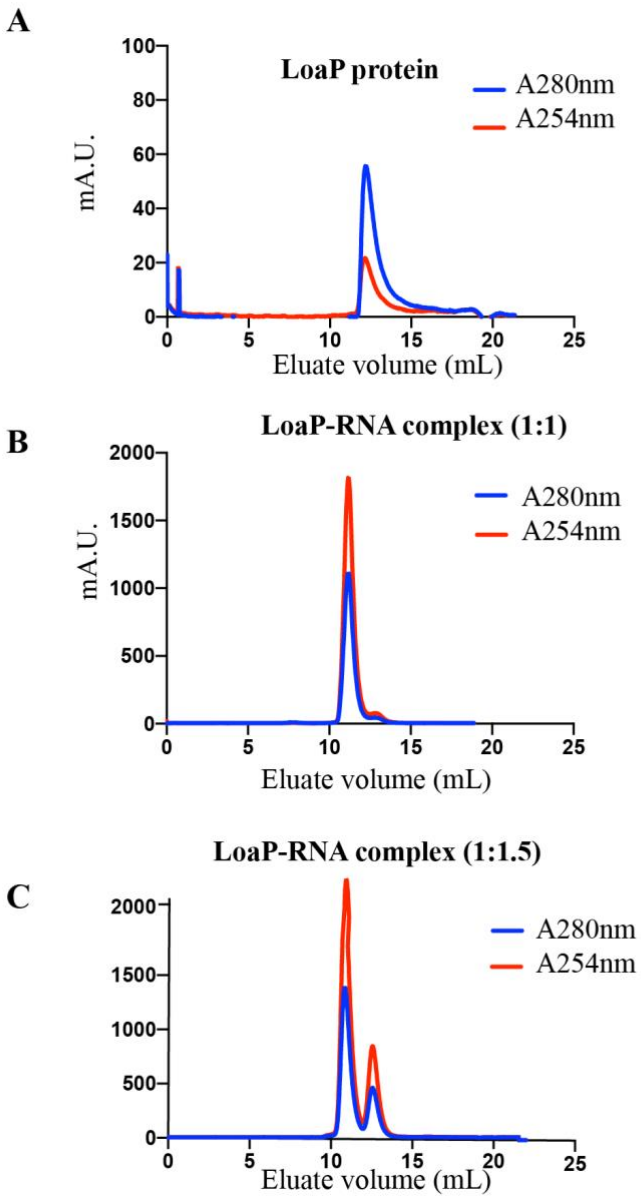


Figure 4-2. Preparation of LoaP-RNA complex for crystal screening. LoaP sequence from thermophilic bacterium *Thermoanaerobacter pseudethanolicus* was purified as previously described then loaded onto Superdex 75 increase size-exclusion column to exchange into crystallization buffer (10 mM Tris.HCl pH 7.2, 100 mM NaCl, 0.1 mM EDTA, 1 mM DTT). LoaP-RNA complexes were formed by incubating purified LoaP with *dfn* RNA hairpin in crystallization buffer at room temperature for 45 minutes. LoaP-RNA complexes formed at 1:1 (B), 1:1.5 (C) molar ratio were then analyzed on Superdex 75 increase size-exclusion column to assess complex formation. Purified LoaP protein as used as a size marker (A).

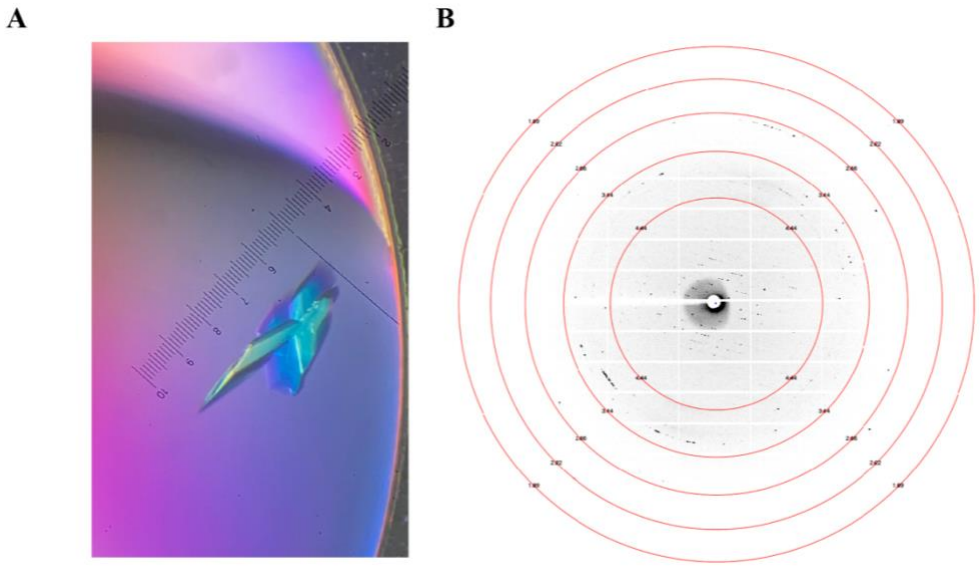


Figure 4-3. LoaP-RNA complex formed crystalline material in a few crystallization buffers. Robotic crystallization trials were performed for LoaP-RNA complex using commercially available screens from Hampton Research and Molecular Dimensions. Crystal screens were set up using sitting-drop diffusion technique: 0.2 μL 100 μM molar LoaP-RNA (1:1) complex was mixed with 0.2 μL crystallization buffer at room temperature, and then incubated at 15°C and 21°C until crystalline material formed in the wells. 950 conditions were screened for LoaP-RNA complex and conditions yielding crystalline material were further optimized to improve crystal growth. One such condition (A) produced crystalline material that grew up to 150 μm x 50 μm in size. These crystals were then harvested for X-ray diffraction data collection. (B) Diffraction data revealed a resolution limit approximately around 2.45 Å.

presence of RNA and protein, crystals were analyzed on PAGE gels. Crystals were first transferred to a 0.1 μm filter and centrifuged at 500 x*g* at 8ºC to get rid of mother liquor (flow through). Crystals were washed three times with ice-cold reservoir solution (wash 1-3), then dissolved in 10 mM EDTA pH 8.0. Samples containing flowthrough, wash, and elution fractions were collected and analyzed by PAGE to visualize proteins and RNA separately. Analysis by 4-20% denaturing SDS-PAGE revealed protein bands (fig. 4-4) corresponding to untagged LoaP (21.0 kDa) and tagged LoaP (32.7 kDa). Both bands were detected in flowthrough and elution fractions indicating that these proteins are present in the mother liquor, and inside the crystals. Since the purity of purified LoaP proteins was estimated by ESI-MS around 97%, we concluded that the observed contaminants are most likely overrepresented as a result of the exceptional sensitivity of the silver-staining method.

The same fractions were then analyzed on 15% 19:1 PAGE, which was stained with ethidium bromide for nucleic acid detection. The results revealed a single band in the elution fraction which migrated on the gel for the same distance as *dfn* RNA standard. Moreover, flowthrough and wash fractions revealed some RNA degradation products. This is most likely due to contaminating RNases which co-purified with LoaP proteins. Free RNA molecules in solution were partially degraded while crystallized RNA molecules were inaccessible to RNases, and therefore protected.

LoaP was labeled with L-selenomethionine to obtain macromolecular phases

To obtain macromolecular phases, we opted to uniformly label LoaP protein with L-selenomethionine (L-SeMet) as all our previous attempt at labeling RNA

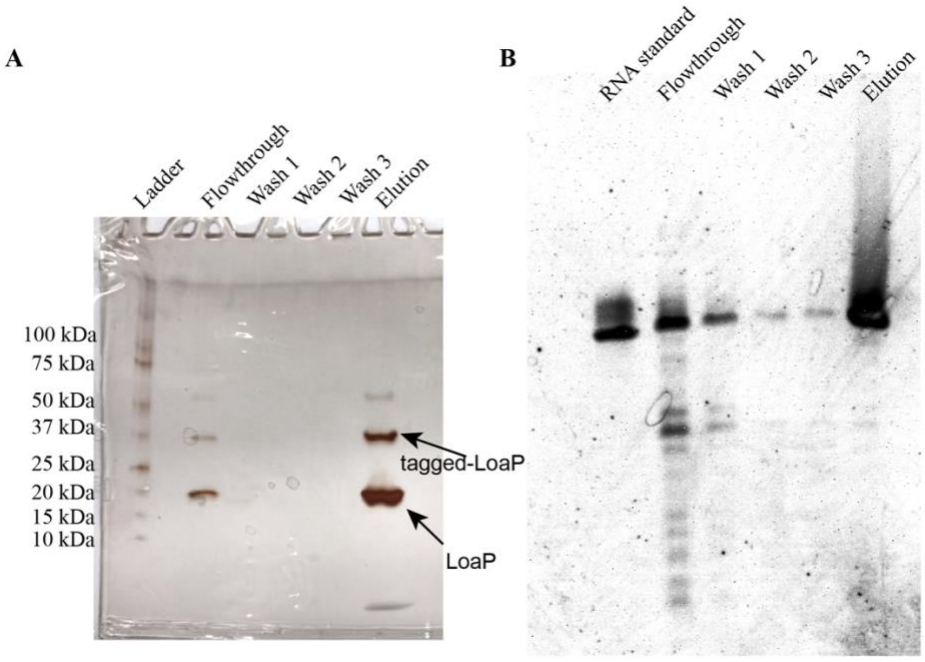
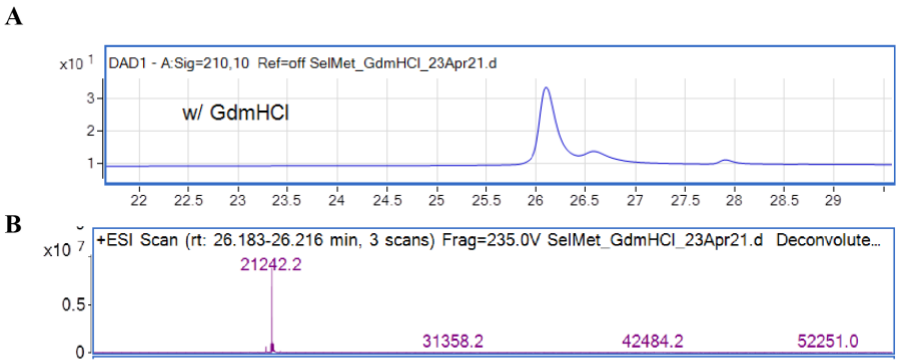


Figure 4-4. Analysis of crystal composition by PAGE reveals protein and RNA content. Crystals which were harvested from LoaP-RNA crystal screen were transferred to 0.1 mm spin filters and washed using ice-cold crystallization buffer to get rid of mother liquor. After several washing steps, crystals were then dissolved in 10 mM EDTA pH 8.0 by heating at 37°C for 10 minutes. Fractions collected from wash and elution steps were analyzed on 4-20% SDS-PAGE (A), and 15% (19:1) acrylamide:bisacrylamide PAGE (B). (A) SDS-PAGE gel stained with silver-staining technique to visualize proteins. (B) 15% PAGE gel stained with ethidium bromide to visualize RNA.

molecules with 5-Bromouridine failed to produce crystals. Typically, incorporation of SeMet is achieved by growing *E. coli* methionine auxotroph strain (B834) in minimal media supplemented with amino acids and L-SeMet. This approach is very successful, often resulting in close to 100% SeMet incorporation. However, a major disadvantage is that protein yield is typically <20% of that of the native protein (184), and it could take up to 24 hours for cultures to reach mid-log phase.

To circumvent these issues, we decided to use an alternative strategy to incorporate SeMet using non-auxotroph stains, albeit with less incorporation efficiency. This was a reasonable compromise since LoaP is a relatively small protein (182 amino acids) but contains 5 methionine residues; therefore, 100% SeMet incorporation is not required. Briefly, *E. coli* BL21 strains carrying an IPTG inducible copy of *T. pseudethanolicus* LoaP were grown in M9 minimal media (see Materials and Methods section for M9 recipe) until OD_{600nm} reached 0.5-0.6, at which point 100 mg/L each of phenylalanine, threonine, lysine, and DL-selenomethionine, and 50 mg/L each of valine, isoleucine, leucine were added directly to cultures in powder form and incubated shaking at 37°C for 15 minutes. The addition of these amino acids shortly before IPTG induction inhibits methionine biosynthesis pathways. In this case, L-selenomethionine supplemented in the media acts as a substitute for methionine and is incorporated during protein synthesis.

Following induction with IPTG, culture growth and protein purification were carried out as previously described in Chapter 1, except that DTT concentration was maintained at 5 mM throughout the purification process as SeMet is sensitive to oxidation. SeMet labeling efficiency was determined using ESI-MS which revealed a



Response Units (%) vs. Acquisition Time (min)

$$MW\ of\ native\ LoaP = 21006.94 \frac{g}{mol}$$

$$Calculated\ MW\ of\ 100\%\ SeMet - labeled\ LoaP$$

$$= 21006.94 - 5(32.06)\ Sulfur + 5(78.96)\ Selenium$$

$$= 21241.4\ g/mol$$

Figure 4-5. L-selenomethionine labeling strategy yielded 100% labeled protein. (A) UV chromatogram showing elution of LoaP protein from C18 using 0-70% Acetonitrile linear gradient at 2% per min. (B) Deconvoluted mass spectra of eluted LoaP peak analyzed on Agilent 6224 TOF-LC/MS.

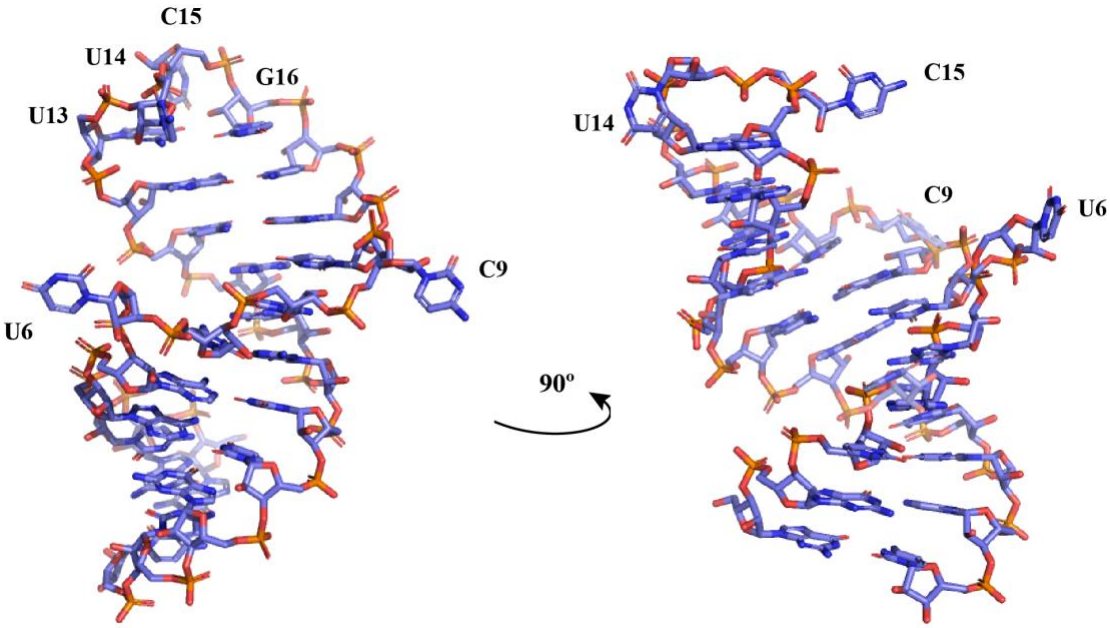


Figure 4-6. The crystal structure of the *dfn* RNA from an initial model. Crystal structure of the *dfn* RNA hairpin in complex with LoaP is shown from two different angles. In this model, *dfn* RNA adopts a UNCG-type hairpin structure in which residues U13-G16 form the UUCG tetraloop. Two RNA residues in the stem (U6 and C9) do not form base pair interactions and are flipped outwards.

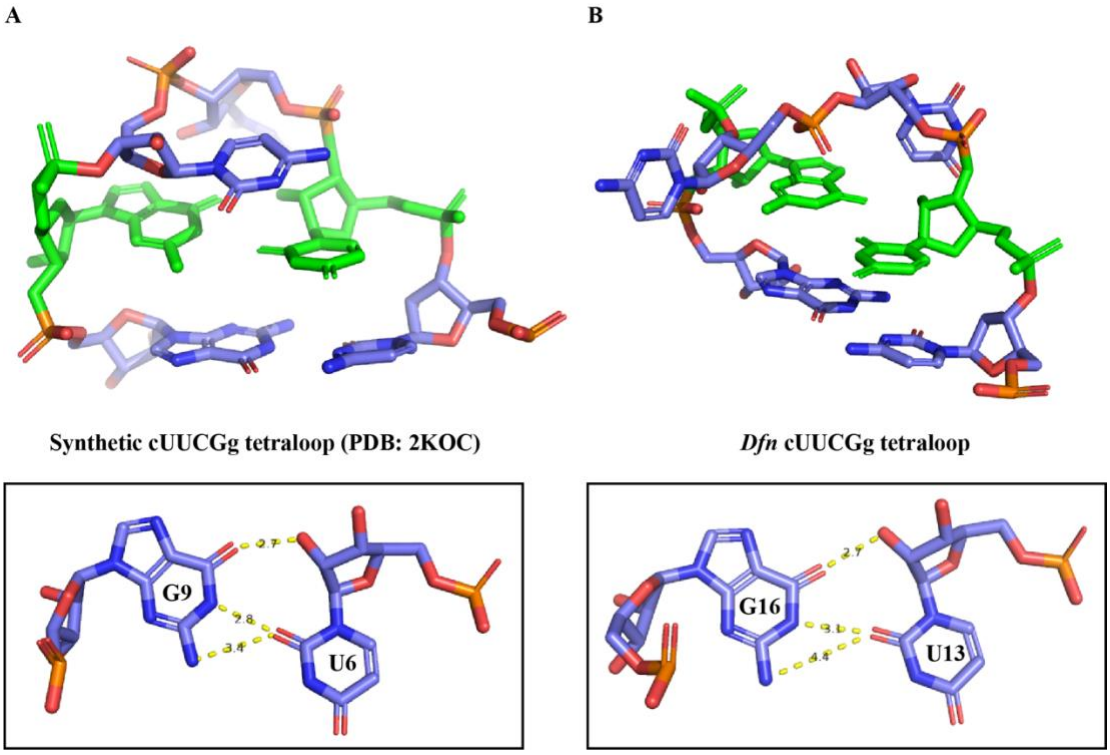


Figure 4-7. A comparison between the structure of the *dfn* tetraloop versus an NMR structure of cUUCGg tetraloop from a synthetic RNA. *Top:* cartoon representation of the three-dimensional structure of a UNCG-type tetraloop from the NMR solution structure of a synthetic RNA (left, PDB: 2KOC) and from the crystal structure of the *dfn* RNA (right). U1 and G4 residues in the tetraloops which form a non-canonical wobble base pair are colored in green. *Bottom:* hydrogen bonding interactions between the first and last base pairs in the UUCG tetraloops.

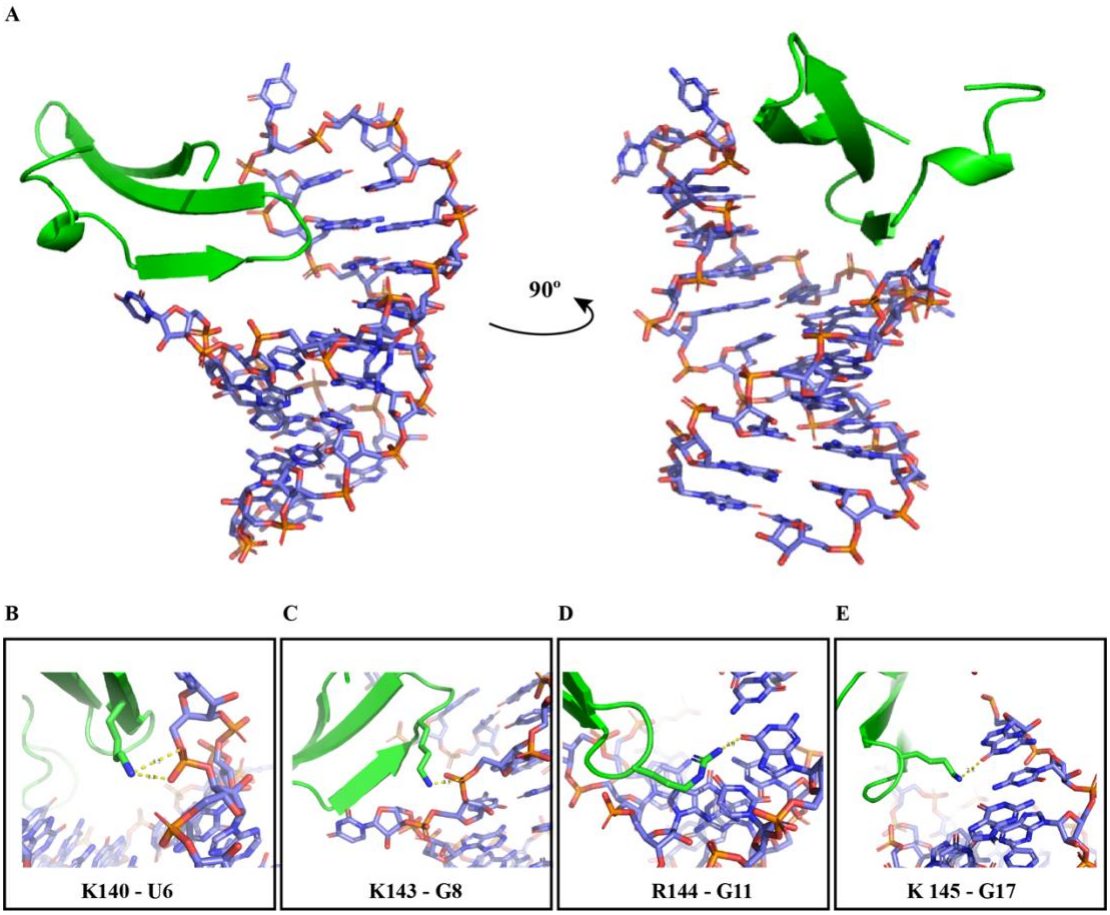


Figure 4-8. Structural representation of the interactions formed between the *dfn* RNA and residues within the LoaP CTD domain. (A) Cartoon representation of the three-dimensional structure of the *dfn* RNA hairpin and protein residues within the LoaP CTD domain viewed from two different angles. (B-E) Hydrogen bonding interactions observed in the initial model between conserved residues within the LoaP CTD domain and the RNA hairpin.

single mass peak in the purified protein stock corresponding to the mass of 100% SeMet-labeled protein (21242.2 g/mol) (fig. 4-5).

Discussion

As discussed in Chapter 2, LoaP proteins are broadly characterized by a preponderance of positively-charged arginine (R) or lysine (K) residues relative to core NusG, which instead typically contains negatively charged residues at the corresponding positions. Our biochemical data indicate that LoaP proteins exhibit an unprecedented capability within the NusG family to bind a small RNA hairpin with high affinity and specificity. This RNA hairpin is commonly found in leader regions of operons which are regulated by LoaP factors, suggesting a functional importance of the LoaP-RNA interaction. Furthermore, binding to the RNA hairpin is mediated by a few basic residues localized within the flexibly-linked CTD. Interestingly, manual inspection of this region reveals striking similarities with a known RNA-binding motif called Arginine-Rich Motif.

Arginine-rich motifs (ARM) are employed by a variety of RNA-binding proteins (170, 171, 185) to recognize cognate RNA hairpins. While few determinants are explicitly diagnostic of these motifs, they are generally characterized by multiple basic residues (185) that are proximally clustered (~8-20 residues). It is thought that these motifs have arisen independently throughout phylogeny, and function as independent domains, separate from the proteins in which they are found (171). As such, they have proved to be excellent models for structural studies of protein-RNA complexes (186). Common examples of ARM peptides include the human

immunodeficiency virus (HIV) Rev (187), HIV Tat (188), bovine immunodeficiency virus (BIV) Tat (189), and phage λN antiterminator (35), all of which bind small cognate RNA hairpins with high affinity and specificity. In fact, λN peptide binds its cognate RNA hairpin *boxB* 16-fold better than related phage P22 *boxB* RNA (171).

Structural characterization (98, 190, 191) of ARM peptides indicate that they are intrinsically flexible, but they have a tendency to bind specifically to their cognate RNA hairpins in an α-helical conformation. Moreover, these peptides preferentially associate with RNA structures containing characteristic wide major grooves (192), typically found near loops or bulged regions. For example, N peptide, which is required for the assembly of λN transcription antitermination complex, is structurally disordered in its free form (98), but binding to *boxB* RNA induces a conformational change to an extended α-helical conformation. A few exceptions to this trend were found in BIV Tat ARM (193), which assumes a β-hairpin structure, and the Jembrana disease virus (JDV) Tat protein, which binds both BIV and HIV-1 TAR RNAs in a β-hairpin structure, or an extended α-helical conformation, respectively (194). Taken together, these structural studies seem to suggest that in addition to their structural flexibility, the folds of ARM peptides could be dictated by binding to the cognate RNA hairpins.

To gain insights into the LoaP-RNA ribonucleoprotein complex, we attempted to solve the complex structure by X-ray crystallography. No crystals were observed for LoaP protein alone, and only two conditions (out of 950 total conditions) yielded crystals for the LoaP:RNA complex at 1:1 molar ratio (fig. 4-3). Analysis by PAGE revealed that those crystals included protein and RNA components (fig. 4-4), the

identity of which was confirmed by comparison to LoaP and RNA standards by PAGE.

Electron density maps for the complex (see Table 1 for data collection and statistics) were generated by the single-wavelength anomalous dispersion (SAD) method using data from L-selenomethionine labeled LoaP-RNA complex. The crystallographic asymmetric unit (A.U.) contains three similar RNA molecules and two protein molecules. The RNA models were manually built in Coot and refined in Phenix.refine program. In this initial model, the *dfn* RNA molecules adopt a canonical UNCG-type fold (fig. 4-6). UNCG tetraloops comprise one of the most abundant classes of RNA tetraloops – the other class being GNRA tetraloops –, while other folds are typically less abundant and widespread in the database (162, 195). Structural characterization of UNCG tetraloops by NMR spectroscopy and X-ray crystallography (195-199) revealed that UNCG and GNRA tetraloops exhibit exceptional thermodynamic stability conferred by stacking interactions, specific non-Watson and Crick base pairing, and distinct backbone conformations.

The most diagnostic structural feature of UUCG tetraloops is the formation of a *trans*-wobble G-U base pair (198) between the first and last residues in the tetraloop, while the other two resides are unpaired. This interaction is further stabilized by a hydrogen bond interaction between the ribose 2'-hydroxyl group (2'OH) in U1 and the carbonyl oxygen in G4(O6). Indeed, the replacement of U1 in the UUCG tetraloops by its deoxyribose analog (200, 201) strongly destabilizes the loop, while the replacement of the ribose sugar of residue U2 results in marginal destabilization effects.

The initial model of the *dfn* RNA hairpin reveal that the first and last residues (U13 and G16) in the tetraloop are facing inwards, stacking on top of the C-G closing Watson-Crick base pair, while the second and third base are unpaired (fig. 4-7B). Compared to a high-resolution NMR structure (202) of a UUCG tetraloop from the database (fig. 4-7A, PDB: 2KOC), the *dfn* tetraloop appears to adopt an identical conformation, consistent with the interactions which are typically observed in UUCG tetraloops. In particular, the non-canonical U-G base pair is clearly seen in a conformation that is identical to other UUCG tetraloops (198, 199, 203, 204) in which U13(O2) is the acceptor group for N1 and N2 Watson-Crick positions of G16. Additionally, the 2'OH group of U13 is the donor group in a hydrogen bond with G16(O6) (figure 4-7B, bottom panel). The *dfn* RNA stem adopts a canonical helix conformation (fig. 4-6) in which two single nucleotides (U5 and C9) are unpaired and are flipped out towards the solvent.

We were unable to solve the structure of LoaP using molecular replacement in Phenix software using the published crystal structures of NusG and RfaH proteins (PDB: 5TBZ and 2OUG, respectively) as search models. It is likely that LoaP exhibits unique structural features which are distinct from core NusG and NusG-paralog RfaH, given its unique RNA-binding activity. Currently, we obtained a structure model containing 30 amino acid residues within the LoaP CTD. This model was manually built in Coot, refined in Phenix.refine, and used as a search model using the program Phaser in Phenix. The top solution had a TFZ score of 28.2.

Interestingly, the structure of these amino acids residues suggests that LoaP CTD adopts a three-stranded β-sheet conformation which is typical of the CTD of

core NusG proteins. Intriguingly, a β-hairpin structure within the CTD is situated in an edge-on position in the major groove of the RNA (fig. 4-8A), where it is almost completely buried. The tight fit potentially results from the widening of the major groove of the RNA due to the presence of the two unpaired residues. Surprisingly, these unpaired nucleotides (U6 and C9) do not appear to form any discernable interactions with amino acid residues within this portion of the CTD. Instead, several conserved residues located within the flexible loop connecting the β-strands form direct interactions with two RNA bases (G11 and G17) and the RNA backbone. In particular, residues K140 and K143 appear to interact directly with the non-bridging oxygens in the phosphate backbone (fig. 4-8 A and B), while residues R144 and K145 form hydrogen bond interactions with two base-paired residues in the RNA stem: G11 and G17, respectively (fig. 4-8 C and D). These specific interactions are mediated by the guanidino group in R144 and G11(O6), and the amino group in K145 and G17(O6).

These observations from the crystal structure are consistent with our prior biochemical analysis which revealed that LoaP RNA-binding activity was significantly destabilized when residues K148 and R149 in *B. velezensis* LoaP, (residues K143 and R144 in *T. pseud* LoaP, respectively) were mutated to alanine, indicating that they are required for binding RNA (fig. 3-7). A less destabilizing effect was observed when residue K145 (K140 in *T. pseud* LoaP) was mutated, suggesting that this residue plays a secondary role in the LoaP-RNA interaction. This is consistent with the observation that K140 in the crystal structure interacts with the

non-bridging oxygens in the phosphate backbone, and potentially stabilizes the complex.

It is still not clear whether additional interactions are required for the formation of LoaP-RNA complex as model building is still underway. However, an additional residue K150 (K145 in *T. pseud* LoaP) appears to interact specifically with G17(O6) and potentially plays a key role in the binding event. This amino acid residue was not included in our prior alanine scanning mutagenesis studies, and therefore its functional role is yet to be determined.

Another interesting observation from the crystal structure is the absence of binding interactions between the unpaired nucleotides (U6 and C9) and amino acids residues within this loop. Equilibrium binding data (see fig. 3-4) strongly argue that these bulged residues are required for binding LoaP as their deletion resulted in complete loss of binding interactions. This observation is further supported by competition binding assays in which *dfn* RNA hairpin containing only one bulged residue, either U5 or C9, moderately competed with wildtype RNA hairpin for binding LoaP, but were unable to completely displace bound wildtype hairpin from LoaP-RNA complex (fig 3-5). A reasonable explanation is that the structural importance of these bulged residues arise from the widening of the major groove in the RNA hairpin, which is necessary to accommodate the β-hairpin structure. In the absence of the bulged residues, the RNA major groove is perhaps be too narrow to tightly fit the β-hairpin and thereby preclude the formation of important LoaP-RNA interactions.

In conclusion, the structural data presented here are so far in good agreement with our prior biochemical findings on the specific interaction between LoaP and the cognate RNA hairpin. This ribonucleoprotein complex is a key part of an overall genetic regulatory mechanism, which presently remains elusive. Solving the structure of LoaP-RNA complex would constitute a major progress in understanding the mechanistic role of LoaP antiterminator as well as expand the mechanistic diversity within NusG family of regulators.

Materials and Methods

M9 minimal media

M9 minimal media was prepared by dissolving 6g Na_2HPO_4, 3.0 g KH_2PO_4, 0.5 g NaCl in 1 L deionized water. The solutions were then autoclaved for 1 hour and stored at room temperature. Prior to inoculating the M9 minima media, the following reagents were added: 1 mM $MgSO_4$, 10 μg/mL thiamine, 10 μM $CaCl_2$, 1.0 g NH_4Cl, and 4.0 g glucose.

RNA preparation and purification

RNA molecules used in the preparation of LoaP-RNA complex were transcribed *in vitro* using T7 RNA-polymerase and purified by electrophoresis on 12% polyacrylamide (19:1 acrylamide/bisacrylamide), 1x TBE, 8 M Urea gels; electroeluted from gel slices; washed twice with 1 M NaCl; desalted by ultrafiltration, filtered (0.1 μm cutoff, Amicon Ultrafree-MC, Millipore), and stored at -20℃.

Figure 4-9. Sequence alignment of LoaP protein used in the crystallization of LoaP-RNA complex vs core NusG proteins from the PDB database. Highly conserved residues are highlighted in green. Methionine positions which were labeled with L-selenomethionine are highlighted in purple.

	SeMet_LoaP-RNA complex
Resolution range (Å)	48.28 - 2.45(2.538 - 2.45)
Space group	*C 1 2 1*
Cell dimensions	
a, b, c (Å)	68.951, 87.451, 148.793
α, β, γ (°)	90, 103.24, 90
Total reflections	62684 (5748)
Unique reflections	31602 (3030)
Multiplicity	2.0 (1.9)
Completeness (%)	91.71 (71.67)
<I> / <σ(I)>	14.45 (1.04)
Wilson B-factor (Å^2)	49.71
R-merge	0.05584 (0.7893)
R-meas	0.07896 (1.116)
R-pim	0.05584 (0.7893)
CC1/2	0.997 (0.413)
*CC**	0.999 (0.764)
Reflections used in refinement	29130 (2282)
Reflections used for R-free	1857 (155)
R-work	0.4468 (0.5188)
R-free	0.4416 (0.5065)
CC(work)	0.723 (0.258)
CC(free)	0.884 (0.328)
Number of non-hydrogen atoms	3195
macromolecules	3195
ligands	0
solvent	0
Protein residues	188
*RMS (bonds) (*Å)	0.011
RMS (angles) (°)	1.48
Ramachandran favored (%)	87.78
Ramachandran allowed (%)	8.89
Ramachandran outliers (%)	3.33
Rotamer outliers (%)	8.33
Clashscore	13.08
Average B-factor (Å^2)	39.94
macromolecules	26.77

*Statistics for the highest-resolution shell are shown in parentheses.

Table 1. Data collection and refinement statistics.

Crystallization and diffraction data collection

Dfn RNA was thawed on ice from -20℃ stock and diluted in binding buffer (10 mM Tris.HCl pH 7.2, 100 mM NaCl, 0.5 mM EDTA, 1 mM DTT). The diluted RNA solution (300 μM) was heated at 92℃ for 3 minutes and then immediately transferred to an ice-bath for 10 minutes. *T. pseud* LoaP samples were thawed from -80℃ stocks and loaded onto Superdex 75 increase 10/300 GL size-exclusion column, which was pre-equilibrated in binding buffer. Fractions containing LoaP were collected and concentrated to ~500 μM (3 kDA MWCO, Amicon Ultra Centrifugal filter, Millipore). LoaP-RNA complex was prepared by mixing 100 μM purified *dfn* RNA with 100 μM *T. pseud* LoaP (1:1 molar ratio) in binding buffer, and incubated at room temperature for 45 minutes. For crystallization, 0.4 μL of LoaP-RNA complex (100 μM) and 0.4 μL of reservoir solution (0.250 M Na Acetate pH 5.2, 12% (w/v) PEG 3350) were mixed and equilibrated at 294 K by sitting drop vapor diffusion. Tetragonal pyramidal crystals grew in 5-8 days to maximum dimensions of 200 x 50 x 50 μm^3. Cryoprotection was performed by adding 2 μL 30% (w/v) PEG 400 directly to the mother liquor and incubated for 5 minutes at room temperature. Crystals were then mounted in a nylon loop and vitrified by plunging into liquid nitrogen. Data were collected at 100K at ALS beamline 501 using 1.105 Å X-radiation. Data were reduced in HKL2000 (205) with 10% of reflections flagged for R_{Free} calculations.

Structure determination and refinement

SeMet-labeled LoaP-RNA complex for phasing was crystalized under the same conditions. Crystals were soaked in the cryoprotectant conditions described above

and vitrified by plunging in liquid nitrogen. Data sets reporting a significant anomalous signal by XDS were further analyzed for phasing. SHELXC (206) reported significant anomalous signal extending to 2.45 Å from a single crystal. Heavy atom sites from SHELXD (206) were refined using AutoSol (207) yielding a mean overall figure of merit of 0.229. Density modification using Resolve (208) resulted in an electron density map into which 26 nucleotides of RNA could be built manually using Coot. This model was then refined using Phenix.refine and molecular replaced into the data set using the program Phaser (209). The top solution had TFZ score of 16.1. Manual building and refinement was performed in Coot and Phenix.refine, respectively. HL coefficients were used as a target during refinement.

Chapter 5: Conclusions and perspectives

LoaP-mediated antitermination mechanism

Transcription-translation coupling has been considered a general hallmark of gene processing pathways in bacteria. In this context, several transcription factors exert genetic regulatory effects by facilitating the formation of this linkage to ensure uninterrupted transcription-translation output. For example, NusG paralog RfaH prevents Rho-dependent termination (134, 135, 142) by acting as a molecular tether bridging the transcription apparatus to the leading ribosome. However, RfaH does not promote antitermination of intrinsic terminators, and therefore lacks the ability to act as a processive antitermination factor.

A recent study (159) presented evidence arguing that the interplay between transcription and translation in some bacterial species could be stinkingly different from the currently established paradigm. In particular, it was shown that in the Gram-positive bacterium *B. subtilis*, the transcription apparatus do not routinely associate with the leading ribosome resulting in a 'runaway transcription' in which the rate of transcription exceeds the rate of protein synthesis. Interestingly, these findings are consistent with the observed diminished role of Rho termination in *B. subtilis*, and more importantly, the prevalence of cis-acting regulatory elements in leader regions (*i.e.* riboswitches).

Our lab recently discovered a novel subfamily of NusG paralogs in *B. velezensis* known as LoaP which is primarily encoded by Gram-positive bacteria, and

is found in several bacterial phyla including Firmicutes, Actinobacteria, and Spirochaetes (131). LoaP is often found adjacent to exceptionally long biosynthetic gene clusters encoding for secondary metabolites. Furthermore, LoaP has been shown to promote processive antitermination of intrinsic terminators located within 5' UTR or interspersed in the intergenic regions hinting at a potential genetic regulatory mechanism governing their gene expression.

B. subtilis and *B. velezensis* are very closely related organisms (131, 160); therefore, analysis of LoaP antitermination is likely to demonstrate how NusG paralogs can regulate transcription elongation in bacteria that do not couple transcription and translation. In particular, the absence of a coupled ribosome during transcription elongation raises the possibility that Rho terminator could be dispensable in some termination/antitermination complexes. Instead, transcription attenuation in these organisms is achieved through interactions with nascent RNA and specific RNA-binding regulators.

Processive antitermination mechanisms often involve specific interactions between regulators and nucleic acids. This is typically a critical event in the assembly of antitermination complexes, as these processes often control gene expression by "action at a distance" effects (97) in which specific regulators associate with RNAP at one site and modulate transcription at another (*i.e.* λN antitermination complex). We speculate that NusG paralog LoaP functions in a manner similar to N peptide; however, in this scenario, binding interactions with some Nus factors (*e.g.* NusG, NusE, and NusB) would not be required. This is because LoaP is capable of binding RNAP via the NTD and simultaneously associate with cognate RNA via the CTD

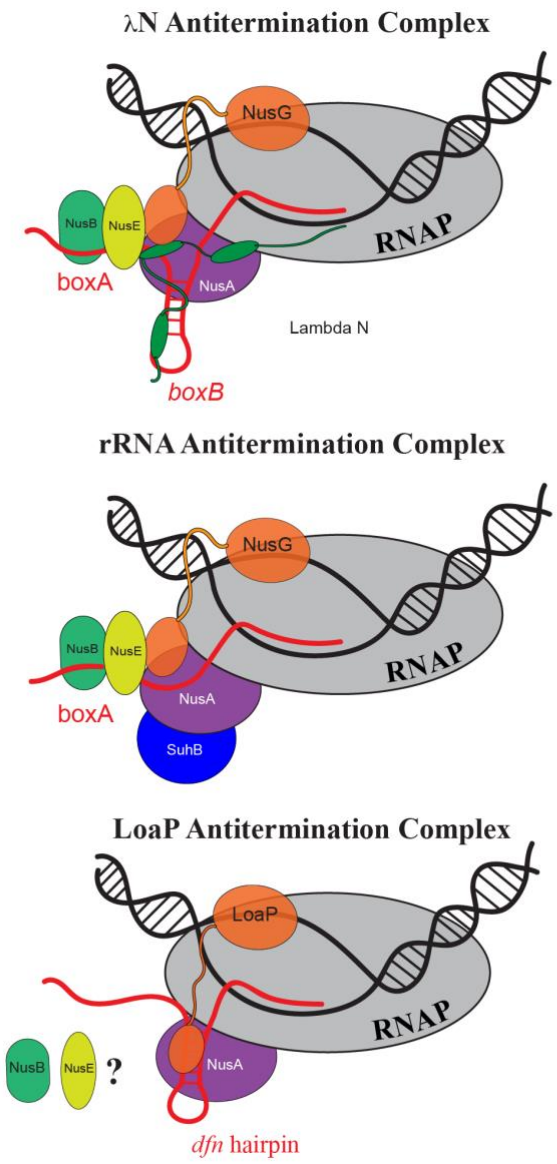


Figure 5-1. Schematic representation of antitermination complexes

thereby forming an RNA loop. It is not yet clear how this RNA-binding activity impacts binding of the CTD with the other cellular factors, but it would not be unexpected if NusA mediates additional contacts to this ribonucleoprotein complex (see fig. 5-1).

Expanding the search for LoaP-regulated operons

As discussed in Introduction chapter, a previous study (148) has shown that the closthioamide (CTA) biosynthetic gene cluster in *R. cellulolyticum* is silent under standard laboratory conditions. However, the overexpression of LoaP (Ccel_0849) in *R. cellulolyticum* from a plasmid under inducible conditions activates the production of Closthioamide and its derivates (210). Although the CTA biosynthetic gene cluster was verified using genome editing approaches, the heterologous expression of the CTA operon in the Gram-negative facultative anaerobe *E. coli* was unsuccessful. The lack of detection of CTA and several precursor intermediates in *E. coli* is indicative of a defective operon – this is typically the case when biosynthetic gene clusters are expressed outside of their native bacterial species. This also suggests that some regulatory genes, which exert regulatory control over the biosynthetic gene cluster, and are absolutely essential for the complete synthesis of closthioamide, are absent both within the CTA operon and in the *E. coli* chromosome.

Interestingly, we discovered that *C. cellulolyticum* LoaP used in this study also exhibited specific binding interactions to *dfn* RNA hairpin (fig. 5-1) indicating that a homologous RNA hairpin is located within the genome and is likely associated with LoaP-mediated antitermination. We hypothesize that LoaP is required for the production of closthioamide and its derivatives in *R. cellulolyticum*.

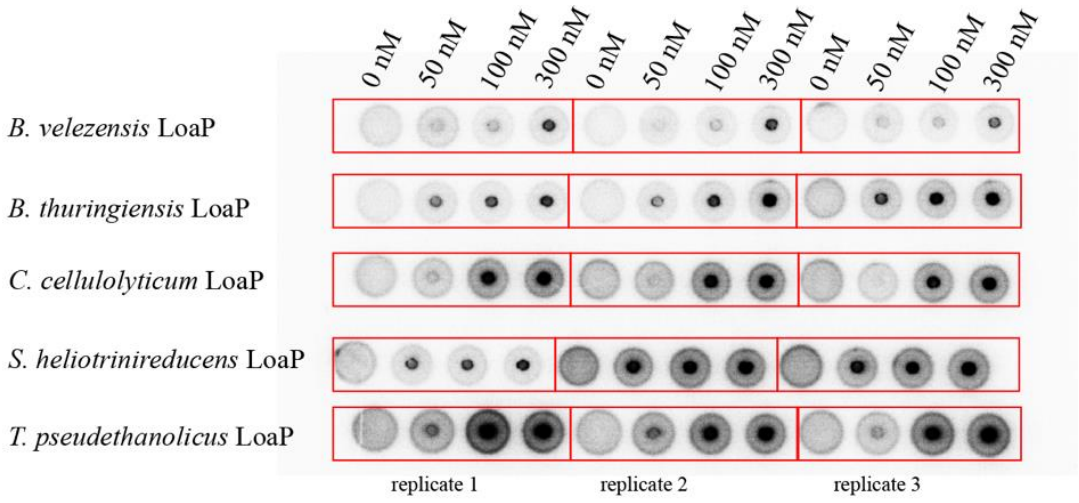


Figure 5-2. Equilibrium binding DRaCALA screen testing RNA-binding activity of LoaP proteins from five different organisms. LoaP proteins at varying concentrations were incubated with radiolabeled *dfn* RNA hairpin at room temperature for 45 minutes. 2 μL aliquots from binding reactions were spotted on nitrocellulose paper and visualized by exposure to phosphor screens for 20 minutes. Dark spots indicate the association of proteins with radiolabeled RNA at any given concentration. Binding reactions were performed in triplicates.

It remains elusive whether LoaP regulates the CTA gene cluster directly, or if it regulates a subset of participatory genes involved in CTA biosynthesis. Therefore, while the mechanistic role of the LoaP-RNA complex still remains to be discovered, our data suggest that there is likely to be considerable diversity in the regulatory mechanisms employed by NusG specialized paralogs. LoaP proteins exhibit unprecedented ability of binding specifically and with high affinity to a cognate UNCG-type RNA hairpin. This unique RNA-binding activity appear to be unique to LoaP subfamily as both core NusG and NusG-paralog RfaH lack the ability to bind RNA. Furthermore, LoaP seems to exhibit functional and structural characteristics that distinguishes it as a specialized NusG paralog operating with a newly uncharacterized mechanism. Understanding these distinct functions of LoaP is important because it expands the mechanistic range of NusG family proteins, and it allows us to gain a wider insight into antibiotic regulation on a genetic level in bacteria.

Appendix

RNA sequences used in the equilibrium binding assays

1. WT *dfn* sequence

5’ – GGAAAGGCAAUCGCUGCUUCGGCACGUUGCC

2. M4 *dfn* sequence

5’ – GGAAAGGCAACGUGCUUCGGCACGUUGCC

3. M10 *dfn* sequence

5’ – GGAAAGGCAAUCGCUGCGAAAGCACGUUGCC

4. *boxB* sequence

5’ – GGAAAGGGCCCUGAAGAAGGGCCC

DNA sequence used in equilibrium binding assays

A. WT dfn sequence

5’ – GGAAAGGCAATCGCTGCTTCGGCACGTTGCC

LoaP protein sequences used in the equilibrium binding assays

A. *B. velezensis* LoaP

AGTMKWYALFVESGKEETVQKFLRLQFDEQALYSIIPKKKVTERKAGIKYEA
LKKMFPGYVLFKTKMTERTFHKIKELPISCRIVNNGAYYSKERKTYFTTIKDE
EILPIIRLIGEGDTVDYSKVYIENSKVTVASGPLKGMEGIIKKIDKRKRRAKICL
SFMGLDKMVNVGIEVLSKP

B. *S. heliotrinireducens* LoaP

AGTMWYVIQVGTNQEDRVIGLIRSFVGKDVLKEAFVPQVEVMRRSRGQWQ
KRKELLLPGYVFVIATDPEKLNQALIDVPAFTRLLGNDVSFTPLLDDEIKFLEA
FTAPDRRIVRMSKGVIEGDQIIINEGPLRGQTGLIKRIDRHKRLAYLEMTVMG
RKKMIKVGLEIVSKS

C. *T. pseudethanolicus* LoaP

AGTMKKWYVIFTRSGYENKVRDIIENCFKEEVKLLIPKRKIIERVKGQPVEKIK
LLFPGYVFVNAEMSDDLYYKISEVLKRGIFLKEGKRPAFVKEEEMKIILSLTK
NSDLIDLSKGIMEGERVKIIEGPLKGYEGLIKKIDKRKKRAKVIFSIAGELKSVD
LAIEVMENVSEQQRSLVYAC

DNA sequences used for *in vitro* transcription of *dfn* RNA

A. Hammerhead construct fused to wildtype *dfn* RNA

5’ – TAATACGACTCACTATA GGG AGA CAG CGA TTG CC CTG ATG AGT CCG TGA GGA CGA AAC GGT ACC CGG TAC CGT C**GG CAA TCG CTG CTT CGG CAC GTT GCC**

*The underlined sequence is the T7 RNA-polymerase promoter.
**Hammerhead sequence is colored in blue.
***_dfn_ RNA sequence is shown in bold font.

Bibliography

1. A. L. Demain, S. Sanchez, Microbial drug discovery: 80 years of progress. *J Antibiot (Tokyo)* **62**, 5-16 (2009).
2. G. Minotti, P. Menna, E. Salvatorelli, G. Cairo, L. Gianni, Anthracyclines: molecular advances and pharmacologic developments in antitumor activity and cardiotoxicity. *Pharmacol Rev* **56**, 185-229 (2004).
3. J. F. Borel, History of the discovery of cyclosporin and of its early pharmacological development. *Wien Klin Wochenschr* **114**, 433-437 (2002).
4. F. Ikeda *et al.*, Role of micafungin in the antifungal armamentarium. *Curr Med Chem* **14**, 1263-1275 (2007).
5. B. Ruiz *et al.*, Production of microbial secondary metabolites: regulation by the carbon source. *Crit Rev Microbiol* **36**, 146-167 (2010).
6. M. C. Wu, B. Law, B. Wilkinson, J. Micklefield, Bioengineering natural product biosynthetic pathways for therapeutic applications. *Curr Opin Biotechnol* **23**, 931-940 (2012).
7. B. Wilkinson, J. Micklefield, Mining and engineering natural-product biosynthetic pathways. *Nat Chem Biol* **3**, 379-386 (2007).
8. M. A. Fischbach, C. T. Walsh, Antibiotics for emerging pathogens. *Science* **325**, 1089-1093 (2009).
9. M. Zerikly, G. L. Challis, Strategies for the discovery of new natural products by genome mining. *Chembiochem* **10**, 625-633 (2009).
10. M. L. Metzker, Sequencing technologies - the next generation. *Nat Rev Genet* **11**, 31-46 (2010).
11. D. F. Browning, S. J. Busby, Local and global regulation of transcription initiation in bacteria. *Nat Rev Microbiol* **14**, 638-650 (2016).
12. C. Mejía-Almonte *et al.*, Redefining fundamental concepts of transcription initiation in bacteria. *Nat Rev Genet* **21**, 699-714 (2020).
13. R. Tollerson, M. Ibba, Translational regulation of environmental adaptation in bacteria. *J Biol Chem* **295**, 10434-10445 (2020).
14. G. A. Belogurov, I. Artsimovitch, The Mechanisms of Substrate Selection, Catalysis, and Translocation by the Elongating RNA Polymerase. *J Mol Biol* **431**, 3975-4006 (2019).
15. J. D. Helmann, M. J. Chamberlin, Structure and function of bacterial sigma factors. *Annu Rev Biochem* **57**, 839-872 (1988).
16. E. F. Ruff, M. T. Record, I. Artsimovitch, Initial events in bacterial transcription initiation. *Biomolecules* **5**, 1035-1062 (2015).
17. K. S. Murakami, S. A. Darst, Bacterial RNA polymerases: the wholo story. *Curr Opin Struct Biol* **13**, 31-39 (2003).
18. G. A. Belogurov, I. Artsimovitch, Regulation of Transcript Elongation. *Annu Rev Microbiol* **69**, 49-69 (2015).
19. A. Mustaev, J. Roberts, M. Gottesman, Transcription elongation. *Transcription* **8**, 150-161 (2017).
20. E. Nudler, RNA polymerase active center: the molecular engine of transcription. *Annu Rev Biochem* **78**, 335-361 (2009).

21. F. Blombach *et al.*, Archaeology of RNA polymerase: factor swapping during the transcription cycle. *Biochem Soc Trans* **41**, 362-367 (2013).
22. A. Feklístov, B. D. Sharon, S. A. Darst, C. A. Gross, Bacterial sigma factors: a historical, structural, and genomic perspective. *Annu Rev Microbiol* **68**, 357-376 (2014).
23. K. S. Murakami, X-ray crystal structure of Escherichia coli RNA polymerase σ70 holoenzyme. *J Biol Chem* **288**, 9126-9134 (2013).
24. V. Mekler *et al.*, Structural organization of bacterial RNA polymerase holoenzyme and the RNA polymerase-promoter open complex. *Cell* **108**, 599-614 (2002).
25. B. Bae *et al.*, Phage T7 Gp2 inhibition of Escherichia coli RNA polymerase involves misappropriation of σ70 domain 1.1. *Proc Natl Acad Sci U S A* **110**, 19772-19777 (2013).
26. G. Lloyd, P. Landini, S. Busby, Activation and repression of transcription initiation in bacteria. *Essays Biochem* **37**, 17-31 (2001).
27. S. J. W. Busby, Transcription activation in bacteria: ancient and modern. *Microbiology (Reading)* **165**, 386-395 (2019).
28. R. R. Breaker, Prospects for riboswitch discovery and analysis. *Mol Cell* **43**, 867-879 (2011).
29. S. Proshkin, A. Mironov, E. Nudler, Riboswitches in regulation of Rho-dependent transcription termination. *Biochim Biophys Acta* **1839**, 974-977 (2014).
30. C. L. Turnbough, Regulation of Bacterial Gene Expression by Transcription Attenuation. *Microbiol Mol Biol Rev* **83** (2019).
31. M. Varon, N. Fuchs, M. Monosov, S. Tolchinsky, E. Rosenberg, Mutation and mapping of genes involved in production of the antibiotic TA in Myxococcus xanthus. *Antimicrob Agents Chemother* **36**, 2316-2321 (1992).
32. V. Simunovic *et al.*, Myxovirescin A biosynthesis is directed by hybrid polyketide synthases/nonribosomal peptide synthetase, 3-hydroxy-3-methylglutaryl-CoA synthases, and trans-acting acyltransferases. *Chembiochem* **7**, 1206-1220 (2006).
33. M. Chatzidaki-Livanis, K. G. Weinacht, L. E. Comstock, Trans locus inhibitors limit concomitant polysaccharide synthesis in the human gut symbiont Bacteroides fragilis. *Proc Natl Acad Sci U S A* **107**, 11976-11980 (2010).
34. M. Chatzidaki-Livanis, M. J. Coyne, L. E. Comstock, A family of transcriptional antitermination factors necessary for synthesis of the capsular polysaccharides of Bacteroides fragilis. *J Bacteriol* **191**, 7288-7295 (2009).
35. W. A. Rees, S. E. Weitzel, T. D. Yager, A. Das, P. H. von Hippel, Bacteriophage lambda N protein alone can induce transcription antitermination in vitro. *Proc Natl Acad Sci U S A* **93**, 342-346 (1996).
36. T. J. Santangelo, I. Artsimovitch, Termination and antitermination: RNA polymerase runs a stop sign. *Nat Rev Microbiol* **9**, 319-329 (2011).
37. R. A. Mooney, I. Artsimovitch, R. Landick, Information processing by RNA polymerase: recognition of regulatory signals during RNA chain elongation. *J Bacteriol* **180**, 3265-3275 (1998).

38. K. S. Murakami, S. Masuda, E. A. Campbell, O. Muzzin, S. A. Darst, Structural basis of transcription initiation: an RNA polymerase holoenzyme-DNA complex. *Science* **296**, 1285-1290 (2002).
39. D. G. Vassylyev *et al.*, Crystal structure of a bacterial RNA polymerase holoenzyme at 2.6 A resolution. *Nature* **417**, 712-719 (2002).
40. B. Krummel, M. J. Chamberlin, Structural analysis of ternary complexes of Escherichia coli RNA polymerase. Deoxyribonuclease I footprinting of defined complexes. *J Mol Biol* **225**, 239-250 (1992).
41. N. V. Vo, L. M. Hsu, C. M. Kane, M. J. Chamberlin, In vitro studies of transcript initiation by Escherichia coli RNA polymerase. 3. Influences of individual DNA elements within the promoter recognition region on abortive initiation and promoter escape. *Biochemistry* **42**, 3798-3811 (2003).
42. H. Yin *et al.*, Transcription against an applied force. *Science* **270**, 1653-1657 (1995).
43. M. A. Grachev *et al.*, Oligonucleotides complementary to a promoter over the region -8...+2 as transcription primers for E. coli RNA polymerase. *Nucleic Acids Res* **12**, 8509-8524 (1984).
44. A. Ray-Soni, M. J. Bellecourt, R. Landick, Mechanisms of Bacterial Transcription Termination: All Good Things Must End. *Annu Rev Biochem* **85**, 319-347 (2016).
45. J. W. Roberts, Mechanisms of Bacterial Transcription Termination. *J Mol Biol* **431**, 4030-4039 (2019).
46. P. Mitra, G. Ghosh, M. Hafeezunnisa, R. Sen, Rho Protein: Roles and Mechanisms. *Annu Rev Microbiol* **71**, 687-709 (2017).
47. W. M. Holmes, T. Platt, M. Rosenberg, Termination of transcription in E. coli. *Cell* **32**, 1029-1032 (1983).
48. V. Epshtein, D. Dutta, J. Wade, E. Nudler, An allosteric mechanism of Rho-dependent transcription termination. *Nature* **463**, 245-249 (2010).
49. E. Skordalakes, J. M. Berger, Structural insights into RNA-dependent ring closure and ATPase activation by the Rho termination factor. *Cell* **127**, 553-564 (2006).
50. M. S. Ciampi, Rho-dependent terminators and transcription termination. *Microbiology (Reading)* **152**, 2515-2528 (2006).
51. J. W. Roberts, S. Shankar, J. J. Filter, RNA polymerase elongation factors. *Annu Rev Microbiol* **62**, 211-233 (2008).
52. C. J. Cardinale *et al.*, Termination factor Rho and its cofactors NusA and NusG silence foreign DNA in E. coli. *Science* **320**, 935-938 (2008).
53. C. M. Burns, W. L. Nowatzke, J. P. Richardson, Activation of Rho-dependent transcription termination by NusG. Dependence on terminator location and acceleration of RNA release. *J Biol Chem* **274**, 5245-5251 (1999).
54. Z. Pasman, P. H. von Hippel, Regulation of rho-dependent transcription termination by NusG is specific to the Escherichia coli elongation complex. *Biochemistry* **39**, 5573-5585 (2000).
55. J. M. Peters, A. D. Vangeloff, R. Landick, Bacterial transcription terminators: the RNA 3'-end chronicles. *J Mol Biol* **412**, 793-813 (2011).

56. X. Guo *et al.*, Structural Basis for NusA Stabilized Transcriptional Pausing. *Mol Cell* **69**, 816-827.e814 (2018).
57. S. Mondal, A. V. Yakhnin, A. Sebastian, I. Albert, P. Babitzke, NusA-dependent transcription termination prevents misregulation of global gene expression. *Nat Microbiol* **1**, 15007 (2016).
58. J. Zhang, R. Landick, A Two-Way Street: Regulatory Interplay between RNA Polymerase and Nascent RNA Structure. *Trends Biochem Sci* **41**, 293-310 (2016).
59. Z. F. Mandell *et al.*, NusG is an intrinsic transcription termination factor that stimulates motility and coordinates gene expression with NusA. *Elife* **10** (2021).
60. R. A. Weisberg, M. E. Gottesman, Processive antitermination. *J Bacteriol* **181**, 359-367 (1999).
61. J. R. Goodson, W. C. Winkler, Processive Antitermination. *Microbiol Spectr* **6** (2018).
62. L. F. Lau, J. W. Roberts, R. Wu, RNA polymerase pausing and transcript release at the lambda tR1 terminator in vitro. *J Biol Chem* **258**, 9391-9397 (1983).
63. R. K. Montange, R. T. Batey, Riboswitches: emerging themes in RNA structure and function. *Annu Rev Biophys* **37**, 117-133 (2008).
64. W. C. Winkler, R. R. Breaker, Regulation of bacterial gene expression by riboswitches. *Annu Rev Microbiol* **59**, 487-517 (2005).
65. W. C. Winkler, Riboswitches and the role of noncoding RNAs in bacterial metabolic control. *Curr Opin Chem Biol* **9**, 594-602 (2005).
66. W. C. Winkler, Metabolic monitoring by bacterial mRNAs. *Arch Microbiol* **183**, 151-159 (2005).
67. M. Strauß *et al.*, Transcription is regulated by NusA:NusG interaction. *Nucleic Acids Res* **44**, 5971-5982 (2016).
68. C. M. Burns, L. V. Richardson, J. P. Richardson, Combinatorial effects of NusA and NusG on transcription elongation and Rho-dependent termination in Escherichia coli. *J Mol Biol* **278**, 307-316 (1998).
69. M. Worbs, G. P. Bourenkov, H. D. Bartunik, R. Huber, M. C. Wahl, An extended RNA binding surface through arrayed S1 and KH domains in transcription factor NusA. *Mol Cell* **7**, 1177-1189 (2001).
70. A. Eisenmann, S. Schwarz, S. Prasch, K. Schweimer, P. Rösch, The E. coli NusA carboxy-terminal domains are structurally similar and show specific RNAP- and lambdaN interaction. *Protein Sci* **14**, 2018-2029 (2005).
71. K. Schweimer *et al.*, NusA interaction with the α subunit of E. coli RNA polymerase is via the UP element site and releases autoinhibition. *Structure* **19**, 945-954 (2011).
72. K. Liu, Y. Zhang, K. Severinov, A. Das, M. M. Hanna, Role of Escherichia coli RNA polymerase alpha subunit in modulation of pausing, termination and anti-termination by the transcription elongation factor NusA. *EMBO J* **15**, 150-161 (1996).

73. A. V. Yakhnin, P. Babitzke, NusA-stimulated RNA polymerase pausing and termination participates in the Bacillus subtilis trp operon attenuation mechanism invitro. *Proc Natl Acad Sci U S A* **99**, 11067-11072 (2002).
74. S. Shankar, A. Hatoum, J. W. Roberts, A transcription antiterminator constructs a NusA-dependent shield to the emerging transcript. *Mol Cell* **27**, 914-927 (2007).
75. T. F. Mah, J. Li, A. R. Davidson, J. Greenblatt, Functional importance of regions in Escherichia coli elongation factor NusA that interact with RNA polymerase, the bacteriophage lambda N protein and RNA. *Mol Microbiol* **34**, 523-537 (1999).
76. A. Sevostyanova, G. A. Belogurov, R. A. Mooney, R. Landick, I. Artsimovitch, The β subunit gate loop is required for RNA polymerase modification by RfaH and NusG. *Mol Cell* **43**, 253-262 (2011).
77. J. Y. Kang *et al.*, Structural Basis for Transcript Elongation Control by NusG Family Universal Regulators. *Cell* **173**, 1650-1662.e1614 (2018).
78. B. M. Burmann *et al.*, A NusE:NusG complex links transcription and translation. *Science* **328**, 501-504 (2010).
79. S. K. Tomar, I. Artsimovitch, NusG-Spt5 proteins-Universal tools for transcription modification and communication. *Chem Rev* **113**, 8604-8619 (2013).
80. I. Artsimovitch, S. H. Knauer, Ancient Transcription Factors in the News. *mBio* **10** (2019).
81. A. V. Yakhnin, P. Babitzke, NusG/Spt5: are there common functions of this ubiquitous transcription elongation factor? *Curr Opin Microbiol* **18**, 68-71 (2014).
82. J. Drögemüller *et al.*, Determination of RNA polymerase binding surfaces of transcription factors by NMR spectroscopy. *Sci Rep* **5**, 16428 (2015).
83. I. Artsimovitch, R. Landick, Pausing by bacterial RNA polymerase is mediated by mechanistically distinct classes of signals. *Proc Natl Acad Sci U S A* **97**, 7090-7095 (2000).
84. R. A. Mooney, K. Schweimer, P. Rösch, M. Gottesman, R. Landick, Two structurally independent domains of E. coli NusG create regulatory plasticity via distinct interactions with RNA polymerase and regulators. *J Mol Biol* **391**, 341-358 (2009).
85. P. K. Zuber, K. Schweimer, P. Rösch, I. Artsimovitch, S. H. Knauer, Reversible fold-switching controls the functional cycle of the antitermination factor RfaH. *Nat Commun* **10**, 702 (2019).
86. J. M. Peters *et al.*, Rho and NusG suppress pervasive antisense transcription in Escherichia coli. *Genes Dev* **26**, 2621-2633 (2012).
87. C. Wang *et al.*, Structural basis of transcription-translation coupling. *Science* **369**, 1359-1365 (2020).
88. R. S. Washburn *et al.*, Escherichia coli NusG Links the Lead Ribosome with the Transcription Elongation Complex. *iScience* **23**, 101352 (2020).
89. S. Saxena *et al.*, Escherichia coli transcription factor NusG binds to 70S ribosomes. *Mol Microbiol* **108**, 495-504 (2018).

90. J. Greenblatt, J. R. Nodwell, S. W. Mason, Transcriptional antitermination. *Nature* **364**, 401-406 (1993).
91. D. I. Friedman, D. L. Court, Transcription antitermination: the lambda paradigm updated. *Mol Microbiol* **18**, 191-200 (1995).
92. B. M. Burmann, X. Luo, P. Rösch, M. C. Wahl, M. E. Gottesman, Fine tuning of the E. coli NusB:NusE complex affinity to BoxA RNA is required for processive antitermination. *Nucleic Acids Res* **38**, 314-326 (2010).
93. D. I. Friedman, A. T. Schauer, M. R. Baumann, L. S. Baron, S. L. Adhya, Evidence that ribosomal protein S10 participates in control of transcription termination. *Proc Natl Acad Sci U S A* **78**, 1115-1118 (1981).
94. J. R. Nodwell, J. Greenblatt, The nut site of bacteriophage lambda is made of RNA and is bound by transcription antitermination factors on the surface of RNA polymerase. *Genes Dev* **5**, 2141-2151 (1991).
95. E. Burova *et al.*, Escherichia coli nusG mutations that block transcription termination by coliphage HK022 Nun protein. *Mol Microbiol* **31**, 1783-1793 (1999).
96. W. Mathisen, [Urologic aspects of urinary tract infections]. *Nord Med* **92**, 10 (1977).
97. C. R. Conant, J. P. Goodarzi, S. E. Weitzel, P. H. von Hippel, The antitermination activity of bacteriophage lambda N protein is controlled by the kinetics of an RNA-looping-facilitated interaction with the transcription complex. *J Mol Biol* **384**, 87-108 (2008).
98. P. Legault, J. Li, J. Mogridge, L. E. Kay, J. Greenblatt, NMR structure of the bacteriophage lambda N peptide/boxB RNA complex: recognition of a GNRA fold by an arginine-rich motif. *Cell* **93**, 289-299 (1998).
99. A. Das, Control of transcription termination by RNA-binding proteins. *Annu Rev Biochem* **62**, 893-930 (1993).
100. A. Das, How the phage lambda N gene product suppresses transcription termination: communication of RNA polymerase with regulatory proteins mediated by signals in nascent RNA. *J Bacteriol* **174**, 6711-6716 (1992).
101. N. Said *et al.*, Structural basis for λN-dependent processive transcription antitermination. *Nat Microbiol* **2**, 17062 (2017).
102. F. Krupp *et al.*, Structural Basis for the Action of an All-Purpose Transcription Anti-termination Factor. *Mol Cell* **74**, 143-157.e145 (2019).
103. X. Yang, P. J. Lewis, The interaction between RNA polymerase and the elongation factor NusA. *RNA Biol* **7**, 272-275 (2010).
104. X. Yang *et al.*, The structure of bacterial RNA polymerase in complex with the essential transcription elongation factor NusA. *EMBO Rep* **10**, 997-1002 (2009).
105. M. W. Webster *et al.*, Structural basis of transcription-translation coupling and collision in bacteria. *Science* **369**, 1355-1359 (2020).
106. M. R. Lawson *et al.*, Mechanism for the Regulated Control of Bacterial Transcription Termination by a Universal Adaptor Protein. *Mol Cell* **71**, 911-922.e914 (2018).
107. I. Herskowitz, E. R. Signer, A site essential for expression of all late genes in bacteriophage lambda. *J Mol Biol* **47**, 545-556 (1970).

108. S. Barik, A. Das, An analysis of the role of host factors in transcription antitermination in vitro by the Q protein of coliphage lambda. *Mol Gen Genet* **222**, 152-156 (1990).
109. W. S. Yarnell, J. W. Roberts, The phage lambda gene Q transcription antiterminator binds DNA in the late gene promoter as it modifies RNA polymerase. *Cell* **69**, 1181-1189 (1992).
110. J. Shi *et al.*, Structural basis of Q-dependent transcription antitermination. *Nat Commun* **10**, 2925 (2019).
111. Z. Yin, J. T. Kaelber, R. H. Ebright, Structural basis of Q-dependent antitermination. *Proc Natl Acad Sci U S A* **116**, 18384-18390 (2019).
112. M. Bubunenko *et al.*, Nus transcription elongation factors and RNase III modulate small ribosome subunit biogenesis in Escherichia coli. *Mol Microbiol* **87**, 382-393 (2013).
113. B. R. Dudenhoeffer, H. Schneider, K. Schweimer, S. H. Knauer, SuhB is an integral part of the ribosomal antitermination complex and interacts with NusA. *Nucleic Acids Res* **47**, 6504-6518 (2019).
114. X. Luo *et al.*, Structural and functional analysis of the E. coli NusB-S10 transcription antitermination complex. *Mol Cell* **32**, 791-802 (2008).
115. N. Singh *et al.*, SuhB Associates with Nus Factors To Facilitate 30S Ribosome Biogenesis in Escherichia coli. *mBio* **7**, e00114 (2016).
116. Y. H. Huang, N. Said, B. Loll, M. C. Wahl, Structural basis for the function of SuhB as a transcription factor in ribosomal RNA synthesis. *Nucleic Acids Res* **47**, 6488-6503 (2019).
117. K. Shiba, K. Ito, T. Yura, Mutation that suppresses the protein export defect of the secY mutation and causes cold-sensitive growth of Escherichia coli. *J Bacteriol* **160**, 696-701 (1984).
118. R. Yano, H. Nagai, K. Shiba, T. Yura, A mutation that enhances synthesis of sigma 32 and suppresses temperature-sensitive growth of the rpoH15 mutant of Escherichia coli. *J Bacteriol* **172**, 2124-2130 (1990).
119. S. F. Chang, D. Ng, L. Baird, C. Georgopoulos, Analysis of an Escherichia coli dnaB temperature-sensitive insertion mutation and its cold-sensitive extragenic suppressor. *J Biol Chem* **266**, 3654-3660 (1991).
120. R. A. King, S. Banik-Maiti, D. J. Jin, R. A. Weisberg, Transcripts that increase the processivity and elongation rate of RNA polymerase. *Cell* **87**, 893-903 (1996).
121. R. A. King, R. A. Weisberg, Suppression of factor-dependent transcription termination by antiterminator RNA. *J Bacteriol* **185**, 7085-7091 (2003).
122. S. Banik-Maiti, R. A. King, R. A. Weisberg, The antiterminator RNA of phage HK022. *J Mol Biol* **272**, 677-687 (1997).
123. M. Clerget, D. J. Jin, R. A. Weisberg, A zinc-binding region in the beta' subunit of RNA polymerase is involved in antitermination of early transcription of phage HK022. *J Mol Biol* **248**, 768-780 (1995).
124. I. Irnov, W. C. Winkler, A regulatory RNA required for antitermination of biofilm and capsular polysaccharide operons in Bacillales. *Mol Microbiol* **76**, 559-575 (2010).

125. B. Wang, I. Artsimovitch, NusG, an Ancient Yet Rapidly Evolving Transcription Factor. *Front Microbiol* **11**, 619618 (2020).
126. F. Werner, D. Grohmann, Evolution of multisubunit RNA polymerases in the three domains of life. *Nat Rev Microbiol* **9**, 85-98 (2011).
127. D. L. Lindstrom *et al.*, Dual roles for Spt5 in pre-mRNA processing and transcription elongation revealed by identification of Spt5-associated proteins. *Mol Cell Biol* **23**, 1368-1378 (2003).
128. G. Diamant, A. Bahat, R. Dikstein, The elongation factor Spt5 facilitates transcription initiation for rapid induction of inflammatory-response genes. *Nat Commun* **7**, 11547 (2016).
129. R. W. Maul *et al.*, Spt5 accumulation at variable genes distinguishes somatic hypermutation in germinal center B cells from ex vivo-activated cells. *J Exp Med* **211**, 2297-2306 (2014).
130. B. Wang, V. M. Gumerov, E. P. Andrianova, I. B. Zhulin, I. Artsimovitch, Origins and Molecular Evolution of the NusG Paralog RfaH. *mBio* **11** (2020).
131. J. R. Goodson, S. Klupt, C. Zhang, P. Straight, W. C. Winkler, LoaP is a broadly conserved antiterminator protein that regulates antibiotic gene clusters in Bacillus amyloliquefaciens. *Nat Microbiol* **2**, 17003 (2017).
132. Y. Paitan, E. Orr, E. Z. Ron, E. Rosenberg, A NusG-like transcription anti-terminator is involved in the biosynthesis of the polyketide antibiotic TA of Myxococcus xanthus. *FEMS Microbiol Lett* **170**, 221-227 (1999).
133. B. Núñez, P. Avila, F. de la Cruz, Genes involved in conjugative DNA processing of plasmid R6K. *Mol Microbiol* **24**, 1157-1168 (1997).
134. I. Artsimovitch, R. Landick, The transcriptional regulator RfaH stimulates RNA chain synthesis after recruitment to elongation complexes by the exposed nontemplate DNA strand. *Cell* **109**, 193-203 (2002).
135. G. A. Belogurov, R. A. Mooney, V. Svetlov, R. Landick, I. Artsimovitch, Functional specialization of transcription elongation factors. *EMBO J* **28**, 112-122 (2009).
136. C. Wandersman, S. Létoffé, Involvement of lipopolysaccharide in the secretion of Escherichia coli alpha-haemolysin and Erwinia chrysanthemi proteases. *Mol Microbiol* **7**, 141-150 (1993).
137. G. Nagy *et al.*, Loss of regulatory protein RfaH attenuates virulence of uropathogenic Escherichia coli. *Infect Immun* **70**, 4406-4413 (2002).
138. G. Liu, J. E. Olsen, L. E. Thomsen, Identification of Genes Essential for Antibiotic-Induced Up-Regulation of Plasmid-Transfer-Genes in Cephalosporin Resistant. *Front Microbiol* **10**, 2203 (2019).
139. M. P. Stevens, P. Hänfling, B. Jann, K. Jann, I. S. Roberts, Regulation of Escherichia coli K5 capsular polysaccharide expression: evidence for involvement of RfaH in the expression of group II capsules. *FEMS Microbiol Lett* **124**, 93-98 (1994).
140. V. Svetlov, E. Nudler, Reading of the non-template DNA by transcription elongation factors. *Mol Microbiol* **109**, 417-421 (2018).
141. P. K. Zuber *et al.*, The universally-conserved transcription factor RfaH is recruited to a hairpin structure of the non-template DNA strand. *Elife* **7** (2018).

142. K. Hu, I. Artsimovitch, A Screen for rfaH Suppressors Reveals a Key Role for a Connector Region of Termination Factor Rho. *mBio* **8** (2017).
143. B. M. Burmann *et al.*, An α helix to β barrel domain switch transforms the transcription factor RfaH into a translation factor. *Cell* **150**, 291-303 (2012).
144. J. Yuan *et al.*, Antibacterial Compounds-Macrolactin Alters the Soil Bacterial Community and Abundance of the Gene Encoding PKS. *Front Microbiol* **7**, 1904 (2016).
145. M. A. Mondol, J. H. Kim, H. S. Lee, Y. J. Lee, H. J. Shin, Macrolactin W, a new antibacterial macrolide from a marine Bacillus sp. *Bioorg Med Chem Lett* **21**, 3832-3835 (2011).
146. X. H. Chen *et al.*, Structural and functional characterization of three polyketide synthase gene clusters in Bacillus amyloliquefaciens FZB 42. *J Bacteriol* **188**, 4024-4036 (2006).
147. L. Wu *et al.*, Difficidin and bacilysin from Bacillus amyloliquefaciens FZB42 have antibacterial activity against Xanthomonas oryzae rice pathogens. *Sci Rep* **5**, 12975 (2015).
148. T. Lincke, S. Behnken, K. Ishida, M. Roth, C. Hertweck, Closthioamide: an unprecedented polythioamide antibiotic from the strictly anaerobic bacterium Clostridium cellulolyticum. *Angew Chem Int Ed Engl* **49**, 2011-2013 (2010).
149. A. I. Chiriac *et al.*, Mode of action of closthioamide: the first member of the polythioamide class of bacterial DNA gyrase inhibitors. *J Antimicrob Chemother* **70**, 2576-2588 (2015).
150. D. Shi, D. Svetlov, R. Abagyan, I. Artsimovitch, Flipping states: a few key residues decide the winning conformation of the only universally conserved transcription factor. *Nucleic Acids Res* **45**, 8835-8843 (2017).
151. G. A. Belogurov *et al.*, Structural basis for converting a general transcription factor into an operon-specific virulence regulator. *Mol Cell* **26**, 117-129 (2007).
152. B. M. Burmann, U. Scheckenhofer, K. Schweimer, P. Rösch, Domain interactions of the transcription-translation coupling factor Escherichia coli NusG are intermolecular and transient. *Biochem J* **435**, 783-789 (2011).
153. M. J. Bailey, C. Hughes, V. Koronakis, RfaH and the ops element, components of a novel system controlling bacterial transcription elongation. *Mol Microbiol* **26**, 845-851 (1997).
154. R. R. Burgess, Protein precipitation techniques. *Methods Enzymol* **463**, 331-342 (2009).
155. S. Frey, D. Gorlich, A new set of highly efficient, tag-cleaving proteases for purifying recombinant proteins. *J Chromatogr A* **1337**, 95-105 (2014).
156. J. A. Glasel, Validity of nucleic acid purities monitored by 260nm/280nm absorbance ratios. *Biotechniques* **18**, 62-63 (1995).
157. R. M. Cordes, W. B. Sims, C. E. Glatz, Precipitation of nucleic acids with poly(ethyleneimine). *Biotechnol Prog* **6**, 283-285 (1990).
158. K. A. Curtis *et al.*, Unusual Salt and pH Induced Changes in Polyethylenimine Solutions. *PLoS One* **11**, e0158147 (2016).
159. G. E. Johnson, J. B. Lalanne, M. L. Peters, G. W. Li, Functionally uncoupled transcription-translation in Bacillus subtilis. *Nature* **585**, 124-128 (2020).

160. B. Fan *et al.*, dRNA-Seq Reveals Genomewide TSSs and Noncoding RNAs of Plant Beneficial Rhizobacterium Bacillus amyloliquefaciens FZB42. *PLoS One* **10**, e0142002 (2015).
161. W. L. Ruzzo, J. Gorodkin, De novo discovery of structured ncRNA motifs in genomic sequences. *Methods Mol Biol* **1097**, 303-318 (2014).
162. S. Bottaro, K. Lindorff-Larsen, Mapping the Universe of RNA Tetraloop Folds. *Biophys J* **113**, 257-267 (2017).
163. K. G. Roelofs, J. Wang, H. O. Sintim, V. T. Lee, Differential radial capillary action of ligand assay for high-throughput detection of protein-metabolite interactions. *Proc Natl Acad Sci U S A* **108**, 15528-15533 (2011).
164. G. P. Donaldson, K. G. Roelofs, Y. Luo, H. O. Sintim, V. T. Lee, A rapid assay for affinity and kinetics of molecular interactions with nucleic acids. *Nucleic Acids Res* **40**, e48 (2012).
165. D. K. Patel, M. P. Gebbie, V. T. Lee, Assessing RNA interactions with proteins by DRaCALA. *Methods Enzymol* **549**, 489-512 (2014).
166. W. Zhang, J. A. Dunkle, J. H. Cate, Structures of the ribosome in intermediate states of ratcheting. *Science* **325**, 1014-1017 (2009).
167. J. R. Weir, F. Bonneau, J. Hentschel, E. Conti, Structural analysis reveals the characteristic features of Mtr4, a DExH helicase involved in nuclear RNA processing and surveillance. *Proc Natl Acad Sci U S A* **107**, 12139-12144 (2010).
168. K. Wang, R. Samudrala, Incorporating background frequency improves entropy-based residue conservation measures. *BMC Bioinformatics* **7**, 385 (2006).
169. J. A. Capra, M. Singh, Predicting functionally important residues from sequence conservation. *Bioinformatics* **23**, 1875-1882 (2007).
170. D. Lazinski, E. Grzadzielska, A. Das, Sequence-specific recognition of RNA hairpins by bacteriophage antiterminators requires a conserved arginine-rich motif. *Cell* **59**, 207-218 (1989).
171. T. S. Bayer, L. N. Booth, S. M. Knudsen, A. D. Ellington, Arginine-rich motifs present multiple interfaces for specific binding by RNA. *RNA* **11**, 1848-1857 (2005).
172. A. V. Yakhnin, H. Yakhnin, P. Babitzke, Function of the Bacillus subtilis transcription elongation factor NusG in hairpin-dependent RNA polymerase pausing in the trp leader. *Proc Natl Acad Sci U S A* **105**, 16131-16136 (2008).
173. A. V. Yakhnin, K. S. Murakami, P. Babitzke, NusG Is a Sequence-specific RNA Polymerase Pause Factor That Binds to the Non-template DNA within the Paused Transcription Bubble. *J Biol Chem* **291**, 5299-5308 (2016).
174. A. V. Yakhnin *et al.*, NusG controls transcription pausing and RNA polymerase translocation throughout the. *Proc Natl Acad Sci U S A* **117**, 21628-21636 (2020).
175. A. V. Yakhnin, M. Kashlev, P. Babitzke, NusG-dependent RNA polymerase pausing is a frequent function of this universally conserved transcription elongation factor. *Crit Rev Biochem Mol Biol*, 1-13 (2020).
176. A. J. Blythe *et al.*, The yeast transcription elongation factor Spt4/5 is a sequence-specific RNA binding protein. *Protein Sci* **25**, 1710-1721 (2016).

177. A. Missra, D. S. Gilmour, Interactions between DSIF (DRB sensitivity inducing factor), NELF (negative elongation factor), and the Drosophila RNA polymerase II transcription elongation complex. *Proc Natl Acad Sci U S A* **107**, 11301-11306 (2010).
178. P. A. Meyer *et al.*, Structures and Functions of the Multiple KOW Domains of Transcription Elongation Factor Spt5. *Mol Cell Biol* **35**, 3354-3369 (2015).
179. W. Li, C. Giles, S. Li, Insights into how Spt5 functions in transcription elongation and repressing transcription coupled DNA repair. *Nucleic Acids Res* **42**, 7069-7083 (2014).
180. N. C. Kyrpides, C. R. Woese, C. A. Ouzounis, KOW: a novel motif linking a bacterial transcription factor with ribosomal proteins. *Trends Biochem Sci* **21**, 425-426 (1996).
181. T. Steiner, J. T. Kaiser, S. Marinkoviç, R. Huber, M. C. Wahl, Crystal structures of transcription factor NusG in light of its nucleic acid- and protein-binding activities. *EMBO J* **21**, 4641-4653 (2002).
182. J. R. Knowlton *et al.*, A spring-loaded state of NusG in its functional cycle is suggested by X-ray crystallography and supported by site-directed mutants. *Biochemistry* **42**, 2275-2281 (2003).
183. G. Varani, K. Nagai, RNA recognition by RNP proteins during RNA processing. *Annu Rev Biophys Biomol Struct* **27**, 407-445 (1998).
184. H. Walden, Selenium incorporation using recombinant techniques. *Acta Crystallogr D Biol Crystallogr* **66**, 352-357 (2010).
185. R. Tan, A. D. Frankel, Structural variety of arginine-rich RNA-binding peptides. *Proc Natl Acad Sci U S A* **92**, 5282-5286 (1995).
186. C. A. Smith, L. Chen, A. D. Frankel, Using peptides as models of RNA-protein interactions. *Methods Enzymol* **318**, 423-438 (2000).
187. F. Casu, B. M. Duggan, M. Hennig, The arginine-rich RNA-binding motif of HIV-1 Rev is intrinsically disordered and folds upon RRE binding. *Biophys J* **105**, 1004-1017 (2013).
188. B. J. Calnan, S. Biancalana, D. Hudson, A. D. Frankel, Analysis of arginine-rich peptides from the HIV Tat protein reveals unusual features of RNA-protein recognition. *Genes Dev* **5**, 201-210 (1991).
189. L. Chen, A. D. Frankel, An RNA-binding peptide from bovine immunodeficiency virus Tat protein recognizes an unusual RNA structure. *Biochemistry* **33**, 2708-2715 (1994).
190. R. Tan, L. Chen, J. A. Buettner, D. Hudson, A. D. Frankel, RNA recognition by an isolated alpha helix. *Cell* **73**, 1031-1040 (1993).
191. Z. Cai *et al.*, Solution structure of P22 transcriptional antitermination N peptide-boxB RNA complex. *Nat Struct Biol* **5**, 203-212 (1998).
192. J. L. Battiste *et al.*, Alpha helix-RNA major groove recognition in an HIV-1 rev peptide-RRE RNA complex. *Science* **273**, 1547-1551 (1996).
193. N. L. Greenbaum, How Tat targets TAR: structure of the BIV peptide-RNA complex. *Structure* **4**, 5-9 (1996).
194. V. Calabro, M. D. Daugherty, A. D. Frankel, A single intermolecular contact mediates intramolecular stabilization of both RNA and protein. *Proc Natl Acad Sci U S A* **102**, 6849-6854 (2005).

195. C. R. Woese, S. Winker, R. R. Gutell, Architecture of ribosomal RNA: constraints on the sequence of "tetra-loops". *Proc Natl Acad Sci U S A* **87**, 8467-8471 (1990).
196. J. P. Sheehy, A. R. Davis, B. M. Znosko, Thermodynamic characterization of naturally occurring RNA tetraloops. *RNA* **16**, 417-429 (2010).
197. M. Molinaro, I. Tinoco, Use of ultra stable UNCG tetraloop hairpins to fold RNA structures: thermodynamic and spectroscopic applications. *Nucleic Acids Res* **23**, 3056-3063 (1995).
198. E. Ennifar *et al.*, The crystal structure of UUCG tetraloop. *J Mol Biol* **304**, 35-42 (2000).
199. F. H. Allain, G. Varani, Structure of the P1 helix from group I self-splicing introns. *J Mol Biol* **250**, 333-353 (1995).
200. D. J. Williams, K. B. Hall, Unrestrained stochastic dynamics simulations of the UUCG tetraloop using an implicit solvation model. *Biophys J* **76**, 3192-3205 (1999).
201. D. J. Williams, K. B. Hall, Experimental and theoretical studies of the effects of deoxyribose substitutions on the stability of the UUCG tetraloop. *J Mol Biol* **297**, 251-265 (2000).
202. S. Nozinovic, B. Fürtig, H. R. Jonker, C. Richter, H. Schwalbe, High-resolution NMR structure of an RNA model system: the 14-mer cUUCGg tetraloop hairpin RNA. *Nucleic Acids Res* **38**, 683-694 (2010).
203. C. Cheong, G. Varani, I. Tinoco, Solution structure of an unusually stable RNA hairpin, 5'GGAC(UUCG)GUCC. *Nature* **346**, 680-682 (1990).
204. G. Varani, C. Cheong, I. Tinoco, Structure of an unusually stable RNA hairpin. *Biochemistry* **30**, 3280-3289 (1991).
205. Z. Otwinowski, W. Minor, Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol* **276**, 307-326 (1997).
206. G. M. Sheldrick, Experimental phasing with SHELXC/D/E: combining chain tracing with density modification. *Acta Crystallogr D Biol Crystallogr* **66**, 479-485 (2010).
207. T. C. Terwilliger *et al.*, Decision-making in structure solution using Bayesian estimates of map quality: the PHENIX AutoSol wizard. *Acta Crystallogr D Biol Crystallogr* **65**, 582-601 (2009).
208. T. C. Terwilliger, Maximum-likelihood density modification using pattern recognition of structural motifs. *Acta Crystallogr D Biol Crystallogr* **57**, 1755-1762 (2001).
209. A. J. McCoy *et al.*, Phaser crystallographic software. *J Appl Crystallogr* **40**, 658-674 (2007).
210. K. L. Dunbar *et al.*, Genome Editing Reveals Novel Thiotemplated Assembly of Polythioamide Antibiotics in Anaerobic Bacteria. *Angew Chem Int Ed Engl* **57**, 14080-14084 (2018).