2008

Structural and Computational Studies of RNA Polymerase

William J. Lane

STRUCTURAL AND COMPUTATIONAL STUDIES OF RNA POLYMERASE

A Thesis Presented to the Faculty of

The Rockefeller University

in Partial Fulfillment of the Requirements for

the degree of Doctor of Philosophy

by

William J. Lane

June 2008

STRUCTURAL AND COMPUTATIONAL STUDIES OF RNA POLYMERASE

William J. Lane, Ph.D.
The Rockefeller University 2008

Although the multi-subunit RNA Polymerase (RNAP) structures have revolutionized our understanding of transcription, we still do not fully understand the molecular details of bacterial promoter recognition and melting. In addition, our understanding is generally limited to highly conserved elements of the structure, with little focus on the bacterial lineage-specific domain insertions. Furthermore, we lack information about the hidden functional residue networks that underlie the activities of this complex multi-subunit molecular machine. By combining structural and computational methods we:

(1) Used X-ray crystallography to investigate promoter -35 element recognition by domain 4 of the Group IV sigma factors, revealing that conserved positions within -35 element induce an AA/TT-tract like DNA geometry, allowing for indirect promoter recognition despite the absence of direct protein/DNA interactions with several highly conserved DNA bases.

(2) Created comprehensive multiple sequence alignments for the two bacterial large subunits (β/β′) and their homologues from the following multi-subunit RNAPs: bacterial, eukaryotic pol I/II/III, Nuclear-Cytoplasmic Large double-stranded DNA Viruses, archeal, and plant plastid. To aid in the creation of the alignments we also developed a sequence retrieval and processing system termed BlaFA (BLAST to FASTA File to Alignment). As a result of our analysis

we gained insights into shared sequence regions, the bacterial large subunit intergenic spacing, and the bacterial lineage-specific domain insertions.

(3) Used our multi-subunit RNAP alignments and Statistical Coupling Analysis (SCA) to determine co-evolving residue networks within and between the two large subunits, as well as with omega/Rpb6. In addition, we uncovered a previously unidentified principle of co-evolution, namely the role of adapter groups in bridging and coordinating the independently evolving main group networks which were responsible for key aspects of transcription including: catalysis and RNAP interactions with DNA and RNA during initiation, elongation, and termination.

(4) Used structure based modeling to computationally trap and understand promoter -10 element recognition by the Group I sigma factors. Our results revealed an unexpected upstream shift of the -10 element during recognition possibly validating a previously proposed twist and melt mechanism of DNA melting.

(5) Determined the conditions necessary to express and purify a Group IV sigma factor from *Thermus Thermophilus*.

Dedicated to my parents.

Acknowledgements

We thank M. Becker and the staff at the National Synchrotron Light Source Beamline X25 for support, and Tom Muir for access to the electrospray mass spectrometer. We thank Lars Westblade, Tom Muir, Chris Lima, Roberto Sanchez, Elizabeth Campbell, Deepti Jain, Valerie Lamour, Olivier Rivoir, and Stan Leibler for helpful discussions and advice. We also thank the Ranganathan Lab especially Rama Ranganathan, Bill Russ, Tina Vo, Rick McLaughlin, and Alan Poole. W.J.L. would like to thank his previous scientific mentors who help him along way including Sina Rabbany, Shahin Rafii, Sérgio Dias, Marshall Michener, and Leanna Levine. In addition, W.J.L. would also like to thank Sara Tuttle, Robert & Joan Lane, and Michele & Brian Yula for their encouragement, support, and patience. W.J.L. was supported by National Institutes of Health MSTP grant GM07739 and The W.M. Keck Foundation Medical Scientist Fellowship. This work was supported by National Institutes of Health grant GM53759 to S.A.D.

Table of Contents

List of Figures

List of Tables

Chapter 1 - Introduction

Bacterial RNAP

Bacterial transcription is driven by the DNA-dependent RNA polymerase (RNAP), comprising five core subunits ($\alpha_2\beta\beta'\omega$) plus an initiation-specific σ subunit, which binds to the core RNAP to form holoenzyme [1-3].

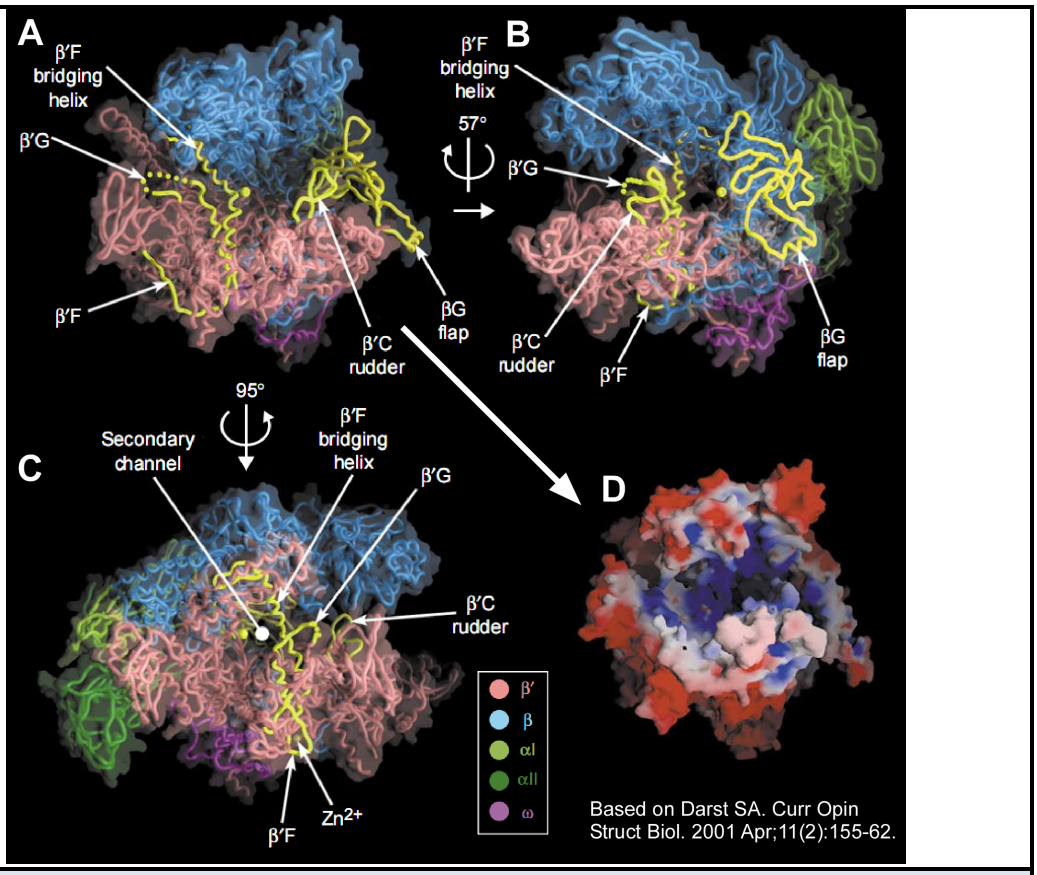


Figure 1.1 – Structure of Bacterial RNAP Core.

(A/B/C) Core RNAP is comprised of 5 subunits shown as cartoon worms and transparent molecular surfaces with α_2 (dark and light green), β (light blue), β' (pink), and ω (purple). The active center Mg^{2+} is shown as a yellow sphere. **(D)** Molecular surface colored according to according to the electrostatic potential with acidic (red), neutral (white), and basic (blue).

The *Thermus aquaticus* (*Taq*) and *Thermus Thermophilus* (*Tth*) crystal structures [4-9] have revealed that bacterial RNAP is shaped like a crab-claw with extensive interactions between the β and β' subunits, which form the two

crab-claw pinchers (Figure 1.1). The central cleft formed by the two large subunits is negatively charged (Figure 1.1D) and interacts with DNA and RNA during the process of transcription. The back wall of the cleft contains the active center Mg^{2+} (MgI) and is the site where incoming nucleotides are added to the growing RNA strand. The α subunit N-terminal dimer binds to the outer surface of β and β' behind the active center and the ω subunit wraps around C-terminus of β' (Figure 1.1).

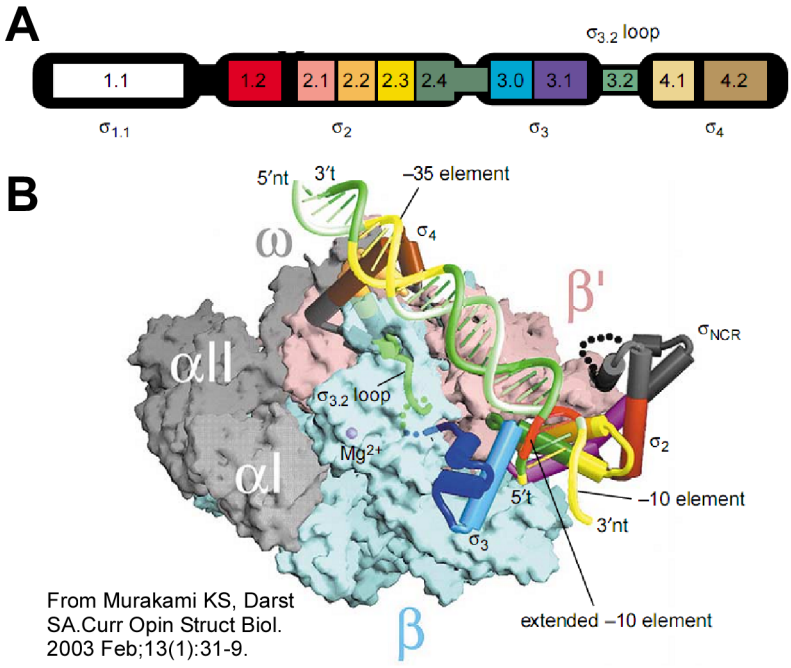


Figure 1.2 – Structure of Bacterial RNAP Holoenzyme.

(A) σ subunit domain architecture. **(B)** Core RNAP subunits are shown as molecular surface with α_2 (gray), β (cyan), β' (pink), and ω (gray). The σ subunit is shown as cylinders with each region color coded as in (A). Fork junction DNA is shown as phosphate-backbone ribbons, with bases indicated schematically as sticks

The bacterial holoenzyme structures [5-7] have revealed that the σ subunit binds to the upstream face of RNAP core opposite the secondary channel (Figure 1.2A and Figure 1.5A). The two globular σ factor domains (σ_2 and σ_4)

responsible for promoter recognition are separated across the face of RNAP with the intervening flexible linker ($\sigma_{3.2}$) entering into core RNAP near the catalytic center (Figure 1.2).

Upon holoenzyme formation RNAP is capable of entering the active stages of the transcription cycle which can be broken into three main steps: initiation, elongation, and termination (Figure 1.3).

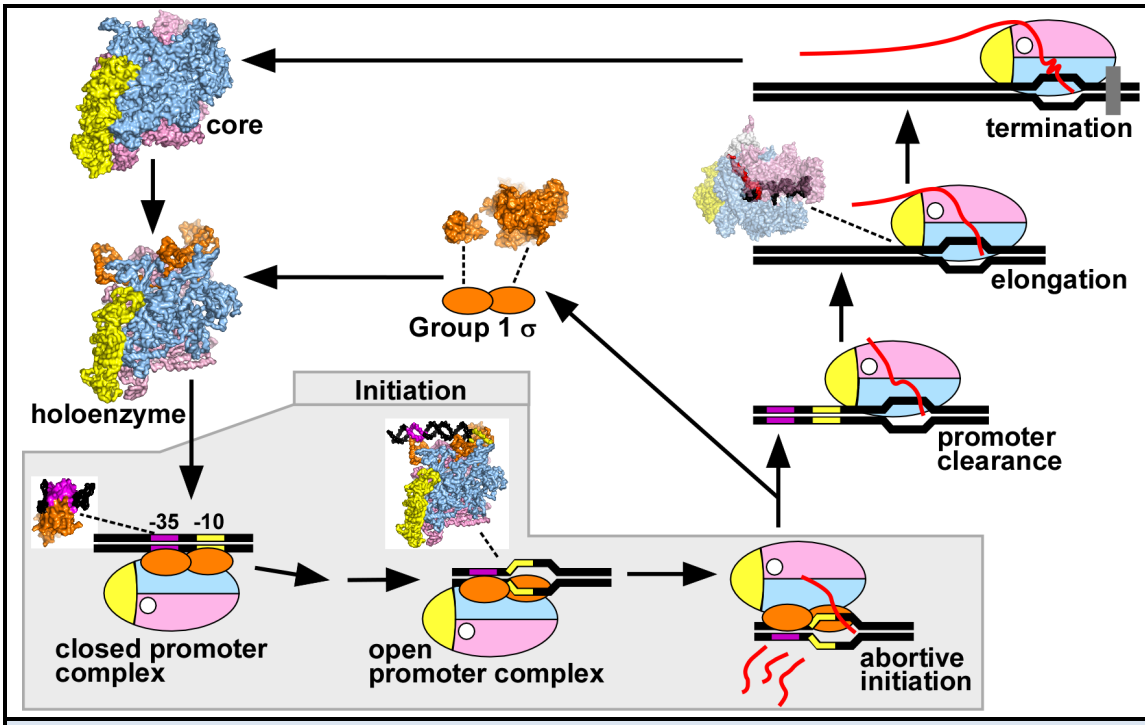


Figure 1.3 – Transcription Cycle.

The process of transcription can be segregated into several distinct biochemical steps. Core RNAP is comprised of 5 subunits with α_2 (yellow), β (light blue), β' (pink), and ω (white), which upon binding the DNA recognition subunit σ (orange) forms the catalytically competent RNAP holoenzyme. Available structures are show a molecular surfaces. Dashed lines indicate partial structures or structures thought to represent a particular state. The gray region highlights the steps involved in transcription initiation.

Sigma (σ) Families

Analysis of the available bacterial genomes has revealed great variation in both the number and type of σ factors each bacterial species possesses [10, 11], allowing for promoter-specific transcription of defined regulons. Most σ factors belong to the σ^{70} family, which can be broadly divided into five subgroups [11, 12]. The Group I (primary) σ factors, such as *Escherichia coli* (*Ec*) σ^{70} and *Taq* σ^{A} (Figure 1.4), direct the transcription of housekeeping genes for which basal levels of transcription are essential for normal cellular processes and survival.

Group II sigma factors are similar to Group I, except that they are not essential. Group III sigma factors promote the transcription of alternative genes necessary for bacterial adaptations such as flagella formation, sporulation, and the cytoplasmic heat shock response. The largest and most diverse subgroup, the Group IV, or ExtraCytoplasmic Function (ECF) σ factors (Figure 1.4), direct the transcription of genes that regulate a wide variety of responses including periplasmic stress, iron transport, metal ion efflux, alginate secretion, and pathogenesis [11, 13-15]. The *Ec* ECF σ factor σ^{E} is an essential protein that directs the response to periplasmic stress [16-19]. Group V sigma factors are responsible for toxic gene expression in *Clostridium difficile* [20].

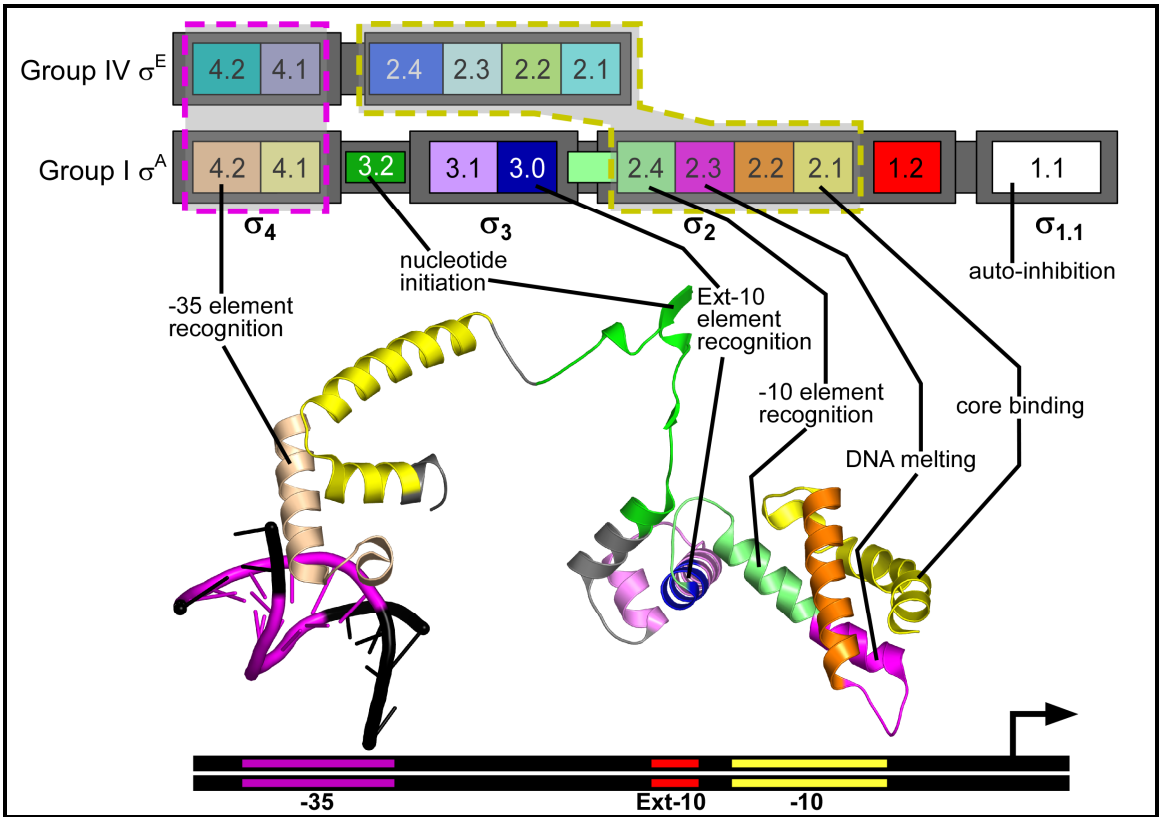


Figure 1.4 – σ Structure and Function.

The structure of the Group I σ factor *Tth* σ^A, from the *Tth* holoenzyme structure is show as a cartoon with each region color coded according to the domain and region schematic above it. In the schematic each σ region is colored separately and domains indicated by larger encompassing dark gray regions above each domain label. The approximate locations of the -10 and -35 interaction sites are indicated below the structure. The interaction between *Tth* σ^A and the -35 element was modeled using the *Taq* σ^A /-35 element structure. The function of each σ^A region is indicated by a line from the schematic to the structure. The domain in common between the Group 1 σ^A and the Group IV σ^E are indicated by purple and yellow dashed regions.

Bacterial RNAP Transcription Initiation

Transcription initiation involved several steps including: closed complex (RP_c) formation, open complex (RP_o) formation, and abortive initiation (Figure 1.5). As shown in Figure 1.5A, promoter-specific transcription initiation first requires the formation of a RP_c in which σ domains 2 (σ_2) and 4 (σ_4) bind sequence-specifically to the -10 and -35 promoter DNA elements, respectively [3, 5, 21].

Upon promoter binding, RP_c quickly and (essentially) irreversibly converts to the RP_o through a series of isomerization steps during which the -10 element is melted and the single stranded template DNA invades the catalytic center (Figure 1.5C). In addition, some promoters contain an extended -10 element (TRTG), located one base upstream of the -10 element, which is thought to stabilize RP_o [22].

Unfortunately, the rapid and irreversible transition from RP_c to RP_o has hindered the experimental study of RP_c promoter recognition and DNA melting. However, the high resolution structure of *Taq* $\sigma^A{}_4$ bound to the major groove of an isolated -35 element has revealed the molecular details of -35 element recognition. In contrast, the exact mechanism of -10 element DNA recognition and melting are still poorly understood. As shown in Figure 1.5B, DNA melting is thought to occur as a result of the natural propensity of the -10 element sequence for thermal breathing, resulting in transiently flipped out bases that can be captured and stabilized by binding to aromatic residues on σ_2 [3, 23]. The low resolution fork junction structure of holoenyme (Figure 1.2B) bound to a mostly non-template single stranded -10 element is thought to represent a partial structure of the RP_o [5].

Upon RP_o formation, RNAP undergoes abortive initiation cycles in which short RNA transcripts are produced and then released with RNAP returning to the transcriptional start site (Figure 1.5C). It is thought that abortive initiation results from a competition between the elongating RNA and a segment of the σ subunit, $\sigma_{3.2}$, which occupies the RNA exit channel of the RNAP [6]. Therefore, in order to

escape abortive initiation, the nascent RNA transcript must displace $\sigma_{3.2}$ (Figure 1.5D) while trying to fully exit RNAP during promoter escape (Figure 1.5E). The displacement of $\sigma_{3.2}$ by the elongating RNA is also thought to contribute to the release of σ from RNAP.

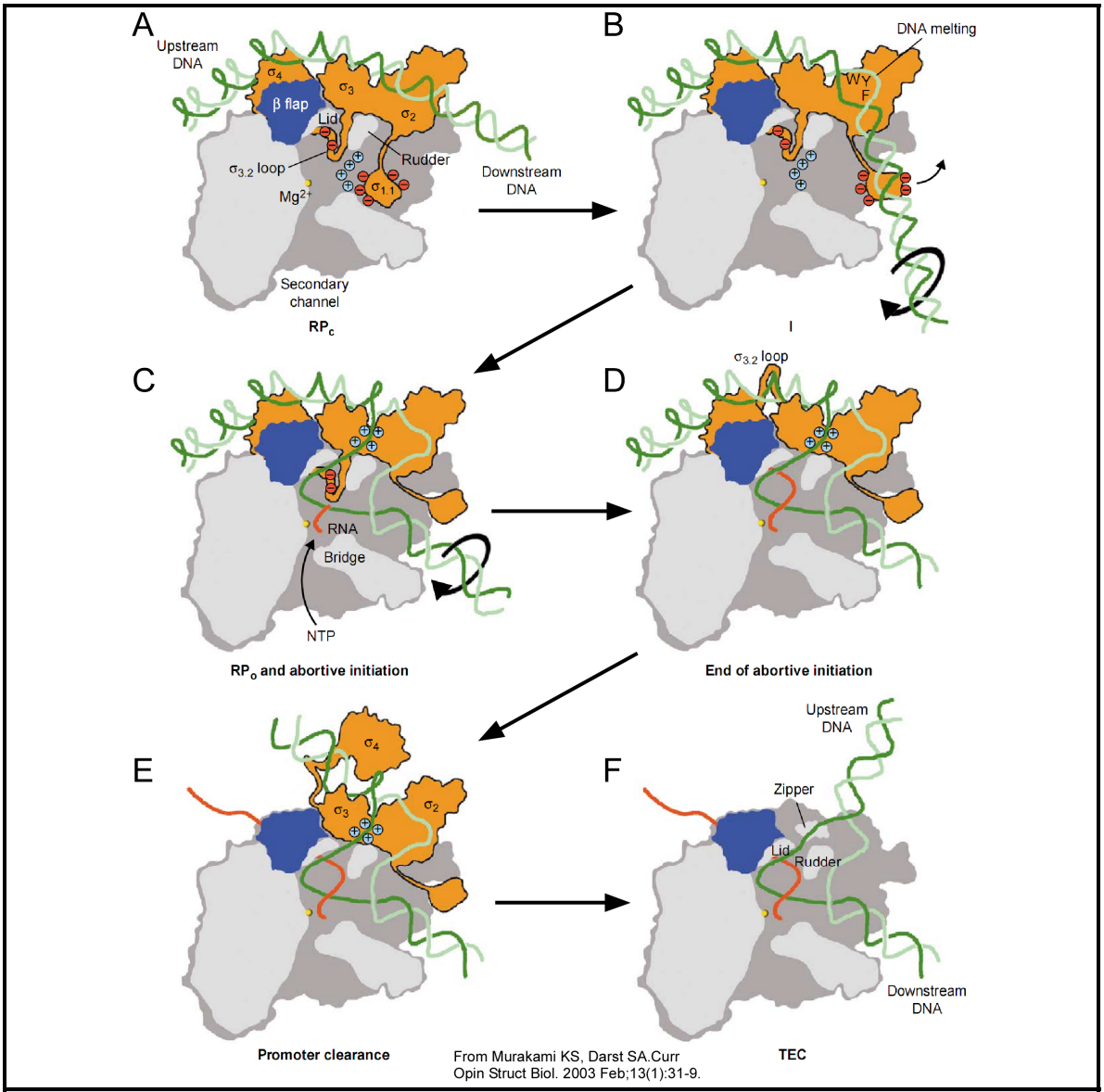


Figure 1.5 – Schematic of Bacterial RNAP Transcription Initiation.

In order to see into the active center RNAP was sliced (horizontally according to view in Figure 1.1) and the top half (mostly β) removed with core subunits (gray), σ subunit (orange), active center MgI (yellow sphere), DNA template strand (dark green), DNA non-template strand (light green), RNA (red), β-flap (blue).

Bacterial RNAP Transcription Elongation

Once abortive initiation is complete, RNAP enters into the elongation phase, which is highlighted by the processive movement of RNAP along the transcribing gene. The recent high resolution *Tth* RNAP core, DNA, and RNA ternary elongation complex (TEC) crystal structure (Figure 1.6) has shed light onto many aspects of elongation [8].

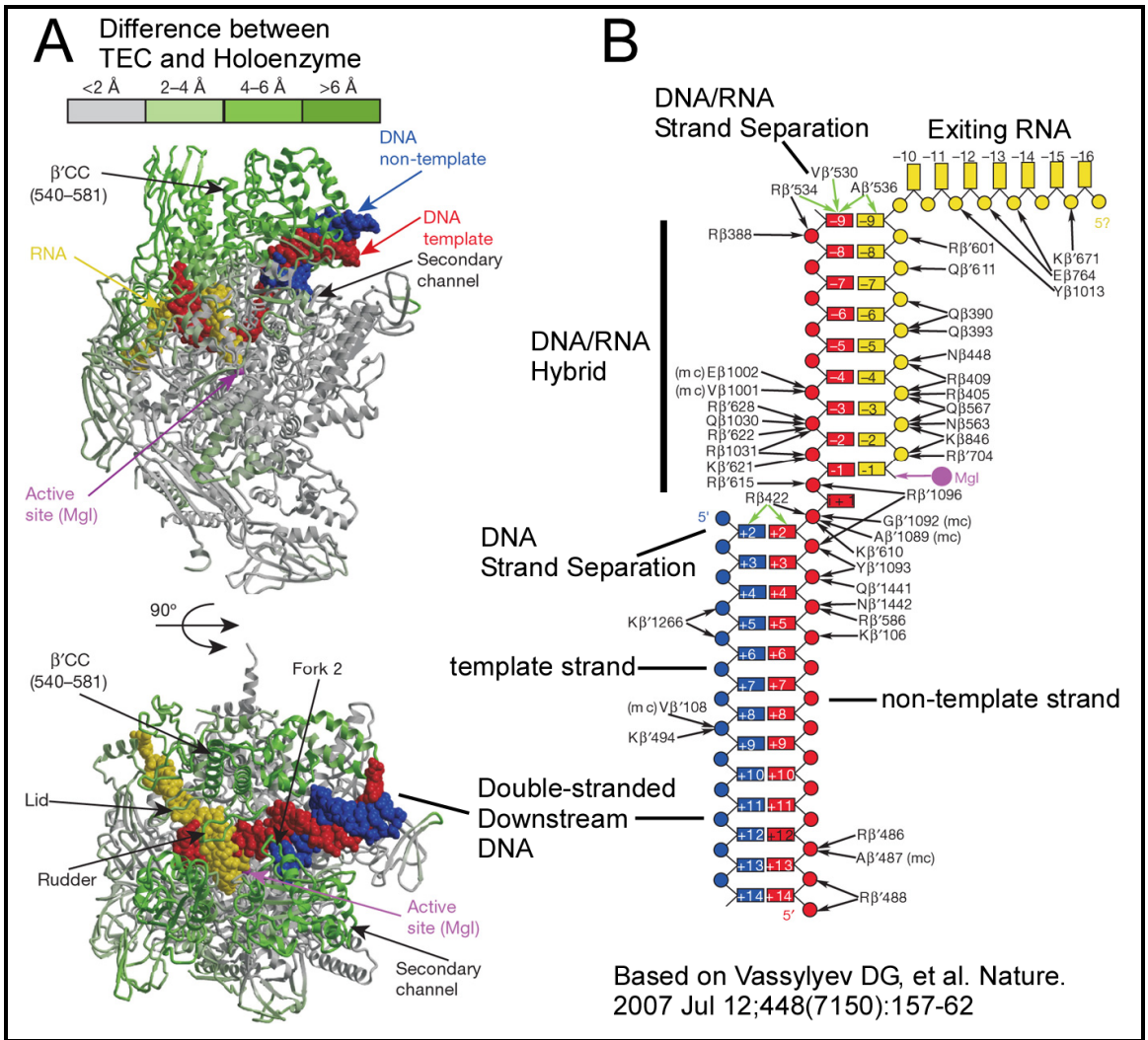
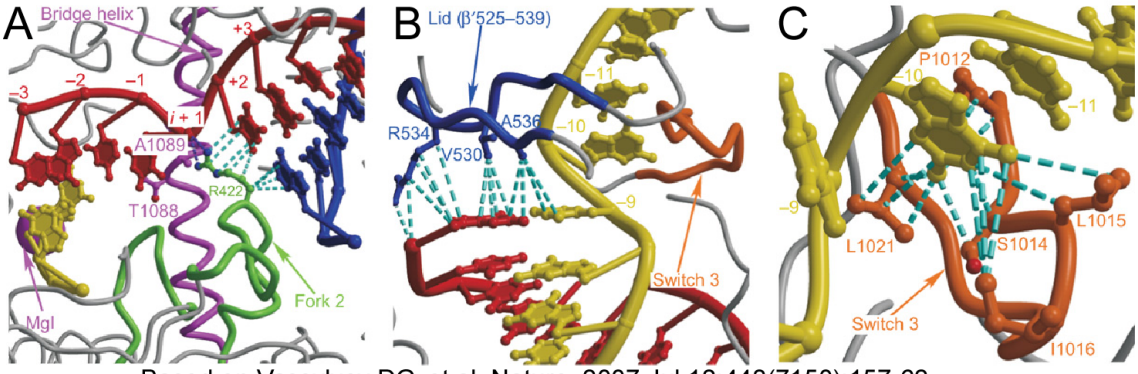


Figure 1.6 – Overview of Bacterial TEC Structure.

(A) *Tth* TEC structure with core subunits as cartoons with residues colored from gray to green according to their deviation from the holoenzyme structure. The DNA (red and blue) and RNA (yellow) are shown as space filled spheres. **(B)** Schematic of the DNA/RNA with protein interactions.

Within the TEC structures the double stranded upstream and downstream DNA are bent 90° to each other, due a kink at the downstream DNA/RNA hybrid [8]. Based on structural modeling using nucleic acid to protein cross linking, the intervening single stranded DNA (12-14 bp) forms a transcription bubble with the template-strand forming a 7-9 bp DNA/RNA hybrid [24]. In the *Tth* TEC structure a 9 bp DNA/RNA hybrid is tightly packed in the active center and positioned with the secondary channel to accept incoming nucleotides [8, 9].



Based on Vassylyev DG, et al. Nature. 2007 Jul 12;448(7150):157-62

Figure 1.7 – Close-up of Bacterial TEC DNA/RNA Interactions.

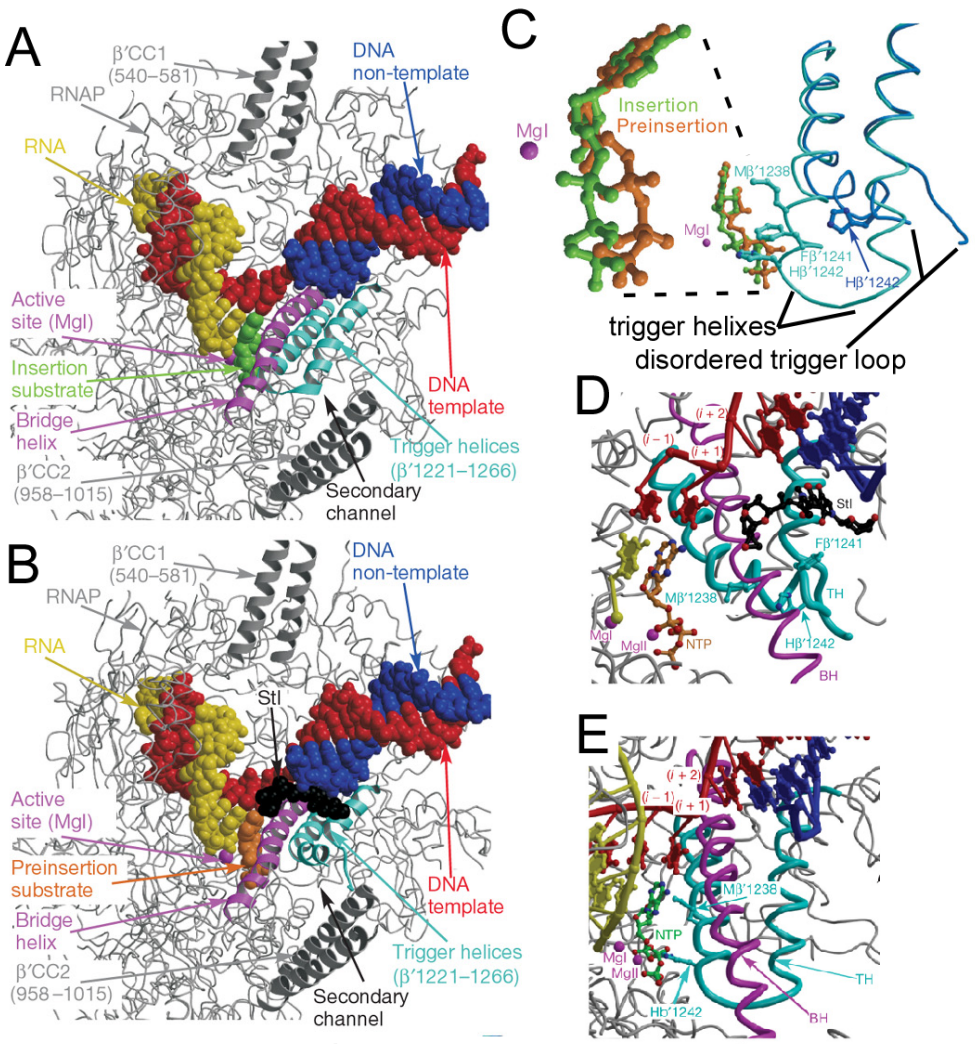
(A) View of *Tth* TEC structure showing β fork double-stranded DNA interaction, which is possibly involved in DNA strand separation. Van der Waals interactions are shown by dashed cyan lines. **(B)** View of *Tth* TEC structure showing β' lid DNA/RNA hybrid interaction, which is possible involved in DNA/RNA strand separation. **(C)** View of *Tth* TEC structure showing β Switch 3 hydrophobic pocket interacting with the first displaced RNA base.

Although the TEC structures did not contain the upstream double-stranded DNA (where the DNA strands re-anneal), it is thought that the upstream edge of the transcription bubble may be stabilized by interactions with the β' lid and zipper domains and the β-flap. Furthermore, even though the downstream edge of the transcription bubble (where the DNA strands separate) is not completely visualized in the TEC structure, the β fork sterically blocks the edge of the non-

template DNA strand and therefore might have a role in DNA strand separation (Figure 1.7A). The rudder loop is positioned between the downstream double-stranded DNA and the DNA/RNA hybrid and thought to play a role in stabilizing the transcription bubble [8, 25].

The RNA strand moves past the β' rudder and lid to exit through the RNA exit channel under the β-flap [8]. As shown in Figure 1.7B, the β-lid forms base stacking interactions with the upstream edge of the DNA/RNA hybrid [8]. The β-lid also sterically blocks the path of the DNA/RNA hybrid possibly aiding strand separation between the upstream template-strand DNA and exiting RNA [8]. In addition the first displaced RNA base is trapped in a hydrophobic pocket formed by the β Switch 3 (Figure 1.7C), which might prevent re-annealing to the DNA [8]. It has also been shown that DNA re-annealing of the upstream double-stranded DNA plays a role in DNA/RNA strand separation [8, 26, 27].

The recent TEC substrate structures (Figure 1.8A,B) [9] have revealed important details about the mechanism of bacterial RNAP nucleotide addition (Figure 1.9). It is thought that bacterial RNAP exists between pre- and post-translation states and that the transition from one to the next is driven by thermal motion. The nucleotide addition cycle starts with the post-translocated state with the incoming NTP enter the secondary channel and binds to RNAP in an open catalytically inactive pre-insertion conformation. It is thought that the incoming NTP acts as a ratchet by stabilizing the post-translation state [9]. In addition, the pre-insertion state acts as an initial substrate specify sieve [9].



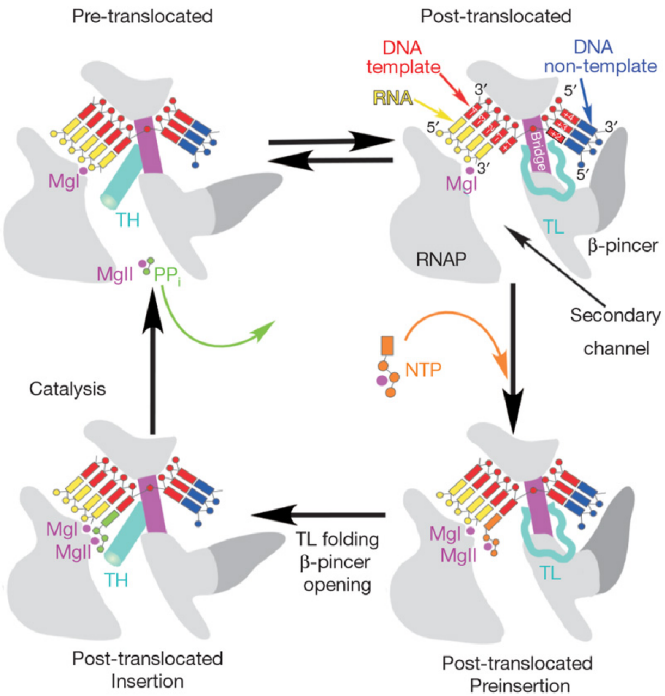
Based on Vassylyev DG, *et al*. Nature. 2007 Jul 12;448(7150):163-8.

Figure 1.8 – Overview of Bacterial TEC NTP Substrate Structures.

(A) Structural view of *Tth* TEC with a non-hydrolysable substrate analogue AMPcPP (green), in which an inactive pre-insertion conformation. **(B)** Structural view of *Tth* TEC with AMPcPP (orange) and the antibiotic streptolydigin (Stl), which is in a catalytically active insertion conformation. **(C)** Close-up view of the NTP substrate analogue in the insertion (green) and pre-insertion (orange). Along with a zoomed out view showing the transition between the trigger loop with gap where disordered to ordered and folded trigger helixes. **(D)** Pre-insertion conformation with ordered part of the trigger loop (TL; cyan). **(E)** Insertions conformation with folded trigger helixes (TH; cyan) and MgI and Mg II properly orientated for catalysis.

In order for catalysis to occur the disordered β' trigger loop (Figure 1.8D) must fold into two anti-parallel trigger helixes (Figure 1.8E), which interact with the NTP substrate orientating (Figure 1.8C) it into a catalytically active closed

insertion conformation [9]. In addition, the folding of the trigger helixes constricts the secondary channel, possibly hindering the dissociation of the inserted NTP while preventing competition from other incoming NTPs [9]. Furthermore, the inherent instability of the trigger loop helices and the opening of the downstream DNA claws mean that the closed insertion conformation might be a high energy state, possibly acting as a second substrate specify sieve [9]. The movement of the NTP into the insertion confirmation also properly orientates the second active site Mg^{2+} (MgII bound by the incoming NTP) with MgI, allowing for nucleotide addition and pyrophosphate release [9]. The subsequent loss of interactions between the trigger helixes and the NTP, destabilize and unfold the trigger helixes back into the trigger loop, thus completing the nucleotide addition cycle.

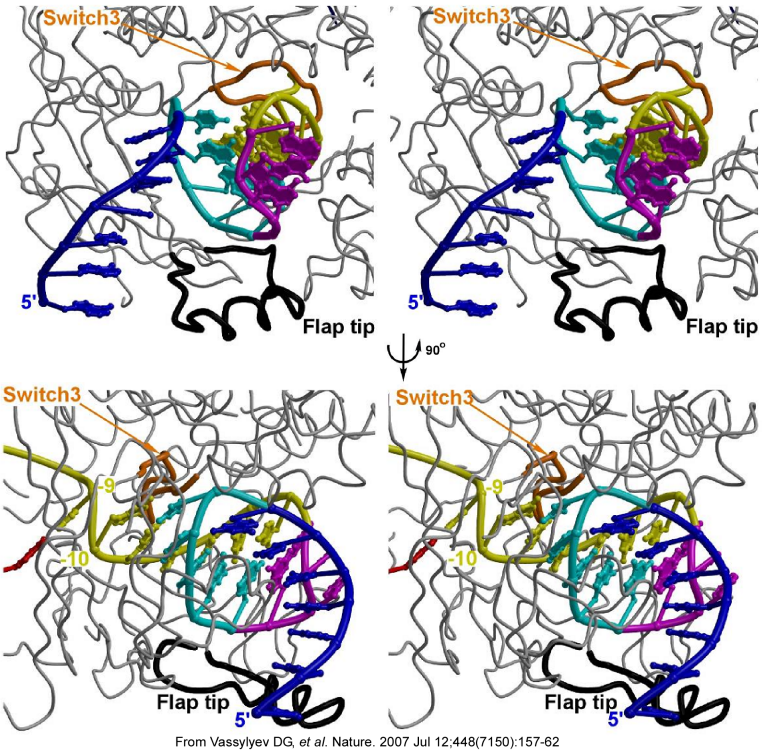


From on Vassylyev DG, *et al.* Nature. 2007 Jul 12;448(7150):163-8.

Figure 1.9 – Bacterial Nucleotide Addition Cycle.
Schematic view of bacterial catalysis with the Trigger Loop (TL) as a cyan line, Trigger Helixes (TH) as a blue cylinder.

Bacterial RNAP Transcription Termination

Upon termination elongating RNAPs disengage from the DNA and release their RNA transcripts. In general there are two types of termination: (1) rho dependent and (2) rho independent or intrinsic termination. In rho dependent termination the ring shaped hexameric rho protein uses ATP hydrolysis to generate a force sufficient to displace the RNA, thus destabilizing the elongating RNAP [28]. On the other hand, intrinsic termination is governed by bulky RNA hairpins, generated by termination signals located at the end of transcriptional units, generally with a DNA sequence containing a GC-rich dyad symmetry element followed by thymine bases [2].



From Vassylyev DG, *et al.* Nature. 2007 Jul 12;448(7150):157-62

Figure 1.10 – Bacterial Paused Hairpin Modeling.

Stereo views of a modeled pause inducing RNA hairpin within the RNA exit channel of the *Tth* TEC structure with the β-flap tip (black), Switch 3 (orange). The RNA has a 5 bp RNA duplex starting at the -11 register with the 5' portion (cyan), 3' portion (yellow), intervening 4 bp loop (magenta), and exiting 5' RNA (blue).

Interestingly, depending on the starting RNA register (Figure 1.6B) a RNA hairpin can either result in a pause (-11/-10) allowing for the interaction of transcriptional regulators or lead to termination (less than -8) [8]. Modeling using the *Tth* TEC structure has shown that the paused hairpin could be accommodated in the RNA exit channel without major structural rearrangements [8]. However, modeling of a termination hairpin in the *Tth* TEC structure was not possible, since it would require major structural rearrangements [8]. Interestingly, these structural movements might involve the lid, coil-coil, and rudder which would destabilize the transcription bubble and lead to termination [8].

Chapter 2 - The Structural Basis for Promoter -35 Element Recognition by the Group IV Sigma Factors*

Introduction

Like many ECF σ factors, *Ec* σ^E is regulated by an anti-σ, RseA [16, 18]. Under normal conditions RseA inactivates σ^E by sequestering it at the cytoplasmic face of the inner-membrane (Figure 2.1). However, when environmental stresses lead to the accumulation of unfolded proteins in the periplasm, a series of proteolytic cleavage reactions release σ^E from RseA [29]. The free σ^E is then able to bind RNAP core and drive the transcription of a core set of genes conserved across most bacteria, as well as a more variable set of species specific genes [30]. The core genes coordinate the assembly and maintenance of the bacterial outer membrane. Many of the variable σ^E regulon members are critical for virulence in important pathogens [31-34]. σ^E also promotes transcription of an operon consisting of itself and three regulatory genes (RseA/B/C).

The structure of *Ec* σ^E bound to the cytoplasmic portion of its anti-σ RseA revealed that, despite little primary sequence identity, domains 2 and 4 of σ^E (σ^E_2 and σ^E_4, respectively) share striking structural similarity to the corresponding domains of *Taq* σ^A (σ^A_2 and σ^A_4) [35]. Domain 4 of all Group I σ factors contain a helix-turn-helix DNA binding motif which recognizes the 6-bp -35 consensus TTGACA [21, 36], while the equivalent domain in *Ec* σ^E_4 is thought to directly recognize the 7-bp -35 element GGAACTT [30]. Taken together, this suggests

*This work has been published in Lane WJ, Darst SA. PLoS Biol. 2006 Sep;4(9):e269.

that the different groups of σ factors share the same general mechanisms of -35 element binding, but that residue changes on the surface of the recognition helix account for differences in promoter specificity. Previous studies using the Group I σ factor *Taq* σ^A have revealed the molecular details of how domain 4 recognizes its consensus -35 element [21].

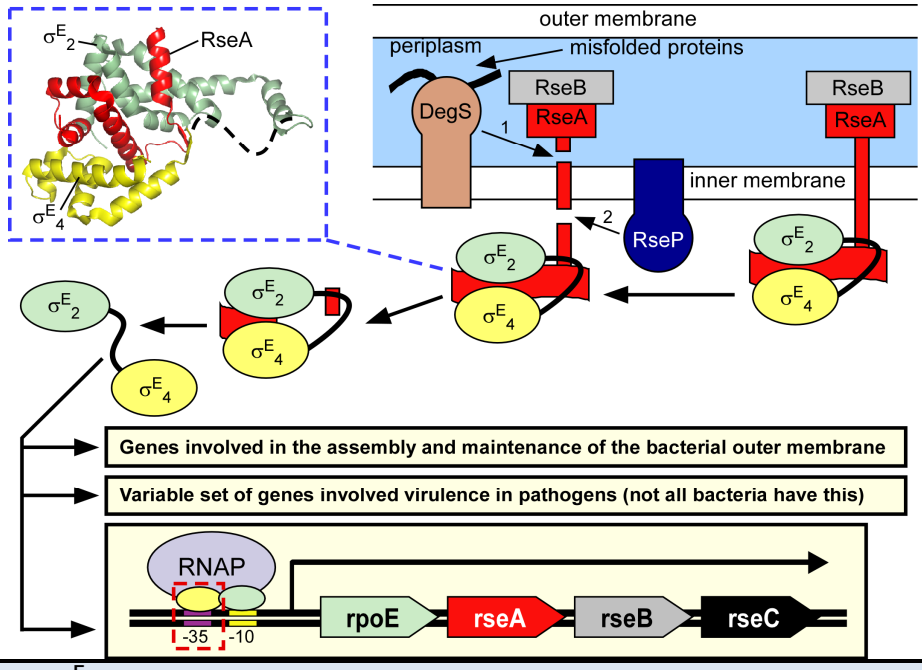


Figure 2.1 – σ^E Regulation.

The prototypical member of the Group IV σ family is *Ec* σ^E. *Ec* σ^E is regulated by binding to an anti-sigma RseA [35]. When environmental stress leads to increased levels of unfolded outer membrane proteins (OMP), a series of proteolytic cleavage reactions cause σ^E to be released from RseA (Reviewed in [29]). σ^E is then free to bind RNAP core and drive the transcription.

Results and Discussion

Progress Timeline

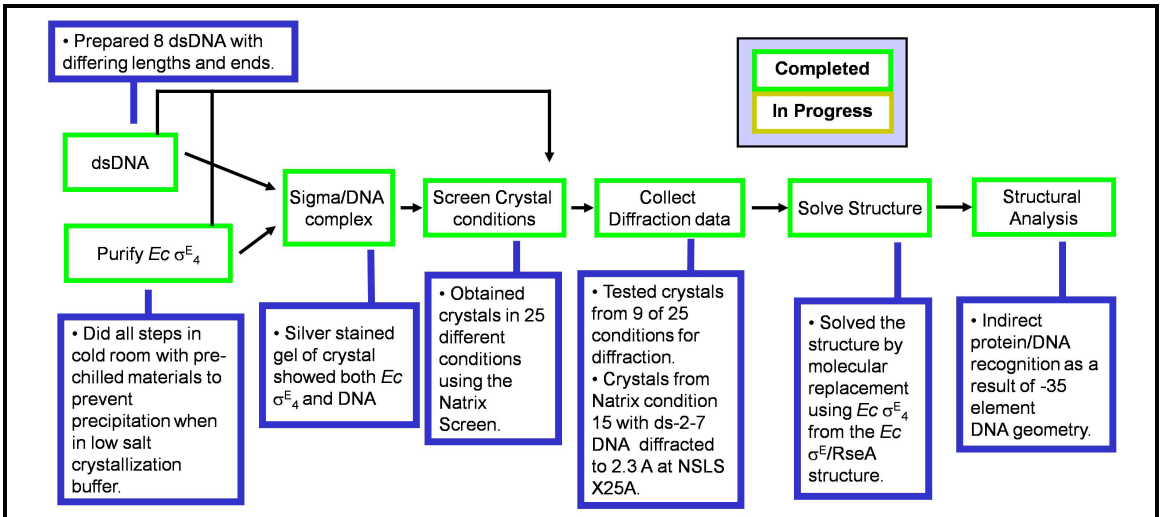


Figure 2.2 – *Ec* σ^E_4 Expression and Purification.

Ec σ^E_4 was expressed and FPLC purified.

Cloning, Expression and Purification

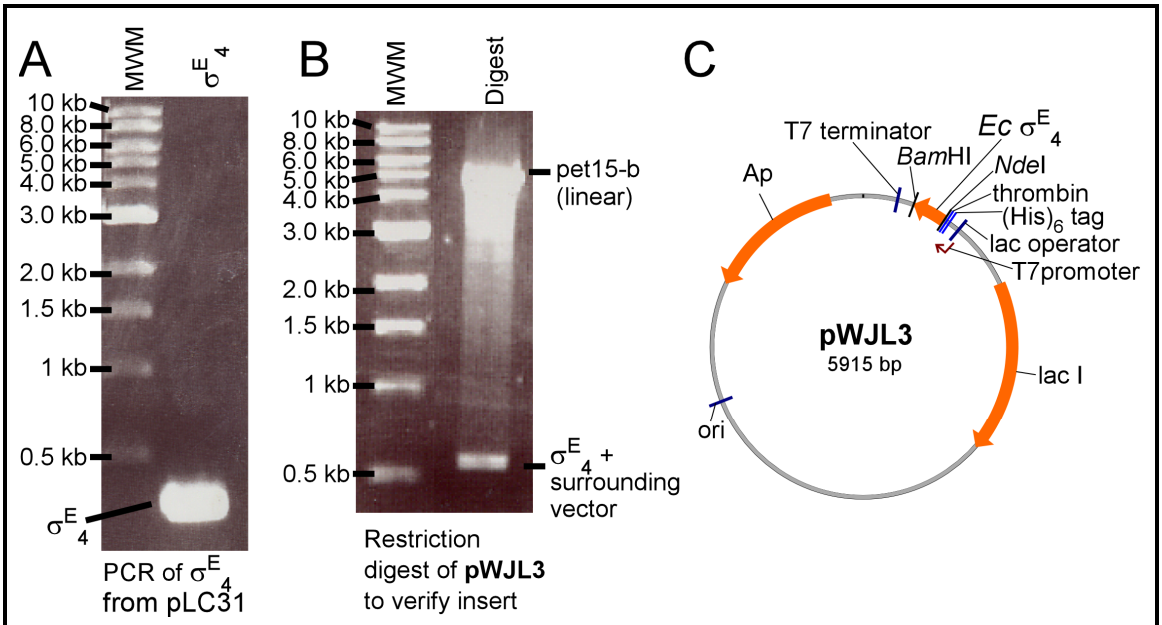


Figure 2.3 – *Ec* σ^E_4 Cloning.

(A) The gene encoding *Ec* σ^E_4 (residues 122-191) was PCR from pLC31 [35]. **(B)** This PCR product from (A) was subcloned into the NdeI/BamHI sites of the pET-15b expression vector (Novagen), creating pWJL3. **(C)** Plasmid map showing important features.

Figure 2.3 shows how the expression vector containing *Ec* σ^E_4 was constructed. Figure 2.4 shows the expression and purification scheme. Please see the Material and Methods (Chapter 6) for more details on the cloning, expression, and purification of *Ec* σ^E_4.

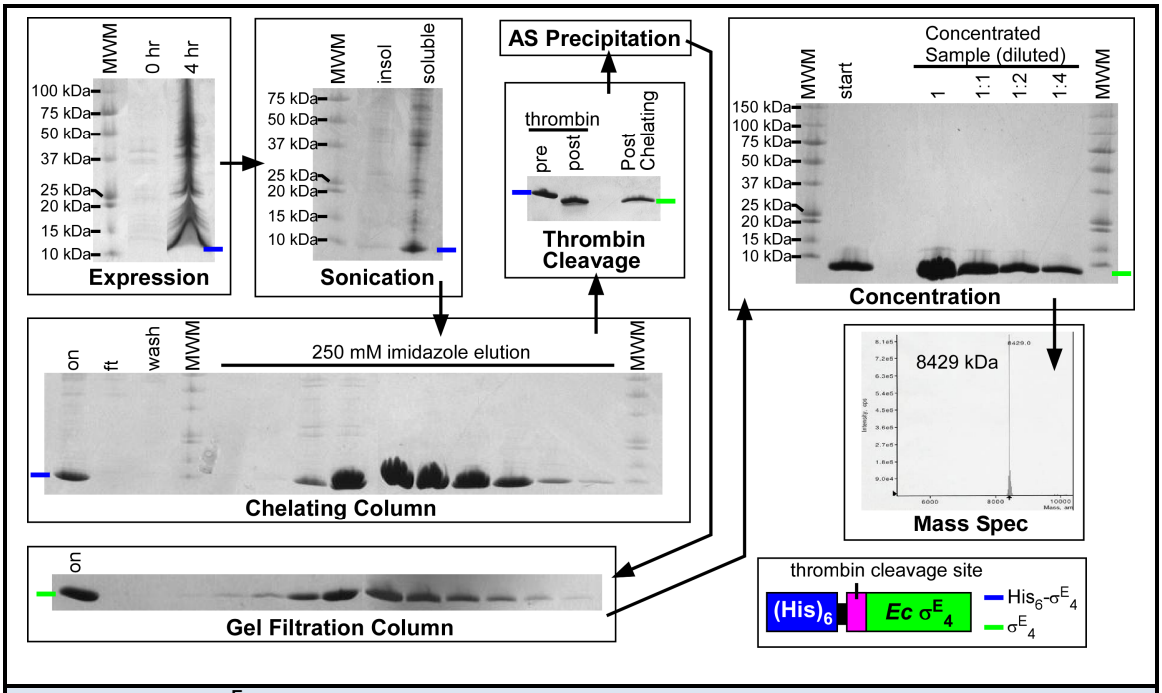


Figure 2.4 – *Ec* σ^E_4 Expression and Purification.

Ec σ^E_4 was expressed and FPLC purified.

Prior to crystallization *Ec* σ^E_4 was exchanged into a low salt crystallization buffer [20 mM Tris-HCl (pH 8), 0.2 M NaCl, 5% glycerol, 0.1 mM EDTA, and 1 mM DTT]. However, in the presence of low salt (<0.3 M NaCl), *Ec* σ^E_4 was extremely susceptible to precipitation at protein concentrations as low as 2 mg/mL (Figure 2.5A). We tried to concentrate the protein in the presence of its cognate -35 element DNA and varied pH, but these had no effect. In the end it was discovered that *Ec* σ^E_4 precipitated in low salt conditions only at room temperature, but it was highly soluble (>30 mg/mL) at 4 °C. The role of

temperature was not initially appreciated, since even if kept cold *Ec* σ^E_4 was so sensitive that it precipitated instantly if all tubes and tips were not pre-chilled. Interestingly, if the precipitation was controlled it could be used to grow crystals. For example, micro-crystals could be grown within seconds by placing a small drop of concentrated *Ec* σ^E_4 in the low salt crystallization buffer on a cover slip and letting it heat to room temperature (Figure 2.5B). The growth of the micro-crystals was sensitive to the size of the liquid drop. Larger crystals could also be grown by leaving eppendorf tubes containing *Ec* σ^E_4 in low salt crystallization buffer at 4 °C for several weeks (Figure 2.5C).

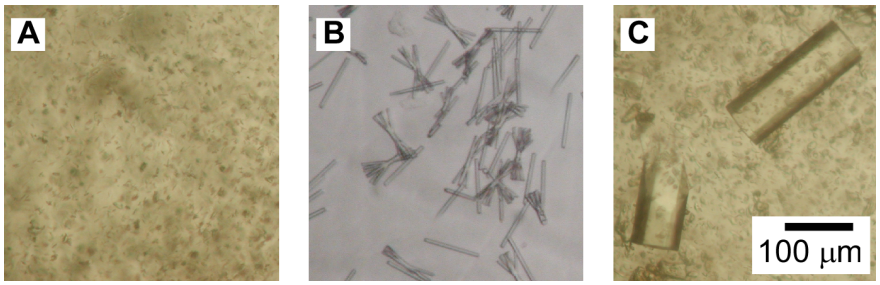


Figure 2.5 – *Ec* σ^E_4 Solubility in Low Salt Conditions.

(A) Magnified view showing a drop of precipitated *Ec* σ^E_4 sample from a tube of concentrated protein in low salt crystallization buffer brought to room temperature.

(B) Magnified view showing a drop of *Ec* σ^E_4 sample that was concentrated successfully at 4 °C, then put on a cover slip and allowed to come to room temperature. The micro-crystals grew within a few seconds.

(C) Magnified view showing a drop of *Ec* σ^E_4 sample that was concentrated successfully at 4 °C and then kept at 4 °C for several weeks in an eppendorf tube.

Crystallization and Structure Determination

We performed vapor diffusion crystallization trials with *Ec* σ^E_4 (residues 122-191) in complex with 8 different double-strand DNA (dsDNA) fragments (Table

2.1) corresponding to the *Ec* σ^E consensus -35 promoter sequence GGAACTT [30].

Table 2.1 – Oligos tested in the crystallization trials.

1st Set			
Oligo	dsDNA	length	Predicted Base Pairing
o-2-1-nt	TCGGAACTTCG (ds-2-1)	11	Watson-Crick
o-2-1-t	GCCTTGAAGCA -> ACGAAGTTCCG	11	
o-2-2-nt	CCGGAACTTCG (ds-2-2)	11	C-GC triple strand Hoogstein
o-2-2-t	GCCTTGAAGCC -> CCGAAGTTCCG	11	
o-2-3-nt	CCGGAACTTCG (ds-2-3)	11	Watson-Crick
o-2-3-t	GCCTTGAAGCG -> GCGAAGTTCCG	11	
o-2-4-nt	TCGGAACTTCA (ds-2-4)	11	Blunt end
o-2-4-t	AGCCTTGAAGT -> TGAAGTTCCGA	11	
o-2-5-nt	TTCGGAACTTCG (ds-2-5)	12	Watson-Crick
o-2-5-t	AGCCTTGAAGCA -> ACGAAGTTCCGA	12	
2nd Set			
Oligo	dsDNA	length	Predicted Base Pairing
o-2-6-nt	CCGGAACTTG (ds-2-6)	10	C-GC triple strand Hoogstein
o-2-6-t	GCCTTGAACC -> CCAAGTTCCG	10	
o-2-7-nt	CCCGGAACTTCG (ds-2-7)	12	C-GC triple strand Hoogstein
o-2-7-t	GGCCTTGAAGCC -> CCGAAGTTCCGG	12	
o-2-8-nt	CCTCGGAACTTCG (ds-2-8)	13	C-GC triple strand Hoogstein
o-2-8-t	GAGCCTTGAAGCC -> CCGAAGTTCCGAG	13	

The crystallization trials using the Natrix Screen (Hampton Research) yielded 25 crystal hits (Figure 2.6). The 1st set of five dsDNA had a fixed DNA length with various ends to promote different base paring interactions. From this 1st set, the ds-2-2 DNA yielded the best crystals; crystals grown in Natrix Screen condition #17 and #44 diffracted to 7 Å and 6.7 Å respectively. Crystals from Natrix Screen condition #44 were crushed and run on an SDS gel which was silver stained to verify the presence of both protein and DNA in the crystal (Figure 2.6). Based on the previous results, a 2nd set of oligos were designed by varying the length of the ds-2-2 DNA.

From the 2nd set, thin rectangular crystals grown (Figure 2.6) using a 12-bp DNA fragment (ds-2-7) in Natrix Screen condition #15 (0.04 M $MgCl_2$, 0.05 M Na Cacodylate pH 6.0, 5% v/v 2-Methyl-2,4-pentanediol [MPD]) diffracted to 2.3 Å-resolution at NSLS X25A (Figure 2.7 and Table 2.2A).

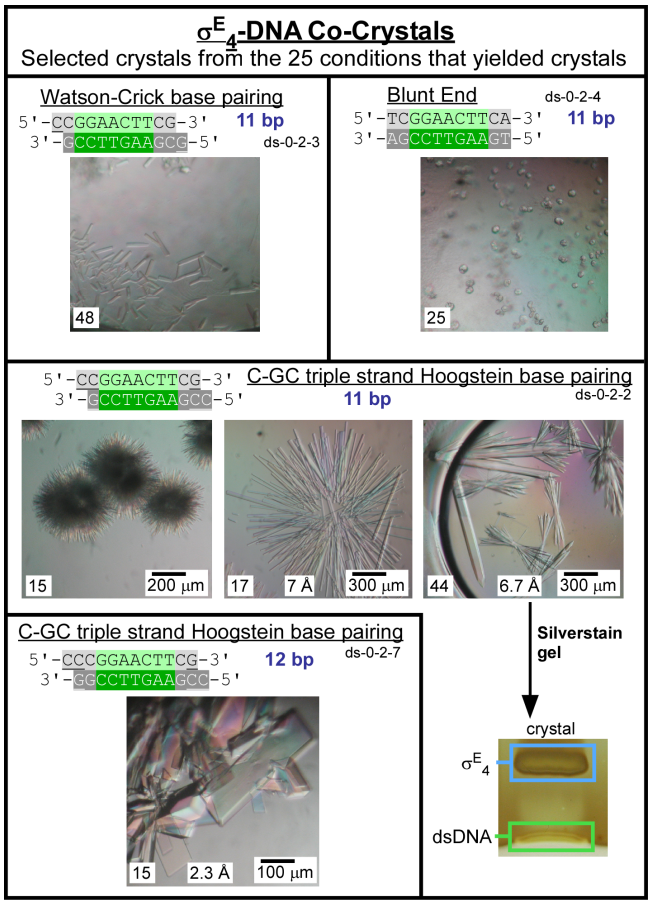


Figure 2.6 – Crystallization of *Ec* σ^E_4/-35 Element DNA.

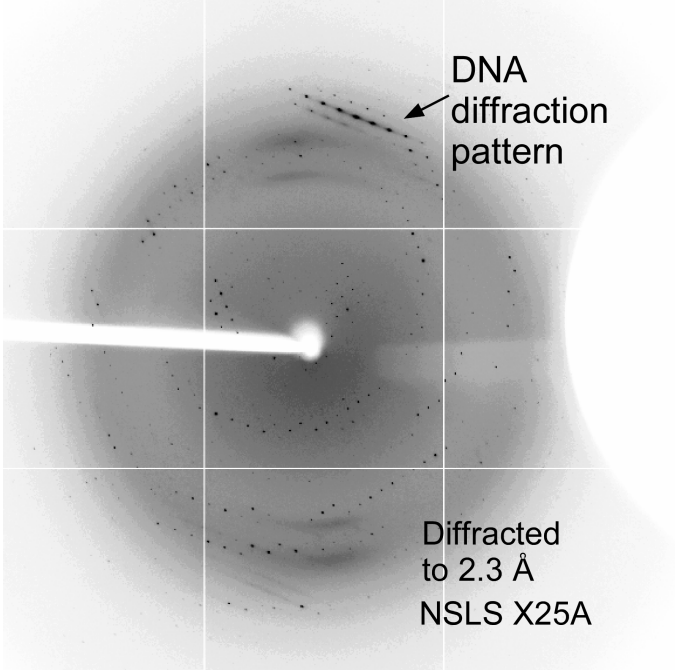


Figure 2.7 – Diffraction Pattern for *Ec* σ^E_4/-35 Element DNA Complex.

The structure was determined by molecular replacement using both a model of *Ec* σ^E_4 from the *Ec* σ^E/RseA complex structure [35] and the 6-bp -35 element from the *Taq* σ^A_4/DNA structure [21] in search models. The crystals contained two σ^E_4/DNA complexes per asymmetric unit, with a solvent content of 65%. Iterative model building and crystallographic refinement converged to an R/R_{free} of 0.241/0.253 (Table 2.2B).

Table 2.2 – X-ray Data Collection and Refinement Parameters.

(A) *Ec* σ^E_4/DNA Diffraction Data

Data Set	Wavelength (Å)	Resolution (Å)	No. of Reflections (Total/Unique)	Completeness (%)	$I/\sigma(I)$	R_{sym}[a] (%)
Native[b]	1.0004	20-2.3 (2.38-2.30)	222,494/19,507	97.5 (96.1)	13.4 (3.8)	5.2 (40.2)

(B) *Ec* σ^E_4/DNA Crystallographic Analysis and Refinement (against native data set)

Space Group	$P2_1$
Unit Cell	a = 55.009 Å, b = 68.709 Å, c = 61.133 Å, α = 90°, β = 101.254°, γ = 90°
Resolution (Å)	20 - 2.3
No. of solvent molecules	136 H_2O
R_{cryst}/R_{free}[c] (%)	24.07/25.28
Rmsd bond lengths	0.009 Å
Rmsd bond angles	1.460°

[a]$R_{sym} = \Sigma|I - \langle I\rangle|/\Sigma I$, where I is observed intensity and $\langle I\rangle$ is average intensity obtained from multiple observations of symmetry related reflections; [b]Dataset was collected at the National Synchrotron Light Source beamline X25; [c]$R_{cryst} = \Sigma||F_{observed}| - |F_{calculated}||/\Sigma|F_{observed}|$, $R_{free} = R_{cryst}$ calculated using 10% random data omitted from the refinement.

Overall Structure

Two σ^E_4 molecules in the asymmetric unit each bound a separate DNA fragment. The crystallographically-related DNA helices packed head-to-tail, forming a pseudo-continuous double helix with the one base-pair overhangs forming Hoogstein base-pairs with the adjacent double helices (Figure 2.8A). Clear electron density could be seen for the entirety of both double-stranded

DNAs (Figure 2.8D, E), excluding the overhanging base at the downstream end of the DNA (Figure 2.8F).

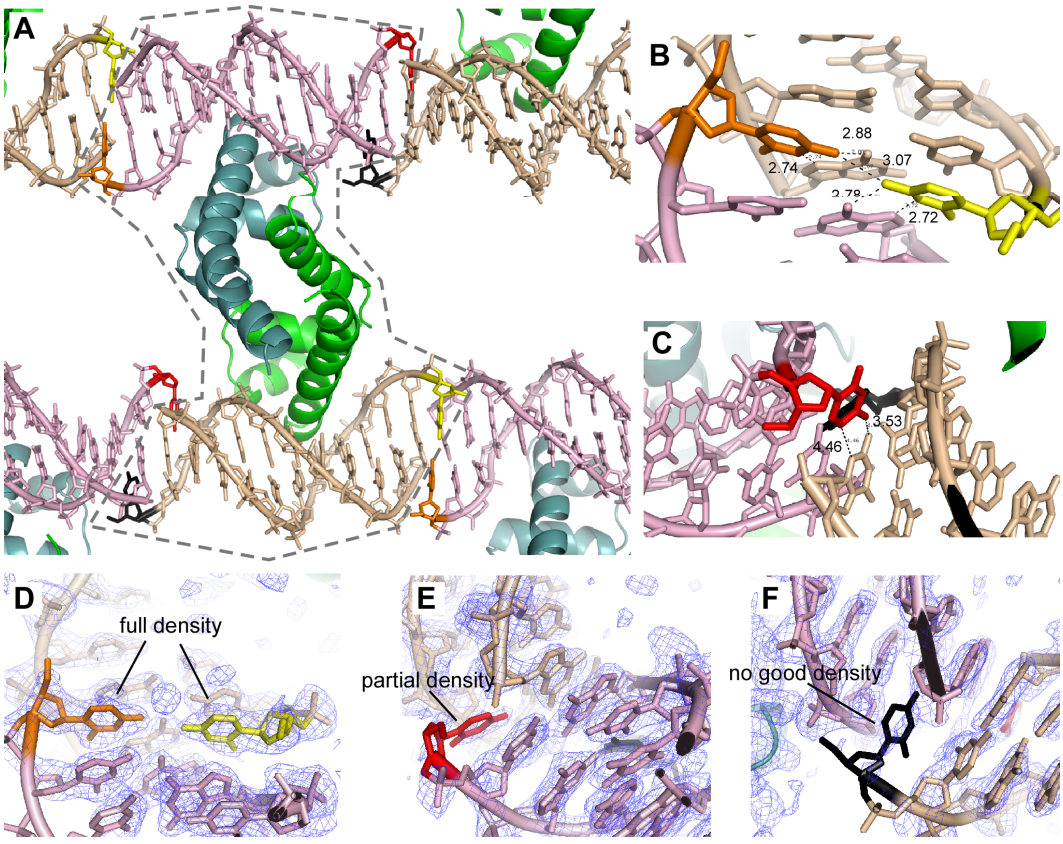


Figure 2.8 – Crystal Packing in the *Ec* σ^{E}_{4}/-35 Element DNA Structure.

(A) Overview of DNA packing. The gray dashed area represents one asymmetric unit.

(B) The yellow and orange C overhang bases form a somewhat distorted Hoogstein like base parings. Instead of paring directly and only with the first double stranded pair the overhangs actually seem to orient towards each other. Nonetheless, the overhangs are within hydrogen bonding distance to the GC double stranded pair.

(C) The red overhang base is out of hydrogen bonding distance, but is in the correct region.

(D) Electron density was fully seen around the yellow and orange bases.

(E) The red base only had partial electron density. So its placement might not be exact.

(F) The black base had no density. The small amount of density that is there was not usable for fitting. In addition no matter what way I placed the black base the Fo-Fc density was not favorable for this base.

As anticipated, the recognition helix of the σ^E_4 helix-turn-helix motif bound in the major groove of the -35 element (Figure 2.9).

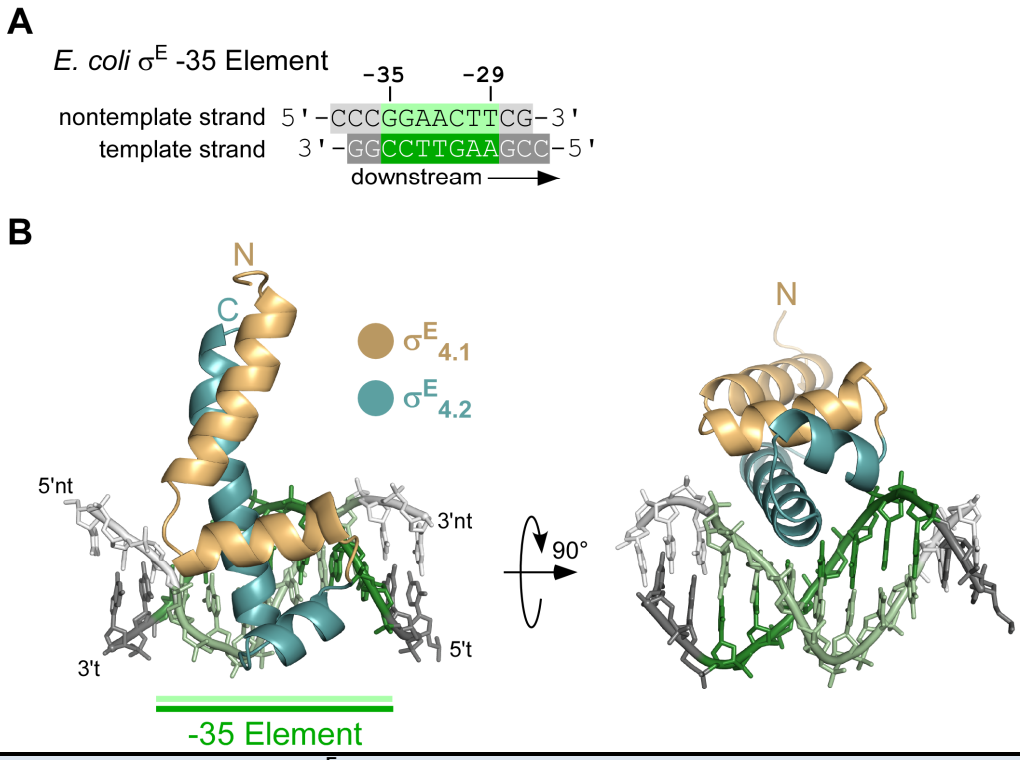


Figure 2.9 – Overview of *Ec* σ^E_4/-35 Element DNA Structure.

(A) Synthetic 12-mer oligonucleotides use for crystallization. The black numbers above the sequence denote the DNA position with respect to the transcription start site at +1. The -35 element is colored light green (nontemplate strand) and dark green (template strand). The flanking bases are colored light grey (nontemplate strand) and dark grey (template strand).

(B) Two views of the *Ec* σ^E_4/-35 element DNA complex, related by a 90° rotation about the horizontal axis as shown. The protein is shown as an α-carbon backbone ribbons, with $\sigma^E_{4.1}$ colored yellow and $\sigma^E_{4.2}$ colored light blue. The DNA is color coded as in (A).

σ^E_4/DNA Interactions

Protein/DNA interactions, which occur exclusively within the major groove, extend from -29 to -36, spanning the entire -35 element as well as one base of upstream DNA (Figure 2.10, Figure 2.11A).

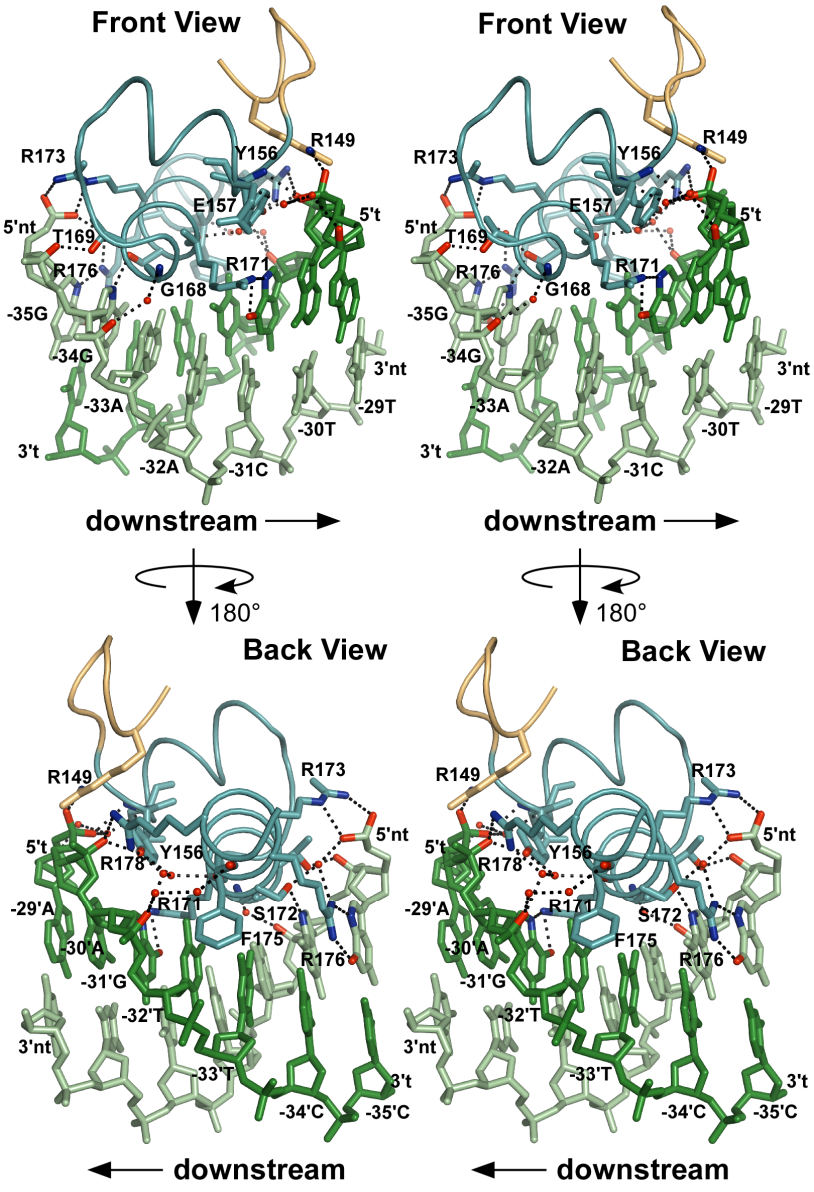


Figure 2.10 – *Ec* σ^{E}_{4}/DNA Contacts; Structural View.

Two stereo views (front and back) of the *Ec* σ^{E}_{4}/-35 element DNA complex, related by a 180° rotation about the vertical axis as shown. The protein is shown as an α-carbon backbone worm, with $\sigma^{E}_{4.1}$ colored yellow and $\sigma^{E}_{4.2}$ colored light blue. Side chains are shown for those residues that make protein/DNA contacts. Carbon atoms of the side chains are colored as the backbone, except atoms involved in polar contacts with the DNA are colored (nitrogen atoms, blue; oxygen atoms, red). The DNA is color-coded as in Figure 2.9A, except atoms involved in polar contacts with the protein are colored (nitrogen atoms, blue; oxygen atoms, red). Water molecules are indicated with red spheres. Dashed black lines indicate hydrogen bonds or salt bridges.

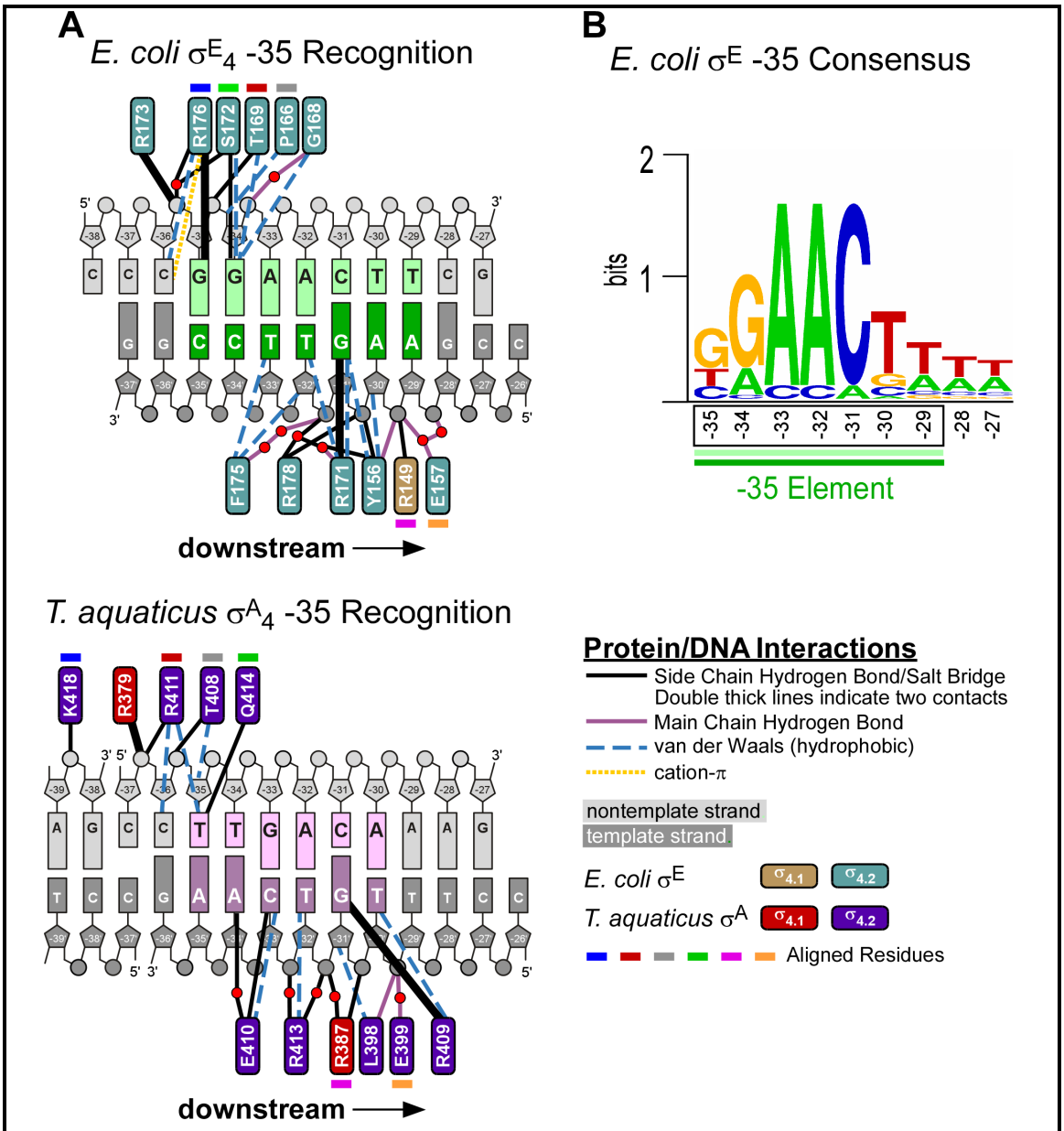


Figure 2.11 – *Ec* σ^{E}_{4}/DNA Contacts; Schematic View.

(A) Schematic representation of σ_4/DNA interactions for *Ec* σ^{E}_{4} (top) and *Taq* σ^{A} (bottom; [21]). The nontemplate / template strand DNA is colored light grey / dark grey (respectively), except the -35 element is colored light green / dark green (for *Ec* σ^{E}_{4}) or pink / magenta (for *Taq* σ^{A}). Colored boxes denote protein residues. Color-coding for the proteins, as well as the meaning of the lines indicating interactions, are explained in the legend (lower right). Double thick solid black lines indicate two hydrogen bonds with the same residue. Water molecules mediating protein/DNA contacts are shown as red circles.

(B) Sequence logo denoting sequence conservation within the *Ec* σ^{E}_{4} -35 element [30, 37].

The protein anchors itself to the DNA by direct and water-mediated side chain and main-chain interactions with the phosphate backbone on the nontemplate strand from -33 to -35 and the template strand from -29′ to -32′ [throughout this chapter, DNA bases will be numbered as in Figure 2.11A, where negative numbers denote base pairs upstream of the transcription start site. Unprimed numbers denote the nontemplate (top) DNA strand, while primes denote the template (bottom) strand]. Specific protein/DNA-base interactions occur through direct hydrogen bonds and van der Waals forces (Figure 2.10, Figure 2.11A). In addition, there is one cation-π interaction between R176 and -36.

Interestingly, the primary base-specific protein/DNA interactions occur at only three positions of the 7-bp -35 element (all Guanines), -35, -34, and -31' (Figure 2.11A). The upstream edge of the -35 element is recognized through a series of hydrogen bonds and van der Waals interactions, mostly between R176 and S172 and the Guanine bases at -35 and -34. R176 forms two hydrogen bonds with the -35G. In addition, R176 forms a cation-π interaction with the -36 DNA base, creating a stair motif along with the -35 hydrogen bonds [38, 39]. S172 forms direct hydrogen bond and van der Waals interactions with the -34G. The protein/DNA-base specific interactions at the -31' position are almost exclusively from R171, which makes two hydrogen bonds and one van der Waals interaction with the -31′G.

In contrast to the numerous base-specific interactions at the -35, -34, and -31' positions, the -33 and -32 positions each contain only one base-specific contact, in the form of van der Waals interactions between the thymidine C5-methyl

groups at -33′ and -32' with F175 and R171, respectively (Figure 2.11A). The structure reveals no base-specific protein/DNA interactions at the -30 and -29 positions.

Geometry of the σ^E_4 -35 Element DNA

At four of the -35 element positions (-33, -32, -30, -29), there are a total of only two protein/DNA-base contacts, both weak, van der Waals contacts (Figure 2.11A). Nevertheless, the -33 and -32 positions are the most highly conserved positions, not only in the *Ec* σ^E -35 consensus, but across all Group IV σ factors with a known promoter specificity (Figure 2.11B, Figure 2.19; [11, 30]). Furthermore, genetic screens for defective transcription due to single nucleotide substitutions in the -35 element of the *Ec* σ^E homologue from *Salmonella enterica* serovar Typhimurium only resulted in the selection of mutants with substitutions at positions -33 and -32 [40]. Therefore, how is it that the most highly conserved and essential positions in the σ^E -35 element are also the same ones that lack strong protein/DNA-base interactions? The answer for this apparent paradox comes from the unique DNA geometry of the σ^E -35 element (Figure 2.12).

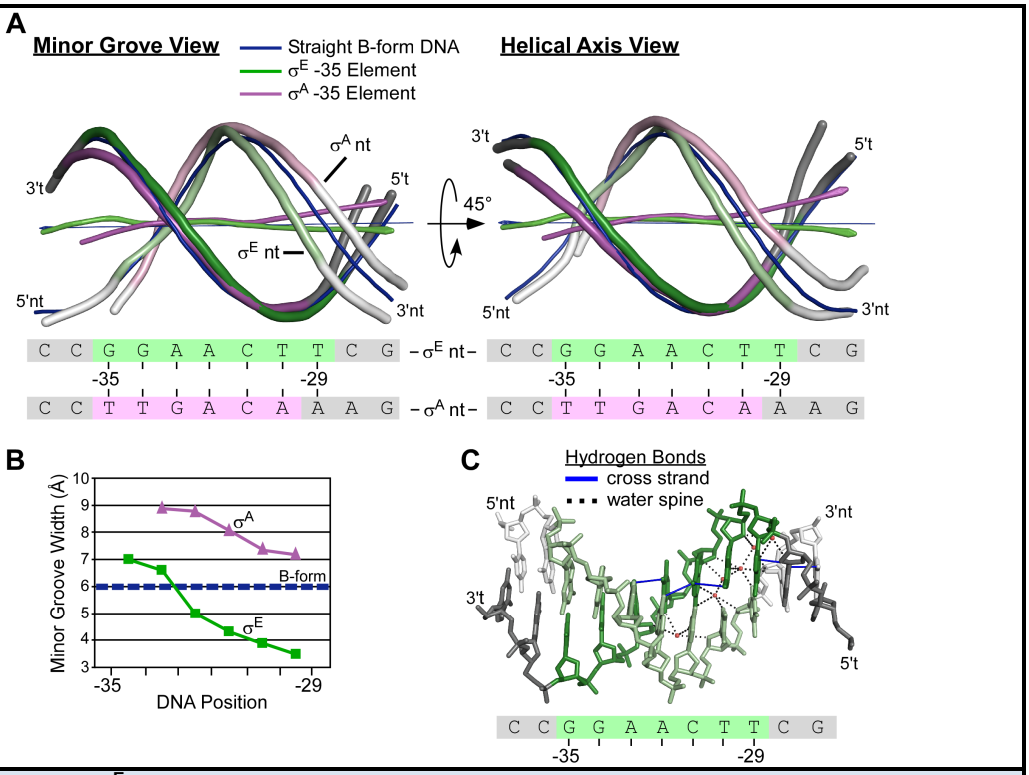


Figure 2.12 – *Ec* σ^E -35 Element DNA Geometry.

(A) Cartoon views of the DNA backbone geometry. The DNA was aligned using the template strand DNA from -35′ to -30′, giving an rmsd of 0.839 over 30 atoms for *Ec* σ^E_4/DNA and *Taq* σ^A_4/DNA. Straight B-form double stranded DNA is blue, *Ec* σ^E -35 element DNA is green, while *Taq* σ^A -35 element DNA is magenta. The paths of the DNA helical axes, calculated using Curves (http://www.ibpc.fr/UPR9080/Curindex.html), are also shown.

(B) Graph showing the DNA minor groove width (calculated using 3DNA [41]) for B-form DNA (blue), *Ec* σ^E_4 -35 element DNA (green), and *Taq* σ^A -35 element DNA (magenta). Minor groove width was calculated as the P-P distance minus 5.8 Å to take into account the radii of the phosphate groups.

(C) View of the hydrogen bonds important in stabilizing the unique geometry of the downstream σ^E -35 element DNA. The waters participating in the spine of hydration are indicated by red spheres. Dashed black lines indicate water-mediated minor groove hydrogen bonds. Dashed blue lines indicate cross strand hydrogen bonds formed between adjacent bases.

The unique DNA geometry induced by oligo(dA)•oligo(dT) tracts, defined by the presence of four to six consecutive A•T bases pairs, is well established [42-46]. Depending on its sequence, oligo(dA)•oligo(dT) tract DNA is rigid and

straight, with a high degree of propeller twist and a very narrow minor groove. Despite not being a true oligo(dA)•oligo(dT) tract as a result of the cytosine insertion at -31, the σ^E -35 element DNA is relatively straight (Figure 2.12A and Figure 2.13C), with a high degree of propeller twist (Figure 2.13A), and the minor groove width begins to narrow at the start of the -33/-32 AA (Figure 2.12B). The narrow minor groove is stabilized by a network of cross-strand hydrogen bonds between adjacent DNA bases, along with a spine of hydration consisting of water-mediated hydrogen bonds between the two strands (Figure 2.12C). The AA at -33/-32 is the most highly conserved feature of the σ^E -35 consensus. After the -31 cytosine insertion, the consensus comprises TT (-30/-29). Furthermore, there is a continued run of two additional conserved T's at -28/-27 (Figure 2.11B; [30]).

Interestingly, the nucleosome structure [47] contains a stretch of DNA (GAAGTT), which closely matches the *Ec* σ^E -35 element from position -34 to -29 (GAACTT) (Figure 2.14). Similar to *Ec* σ^E -35 element DNA, the nucleosome DNA cannot be classified as a typical oligo(dA)•oligo(dT) tracts as a result of the non-A/T base, yet it too displays the hallmark DNA geometry, such as a very narrow minor groove (Figure 2.14B). The presence of similar DNA geometry in two different structural contexts strongly suggests that the oligo(dA)•oligo(dT)-like DNA geometry found in the *Ec* σ^E -35 element DNA complex is an intrinsic property of the DNA sequence and not due to protein induced conformational changes.

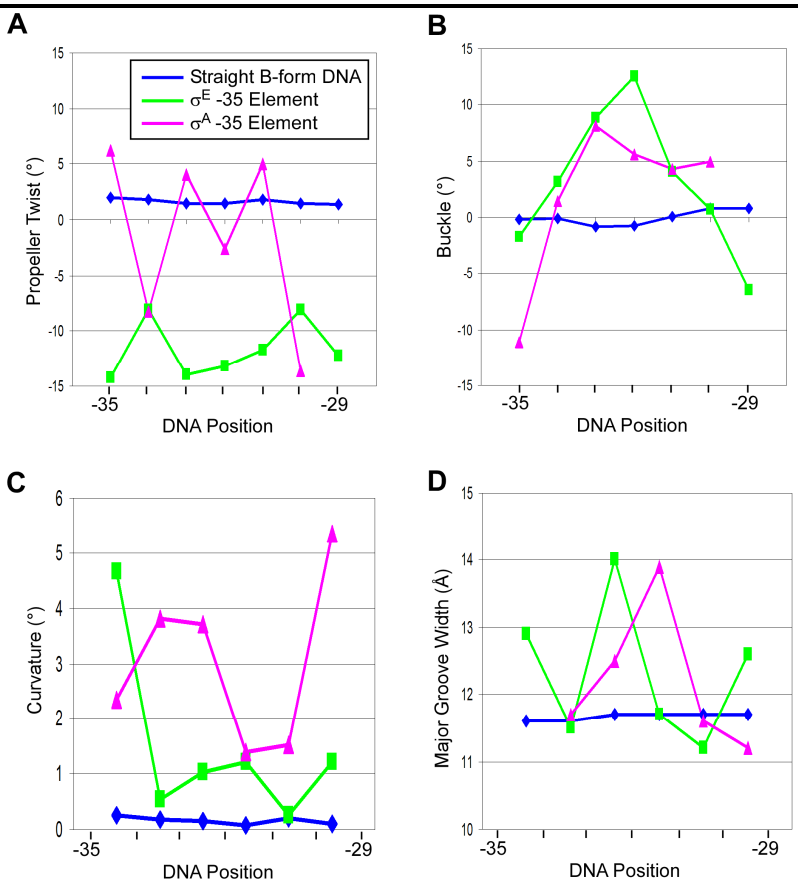


Figure 2.13 – Comparisons of *Ec* σ^E_4 and *Taq* σ^A_4 -35 Element DNA Geometry.
(A) Propeller Twist, **(B)** DNA Buckle, **(C)** Curvature, and **(D)** Major Groove Width calculated using 3DNA.

On its own the absence of strong, base-specific protein/DNA interactions at the -33, -32, and -30 to -27 positions (Figure 2.11A) is conspicuous in light of the high DNA sequence conservation, particularly at the -33/-32 positions (Figure 2.11B). However, combined with the observation that the DNA sequence induces a unique geometry in the -35 element DNA (Figure 2.12), these observations strongly suggest that the DNA sequence is conserved at these positions to enable the global conformation of the DNA, and that this DNA conformation is essential for σ^E_4 binding.

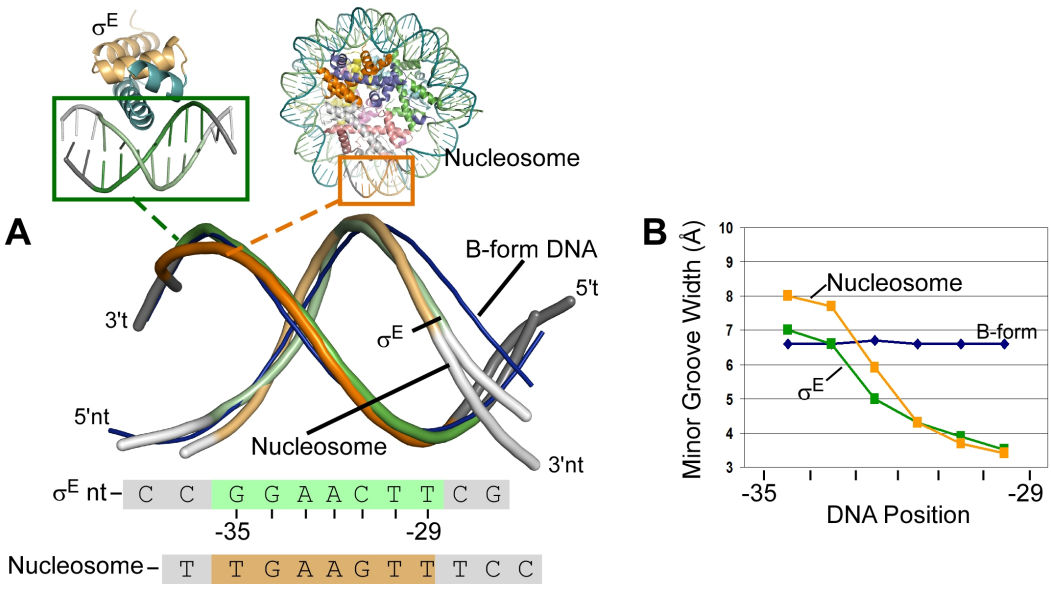


Figure 2.14 – Comparison of *Ec* σ^E_4 -35 Element DNA and nucleosome DNA.

(A) The nucleosome structure (pdb id 1KX4) contains a sequence similar to the *Ec* σ^E_4 -35 Element DNA. Both DNA sequences contain an AA-tract followed by a non-A/T base and then a TT-tract. Despite the non-A/T base, both structures contain narrow minor grooves, which are characteristic of oligo(dA)•oligo(dT) tracts. The DNA structures were aligned using the template strand phosphates. The minor groove narrowing is evident from the location of the non-template strand DNA relative to B-form DNA. The *Ec* σ^E_4 -35 Element DNA is in green and the nucleosome DNA orange.

(B) Graph showing the DNA minor groove width (calculated using 3DNA) for B-form DNA (blue), *Ec* σ^E_4 -35 element DNA (green), and nucleosome DNA (orange). Minor groove width was calculated as the P-P distance minus 5.8 Å to take into account the radii of the phosphate groups.

In this light, the results of the previous genetic screen [40] make good sense. Individual mutations at positions other than the -33 and -32 could be compensated for by both the binding interactions at other -35 element positions and by protein/DNA backbone interactions, which would not be lost at the mutated position. However, substitutions at the -33/-32 positions, which disrupt the highly conserved AA, would in turn disrupt the global DNA geometry necessary for σ^E_4 binding.

Comparison of σ^{E}_{4} and σ^{A}_{4} -35 Element Recognition

Superposition of the DNA from the *Ec* σ^{E}_{4} and *Taq* σ^{A}_{4} [21] -35 element complexes reveals that *Ec* σ^{E}_{4} binds 4 Å further into the major groove than the Group I σ factor *Taq* σ^{A}_{4}, allowing *Ec* σ^{E}_{4} to form more extensive interactions with the DNA (Figure 2.15). In addition, this shift extends the DNA recognition surface of the protein toward the C-terminus of the helix-turn-helix motif recognition helix of *Ec* σ^{E}_{4} (Figure 2.16). For example, even though both promoters have a G at -31′, with *Taq* σ^{A}_{4} it is recognized by R409 and with *Ec* σ^{E}_{4} it is recognized by R171, which is four residues (one helical turn) further towards the C-terminus in the aligned sequences.

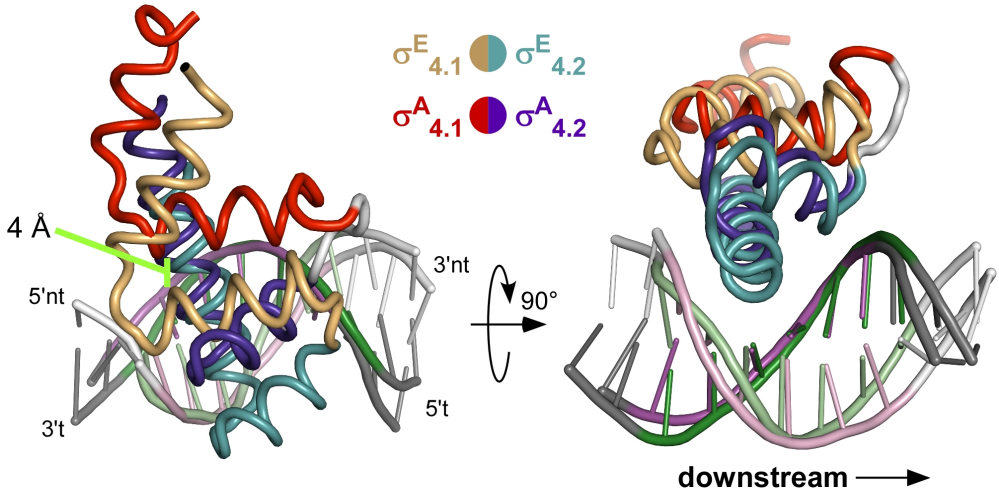


Figure 2.15 – Structural Comparison of *Ec* σ^{E}_{4} and *Taq* σ^{A}_{4} -35 Element Recognition.
Ec σ^{E}_{4}/-35 element DNA and *Taq* σ^{A}_{4}/-35 element DNA complexes were aligned using the template strand DNA from -35′ to -30′, giving an rmsd of 0.839 over 30 atoms. The two views are related by a 90° rotation about the horizontal axis as shown. Proteins are shown as α-carbon backbone worms, color-coded as shown. The *Ec* σ^{E} -35 element DNA is colored light green (nontemplate strand) and dark green (template strand). The *Taq* σ^{A} -35 element is colored pink (nontemplate strand) and magenta (template strand).

Furthermore, the aligned residues *Taq* σ^{A}_{4} K418 and *Ec* σ^{E}_{4} R176 contact the DNA at different positions. Whereas *Taq* σ^{A}_{4} K418 makes contacts upstream of the *Taq* σ^{A} -35 element at -38, *Ec* σ^{E}_{4} R176 forms many important interactions within the σ^{E}_{4} -35 element at -35. Interestingly, *Taq* σ^{A}_{4} makes one van der Waals and four hydrogen bond protein/DNA contacts upstream of the -35 element at -36 and -38, whereas, *Ec* σ^{E}_{4} only makes one van der Waals and one cation-π interaction with the nearby -36 DNA base. In essence the 4 Å shift causes the regions of *Taq* σ^{A}_{4} that were involved in upstream non-promoter element contacts to be involved in sequence specific -35 element contacts in the *Ec* σ^{E}_{4}/DNA structure. For example, in both structures aligned residues K418/R176 (*Taq* σ^{A}_{4}/*Ec* σ^{E}_{4}), T408/P166, R411/T169, and Q414/S172 make up the majority of the upstream nontemplate strand interactions. However, in the case of *Ec* σ^{E}_{4} they all make interactions within the -35 element at -35 and -34, whereas in *Taq* σ^{A}_{4} they make interactions mostly upstream of the -35 element (-38 to -35). Similarly, the aligned residues R387/R149, L398/Y156, and E399/E157 interact in both structures with the downstream template strand DNA backbone. However, in *Ec* σ^{E}_{4} R149 and E157 make their contacts one to two base pairs further downstream than *Taq* σ^{A}_{4} R387 and E399 (Figure 2.16).

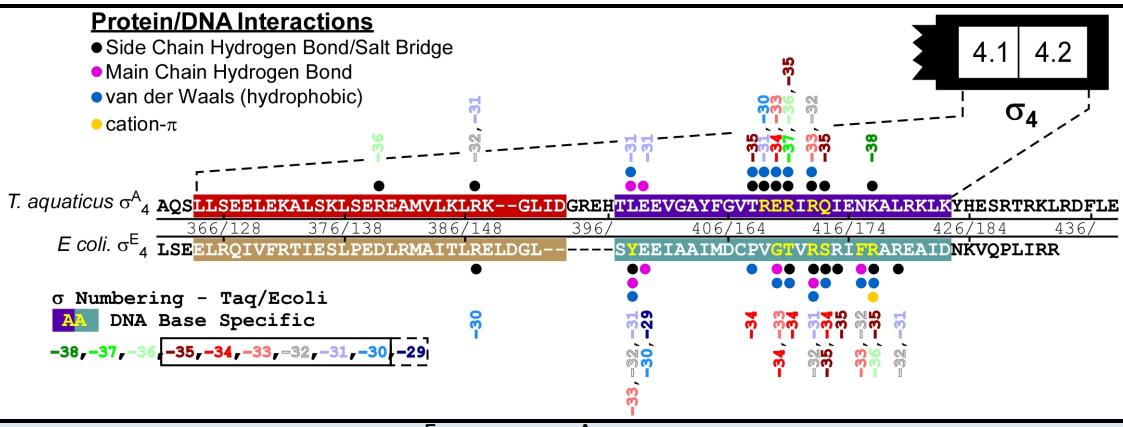


Figure 2.16 – Comparison of *Ec* σ^{E}_{4} and *Taq* σ^{A}_{4} Sequence.

Comparison of the *Ec* σ^{E}_{4} and *Taq* σ^{A}_{4} Protein/DNA interactions. The Cα-backbone of *Ec* σ^{E}_{4} and *Taq* σ^{A}_{4} were aligned using *Ec* σ^{E}_{4} residues 137-150 and 155-182 with *Taq* σ^{A}_{4} residues 375-388 and 397-424, giving an rmsd of 1.00 Å over 42 atoms. Protein residue numbering is shown between the sequences (*Taq*/*Ec*). Residues in $\sigma_{4.1}$ are highlighted in red/yellow (*Taq* σ^{A}/*Ec* σ^{E}) and those in $\sigma_{4.2}$ are colored purple/blue. Red dots denote protein residues that make base-specific DNA contacts. Colored dots denote protein residues that make DNA contacts. Black dots denote hydrogen bonds (<3.2 Å) or salt bridges (< 4 Å) originating from the protein side chain. Magenta dots denote hydrogen bonds originating from the protein main chain. Blue dots denote van der Waals (hydrophobic) contacts (<4.0 Å). Yellow dots denote cation-π interactions. The positions along the DNA that are contacted by each residue are indicated above and bellow the contact circles.

In contrast to the genetic screen for nucleotide substitutions in the σ^{E} -35 element, which only found decreased transcription from mutations at two of the seven promoter positions (-33 and -32; [40]), systematic mutational studies of the *Ec* σ^{70} -35 element have shown decreased transcription from mutations at five of the six promoter positions (-35 to -31; [48]). The two structures also show major differences in the geometry of the -35 element DNA. Whereas *Taq* σ^{A}_{4} bends its -35 element, the protein-bound *Ec* σ^{E}_{4} -35 element DNA is relatively straight (Figure 2.12A). Unlike the σ^{70} -35 element, the *Ec* σ^{E} -35 element itself adopts a unique DNA geometry (described above) that leads to a rigid, straight

DNA segment. In fact, unlike the primary σ factors, which utilize the flexibility of its -35 element DNA, *Ec* σ^E appears to use the rigidity of its -35 element DNA sequence to increase specificity.

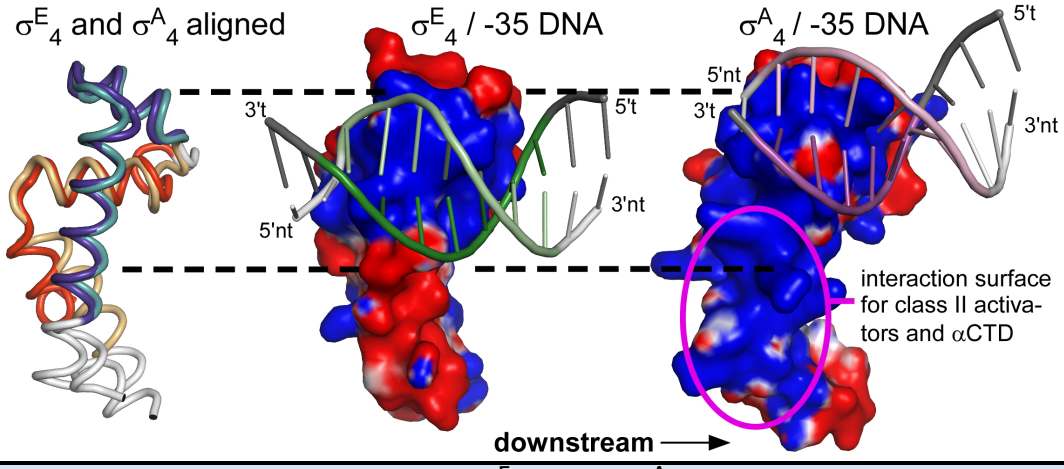


Figure 2.17 – Structural Alignment of *Ec* σ^E_4 and *Taq* σ^A_4.

The protein α-carbon backbones of *Ec* σ^E_4 and *Taq* σ^A_4 were aligned as described in (B). The superimposed proteins, shown as α-carbon backbone worms, are shown on the left, color-coded as in Figure 2.16. The *Ec* σ^E_4/-35 element and *Taq* σ^A/-35 element complexes are shown separately (middle and left, respectively). In these views, the proteins are shown as molecular surfaces, color-coded according to electrostatic surface potential. The DNAs are shown as phosphate-backbone ribbons, with bases indicated schematically as sticks.

Superposition of the proteins from the *Ec* σ^E_4 and *Taq* σ^A_4 -35 element complexes highlights the significant differences in the positioning of the -35 element DNA with respect to the protein, and the different properties of the protein surfaces available for interacting with other proteins bound to the upstream DNA (Figure 2.17). Conserved, basic residues of the Group I σ domain 4 are key targets for interacting with acidic residues of class II transcriptional activators that bind just upstream of the -35 element [21, 49, 50]. The role of transcriptional activators in controlling σ^E transcription is largely unknown.

Implications for -35 element recognition by other Group IV σ factors

The primary sequences of the Group IV σ factors are much more divergent from each other than the members of the other σ^{70}-family subgroups.

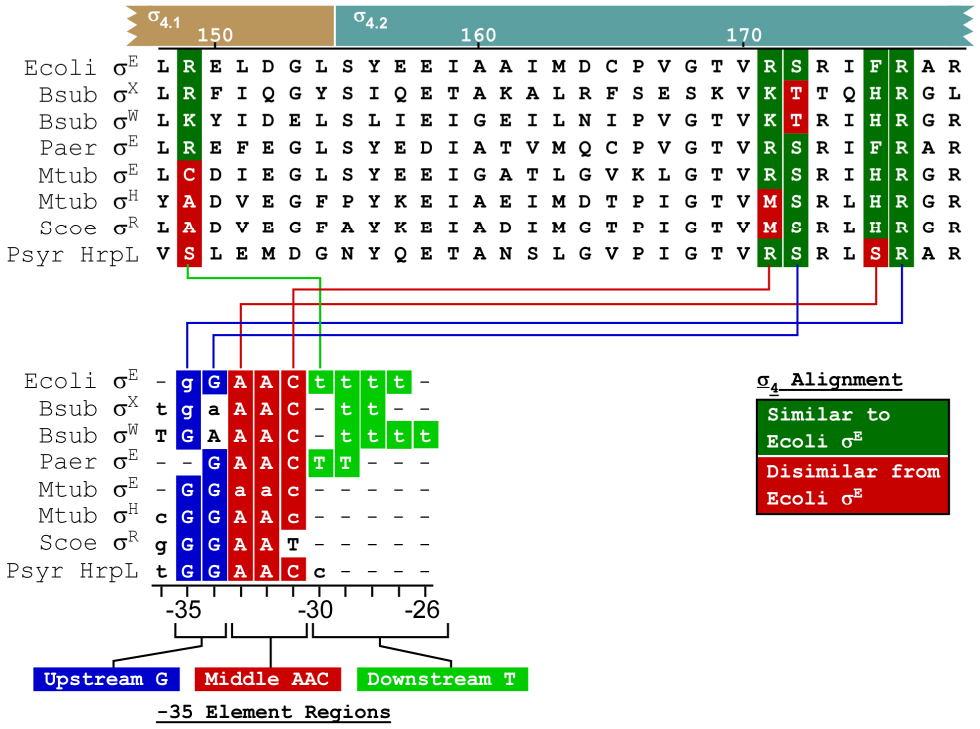


Figure 2.18 – Correlation of σ_4 and -35 element sequences for several Group IV σ factors. The top shows a sequence alignment of the proposed -35 element DNA binding region of several Group IV σ factors. The residue positions that are important in -35 element DNA recognition in the *Ec* σ^E_4/-35 element DNA structure are highlighted green (similar to *Ec* σ^E) or red (dissimilar to *Ec* σ^E). The bottom shows the alignment of the known -35 consensus sequences from several Group IV σ factors. The three -35 element regions are highlighted with the upstream G region (blue), the middle AAC motif (red), and the downstream T rich region (green). Lines connecting the two alignments indicate protein residue/DNA base interactions important for -35 element recognition in the *Ec* σ^E_4/DNA structure.

Furthermore, some genomes contain over 60 Group IV σ factors, each of which can recognize unique, but overlapping, sets of promoter sequences. Nevertheless, the various Group IV σ factors generally share a high degree of conservation in their -35 element sequences, implying that the less conserved

−10 element sequences provide the primary basis for promoter specificity between the different Group IV σ factors, especially within the same species [11, 51, 52]. Therefore, the mechanism of -35 element recognition revealed in the *Ec* σ^{E}_{4}/DNA structure should be relevant to other Group IV σ factors.

Partial to fully characterized regulons have been described for at least eight Group IV σ factors: *Ec* σ^{E} [30], *Bacillus subtilis* (*Bsu*) σ^{X} [53], *Bsu* σ^{W} [54], *Pseudomonas aeruginosa* (*Paer*) σ^{E} [52, 55], *Mycobacterium tuberculosis* (*Mtub*) σ^{E} [56], *Mtub* σ^{H} [57], *Streptomyces coelicolor* (*Scoe*) σ^{R} [58], and *Pseudomonas syringae* (*Psyr*) HrpL [59].

When considering the -35 elements recognized by these Group IV σ factors together, the -35 element can clearly be divided into three distinct regions. The first is an upstream G region, the second is the previously recognized AAC motif [11], and the third is a less well-conserved downstream T-tract (Figure 2.18, Figure 2.19). The differences and similarities between the consensus -35 elements recognized by these Group IV σ factors can be directly explained using the Group IV σ protein sequence alignment and the *Ec* σ^{E}_{4}/DNA structure (Figure 2.15). For example, when consensus sequences for the -35 elements are aligned by the highly conserved AAC motif, all but one of them contain a G at the position equivalent to the *Ec* -35 position. In the structure, this position is recognized by *Ec* σ^{E} R176, which is conserved across all the ECF σ factors. At the -34 position of the promoter consensus, the occurrence of G or A correlates perfectly with the presence of S or T (respectively) at amino acid position 172.

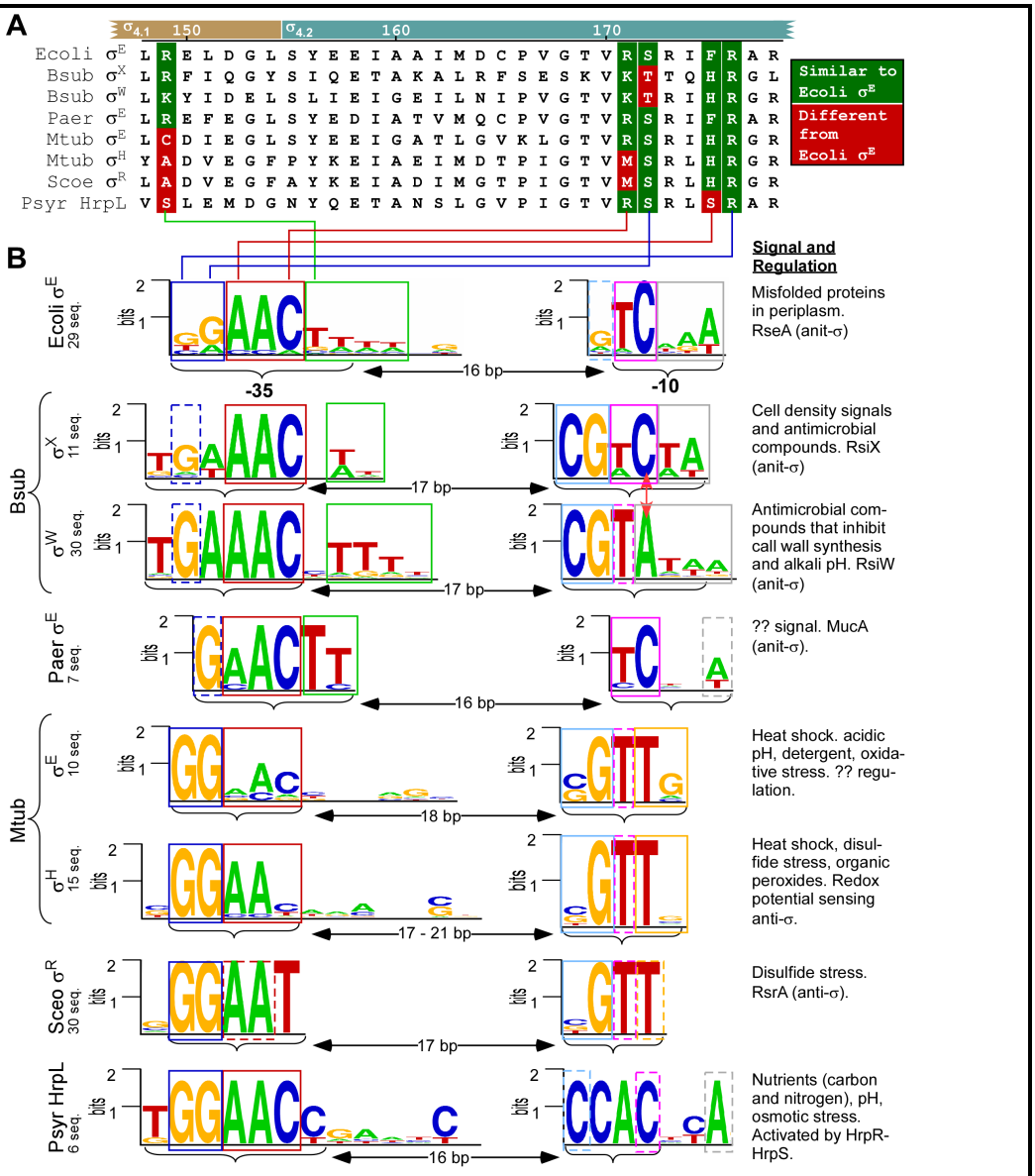


Figure 2.19 – Correlation of σ_4 and -35 element Regulons for several Group IV σ factors. Similar to Figure 2.18, but with the bottom showing the alignment of the known -10 (right) and -35 (left) consensus sequence logos from several Group IV σ factors. The three -35 element regions are highlighted with the upstream G region (blue), the middle AAC motif (red), and the downstream T rich region (green). Lines connecting the two alignments indicate protein residue/DNA base interactions important for -35 element recognition in the *Ec* σ^E_4/DNA structure. Possible regions of similarity within the -10 elements have been highlighted in light blue, magenta, and grey. The single base change thought responsible for the differential gene regulation between *Bsu* σ^X and *Bsu* σ^W is indicated with a red arrow. The column to the right of the sequence logos contains the signal and mechanism of regulation for each σ factor.

In the *Ec* σ^{E}_{4}/-35 element structure, the face of the phenyl-ring of F175 makes van der Waals interactions with the C5-methyl group of the T opposite the absolutely-conserved A at position -33. Consistent with this, all of the ECF σ factors except for *Psyr* HrpL have either an F or an H (which could contribute similar van der Waals interactions) at the equivalent amino acid position.

Amino acid residue R171 of σ^{E}_{4} donates a hydrogen bond to the G opposite the highly conserved C at position -31. Correlating with the conservation of C at this position of the promoter is the occurrence of amino acid residues R or K (which could also donate a hydrogen bond to the complementary G). In the two exceptions, *Mtub* σ^{H} and *Scoe* σ^{R} have M at this amino acid position, and the *Scoe* σ^{R} consensus has a T at this position, while the *Mtub* σ^{H} -35 element has a very weak C/T at this position. Even the downstream T rich sequence, whose primary residue specific interaction is with R149, is found only in the consensus of those σ factors (*Bsu* σ^{X}, *Bsu* σ^{W}, *Paer* σ^{E}) which contain an R or equivalent residue at this position. These correlations suggest that the mechanism of binding found in the *Ec* σ^{E}_{4}/DNA structure can be generalized to other Group IV σ factors.

Conclusions

Despite similar function and structure, the Group I and IV σ factors recognize their -35 elements using distinct mechanisms. The Group IV σ factor *Ec* σ^{E}_{4} binds 4 Å further into the major groove than the Group I σ factor *Taq* σ^{A}_{4}, making more extensive contacts. Unlike *Taq* σ^{A}_{4}, *Ec* σ^{E}_{4} does not bend the DNA. Instead, conserved sequence elements of the σ^{E} -35 promoter induce DNA geometry characteristic of oligo(dA)•oligo(dT)-tract DNA, including pronounced minor groove narrowing. For this reason, the highly conserved AA at -33/-32 is essential for -35 element recognition by σ^{E}_{4}', even in the absence of direct protein interactions with the DNA bases. It appears that these principles of σ^{E}_{4}/-35 element recognition can be applied to a wide range of other Group IV σ factors.

Materials and Methods

Cloning of *Ec* σ^E_4 (pWJL3)

Upstream Primer

5′-GAACCCTGAGAACcatATGTTGTCAGAAGAACTG-3′ ec_sigE_r4_ndeI

Sequence of Full-Length *Ec* rpoE (σ^E)

ATGAGCGAGCAGTTAACGGACCAGGTCCTGGTTGAACGGGTCCAGAAGGGAGATCAGAAAGCCTTTAACTTACTGGT
AGTGCGCTATCAGCATAAAGTGGCGAGTCTGGTTTCCCGCTATGTGCCGTCGGGTGATGTTCCCGATGTGGTACAAG
AAGCTTTTATTAAAGCCTATCGTGCGCTGGATTCGTTCCGGGGAGATAGCGCTTTTTATACATGGCTGTATCGGATT
GCTGTAAATACAGCGAAAAATTACCTGGTTGCTCAGGGGCGTCGTCCACCTTCCAGTGATGTGGATGCCATTGAAGC
TGAAAACTTCGAAAGTGGCGGCGCGTTGAAAGAAATTTCGAACCCTGAGAACTTAATGTTGTCAGAAGAACTGAGAC
AGATAGTTTTCCGAACTATTGAGTCCCTCCCGGAAGATTTACGCATGGCAATAACCTTGCGGGAGCTGGATGGCCTG
AGCTATGAAGAGATAGCCGCTATCATGGATTGTCCGGTAGGTACGGTGCGTTCACGTATCTTCCGAGCGAGGGAAGC
TATTGATAACAAAGTTCAACCGCTTATCAGGCGTTGA

Sequence of E. coli rpoE PCR Product

Nde I

5′GAACCCTGAGAACca|tATGTTGTCAGAAGAACTGAGACAGATAGTTTTCCGAACTATTGAGTCCCTCCCGGAA
GATTTACGCATGGCAATAACCTTGCGGGAGCTGGATGGCCTGAGCTATGAAGAGATAGCCGCTATCATGGATTGTC
CGGTAGGTACGGTGCGTTCACGTATCTTCCGAGCGAGGGAAGCTATTGATAACAAAGTTCAACCGCTTATCAGGCG
TTGAG|GATCC.....T7term-3′

BamH I

pWJL3 Cloning and Expression Region Sequence

- rbs, $(His)_6$, thrombin site, upstream primer, *Ec* σ^E_4, STOP codon, downstream primer

GGTGATGCCGGCCACGATGCGTCCGGCGTAGAGGATCGAGATCTCGATCCCGCGAAATTAATACGACTCACTATAG
GGGAATTGTGAGCGGATAACAATTCCCCTCTAGAAATAATTTTGTTTAACTTTAAGAAGGAGATATACCATGGGCA
GCAGCCATCATCATCATCATCACAGCAGCGGCCTGGTGCCGCGCGGCAGCCAtATGTTGTCAGAAGAACTGAGACA
GATAGTTTTCCGAACTATTGAGTCCCTCCCGGAAGATTTACGCATGGCAATAACCTTGCGGGAGCTGGATGGCCTG
AGCTATGAAGAGATAGCCGCTATCATGGATTGTCCGGTAGGTACGGTGCGTTCACGTATCTTCCGAGCGAGGGAAG
CTATTGATAACAAAGTTCAACCGCTTATCAGGCGTTGAGGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGT
TGGCTGCTGCCACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCT
GAAA

Expression Product

- Molecular weight (PAWS Average Mass): 10308.8 D

MGSSHHHHHHSSGLVPRGSHMLSEELRQIVFRTIESLPEDLRMAITLRELDGLSYEEIAAIMDCPVGTVRSRIFRAR
EAIDNKVQPLIRR

Post Thrombin Cleavage Product

- Molecular weight (PAWS Average Mass): 8426.8 D

GSHMLSEELRQIVFRTIESLPEDLRMAITLRELDGLSYEEIAAIMDCPVGTVRSRIFRAREAIDNKVQPLIRR

Figure 2.20 – Cloning Details for pWJL3.

Cloning of the thrombin cleavable $(His)_6$ tagged *Ec* σ^E_4 expression vector.

We used PCR to generate *Ec* σ^E_4 fragments from pLC31 a pET15-b vector with full-length *Ec* σ^E [35]. The upstream primer was designed to anneal to the start of *Ec* σ^E_4 and introduce a Nde I site for cloning. The upstream primer also introduced an N-term histidine and a silent mutation (GTG to ATG) in second methionine amino acid. A T7 Term primer was used as the downstream primer since pLC31 contains a Bam HI site between the end of the insert and the T7

Term primer. We used the Nde I and Bam HI restriction sites to sub-clone the *Ec* σ^E_4 PCR fragment into a pET-15b expression vector, creating pWJL3 (Figure 2.20). The parent vector for pWJL3 was an ampicillin resistant pET-15b expression vector (Novagen).

Expression and Purification of *Ec* σ^E_4

Figure 2.4 shows the expression and purification scheme. pWJL3 was transformed into *Ec* BL21(DE3)pLysS cells, and transformants were grown at 37 °C in LB medium with amplicillin (100 μg/mL) to an OD_{600} of 0.4-0.6. Protein expression was induced with 1 mM IPTG for 4 hrs. Cells containing the overexpressed protein were harvested and resuspended in lysis buffer [20 mM Tris-HCl (pH 8.0), 0.5 M NaCl, 5% glycerol, 0.1 mM EDTA, 5 mM imidazole (pH 8.0), 0.5 mM β-mercaptoethanol (β-ME), and 1 mM phenylmethylsulfonylfluoride (PMSF)]. Cells were lysed using a sonicator and clarified by centrifugation. Supernatants were applied to 2x5 mL Ni^{2+}-charged HiTrap metal-chelating columns (Amersham Biotech). Lysis buffer with 20 mM imidazole was used to wash the column, followed by elution of the tagged protein using lysis buffer with 250 mM imidazole. To remove the $(His)_6$-tag, samples were diluted into thrombin digestion buffer (20 mM Tris-HCl [pH 8], 0.15 M NaCl, 5% glycerol, 5 mM $CaCl_2$, and 0.5 mM β-ME) and treated with thrombin (500 μg/100 mg protein) at 4 °C. To separate the cleaved (untagged) protein from the thrombin and uncleaved (tagged) the sample was reapplied to the Ni^{2+}-charged HiTrap column in tandem with a 1 mL Benzamidine FF HiTrap column (Amersham), and the flowthrough collected. The sample was then precipitated

using ammonium sulfate (60 g/100 mL sample), centrifuged, and resuspended in gel filtration buffer [20 mM Tris-HCl (pH 8), 0.5 M NaCl, 5% glycerol, and 1 mM DTT]. The resuspended sample was applied to a Superdex 75 gel filtration column (Amersham) equilibrated with gel filtration buffer. The eluted *Ec* σ^E_4 was concentrated to 30 mg/mL by centrifugal filtration (ViaScience) and exchanged into a low salt crystallization buffer [20 mM Tris-HCl (pH 8), 0.2 M NaCl, 5% glycerol, 0.1 mM EDTA, and 1 mM DTT]. Since *Ec* σ^E_4 rapidly precipitated at room temperature when in a low salt buffer (<0.3 M NaCl), all subsequent steps were done in the cold room using pre-chilled supplies. The final purified protein product was aliquoted, flash frozen and stored at -80 °C. Electrospray mass spectrophotometry was used to confirm the mass of the purified product (8427 Da).

Ec σ^E -35 Element Nucleic Acid Preparation

For the purposes of crystallization, several different DNA constructs were designed, based on the *Ec* σ^E_4 -35 consensus. Construct length and flanking bases were varied in an attempt to promote crystallization through end-to-end dsDNA contacts. Lyophilized, tritylated, single-stranded oligonucleotides (Oligos Etc.) were detritylated and purified on an HPLC using a Varian Microsorb 300 DNA column [60]. The purified oligonucleotides were dialyzed into 5 mM TEAB (pH 8.5) and dried on a SpeedVac (Savant). The dried oligonucleotides were resuspended in 5 mM Na cacodylate (pH 7.4), 0.5 mM EDTA, 50 mM NaCl to a concentration of 1 mM. Equimolar amounts of oligonucleotides were annealed by heating to 95 °C for 5 min and then cooling to

22 °C at a rate of 0.01 °C/s. The annealed oligonucleotides were dried in a SpeedVac and stored at -20 °C.

Crystallization and Structure Determination of the *Ec* σ^E_4/DNA Complex

Co-crystals were obtained by vapor diffusion by mixing the duplex DNA (Table 2.1) and *Ec* σ^E_4 (molar ratio 1:1.5) with the final concentration of protein at 1.8 mM (15 mg/mL). The mixture was centrifuged for 30 min, then was mixed with an equal volume of well solution [0.04 M $MgCl_2$, 0.05 M Na-Cacodylate (pH 6.0), and 5% v/v 2-Methyl-2,4-pentanediol (MPD)]. Rectangular crystals (0.3 x 0.1 x 0.06 mm) grew within 5 days. Crystals were prepared for cryocrystallography by soaking in the crystallization solution supplemented with 25% MPD, followed by flash freezing in liquid nitrogen. A native data set was collected to 2.3 Å at The National Synchrotron Light Source (NSLS, Brookhaven National Laboratory, Upton, NY), Beamline X25 (Table 2.2A).

The structure was solved by molecular replacement with Molrep 8.1 [61] using *Ec* σ^E_4 from the *Ec* σ^E/RseA complex structure [35]. Initially, Molrep was used to search for solutions with 2 or 3 molecules per asymmetric unit. Both searches yielded a solution with 2 molecules of *Ec* σ^E_4 arranged in a symmetrical dimer (Molrep Corr=0.252). Though there were some slight clashes between the flexible N- and C-term regions, the crystal symmetry related molecules did not clash and in fact stacked upon one another in one direction. Additionally there was room for the double-stranded DNA. However, when this solution was used to generate an electron density map there was no observable density for the DNA. In an effort to improve the solution, the 2 molecule dimer was used as a search

model to generate a new Molrep solution (Molrep Corr=0.439), which yielded some clear double-stranded DNA density. Molrep was further used to improve the double-stranded DNA density by keeping the *Ec* σ^{E}_{4} dimer fixed and doing two tandem molecular replacement searches using the 6 bp -35 element from the Taq σ^{A}_{4}/DNA structure ([21]; first DNA: Molrep Corr=0.464 and second DNA: Molrep Corr=0.475). In addition to placing the double-stranded DNA into the previously seen DNA density, it extended the density one or two bases past the DNA search model. The solution was further improved by using a one base pair register offset between the two search model DNAs, to generate a 7 bp DNA which was used to do two tandem Molrep molecular replacement searches (first DNA: Molrep Corr=0.469 and second DNA: Molrep Corr=0.487). CNS v1.1 [62] was then used to perform density modification, giving an improved electron density map in which clear density could be seen for the entirety of both double-stranded DNAs, excluding the overhanging base at the downstream end of the DNA. The final DNA was built using a starting template of straight B-form double-stranded DNA corresponding to the crystallization oligos (constructed using Namot2; http://namot.sourceforge.net/). Model building was done using O v9.0.7 [63] and refinement using CNS v1.1 (Table 2.2B).

Protein/DNA contacts were analysed using the program CONTACT. Hydrogen bond and van der Waals contacts were visualized in PyMOL using a custom PyMOL function (show_contacts_man.py), followed by geometric verification using PyMOL v0.98 (http://www.pymol.org). Cation-π interactions were visualized in PyMOL using a custom PyMOL function (show_cation_pi.py)

based on previously determined geometric criteria [38]. DNA geometry was analyzed using 3DNA v1.5 [41] and Curves v5.1 (http://www.ibpc.fr/UPR9080/Curindex.html). Electrostatic surfaces were calculated using APBS: Adaptive Poisson-Boltzmann Solver [64]. All structural figures were prepared using PyMOL.

Chapter 3 - Large Scale Sequence Analysis of the Multi-Subunit RNA Polymerases*

Introduction

Multi-Subunit RNAPs

In all cellular organisms the process of transcription is driven by a large multi-subunit molecular machine, the RNA Polymerase (RNAP) [65]. As mentioned previously, bacteria contain a single DNA-dependent multi-subunit RNAP, comprising five core subunits ($\alpha_2\beta\beta'\omega$) plus an initiation-specific σ factor, which binds to RNAP core to form holoenzyme. Eukaryotes contain three DNA-dependent multi-subunit cellular RNAPs termed pol I/II/III comprising 10 common subunits (Rpb1-3,Rpb5-6,Rpb8-12) plus an additional 4, 2, and 5 subunits respectively [65]. In addition to pol I/II/III, plants contain two additional multi-subunit RNAPs: (1) a cellular pol IV RNAP [66] and (2) an organelle plastid (ie chloroplast) RNAP [67, 68] closely related to cyanobacterial RNAP. Archaea contain only one cellular RNAP composed of 12 subunits, 11 of which are similar to pol II subunits. In general DNA viruses contain single subunit DNA-dependent RNAPs divergent in sequence and structure to the multi-subunit RNAPs found in the other branches of life. However, the Nuclear-Cytoplasmic Large double-stranded DNA Viruses (NCLDV) contain a pol like multi-subunit RNAP presumably acquired from their eukaryotic hosts [69, 70]. The x-ray structures of the multi-subunit bacterial RNAP (Taq) and eukaryotic *Saccharomyces cerevisiae* (*Sce*) pol II RNAP revealed that the two share a high degree of structural similarity [65]. In fact, there are clear homologues for all five of the core bacterial subunits β' / β / α / α / ω which correspond to pol I A190 / A135 / AC40 /

*Paper in Preparation

AC19 / ABC23(Rpb6), pol II B150(Rpb1) / B220(Rpb2) / B44(Rpb3) / B12.5(Rpb11) / ABC23(Rpb6), pol III C160 / C128 / AC40 /AC19 / ABC23(Rpb6), and Archaeal A / B / D / L / K. The central mass of all multi-subunit RNAPs is composed of the two large molecular weight subunits (bacterial β/β') which structurally form two extended pincers around the catalytic center.

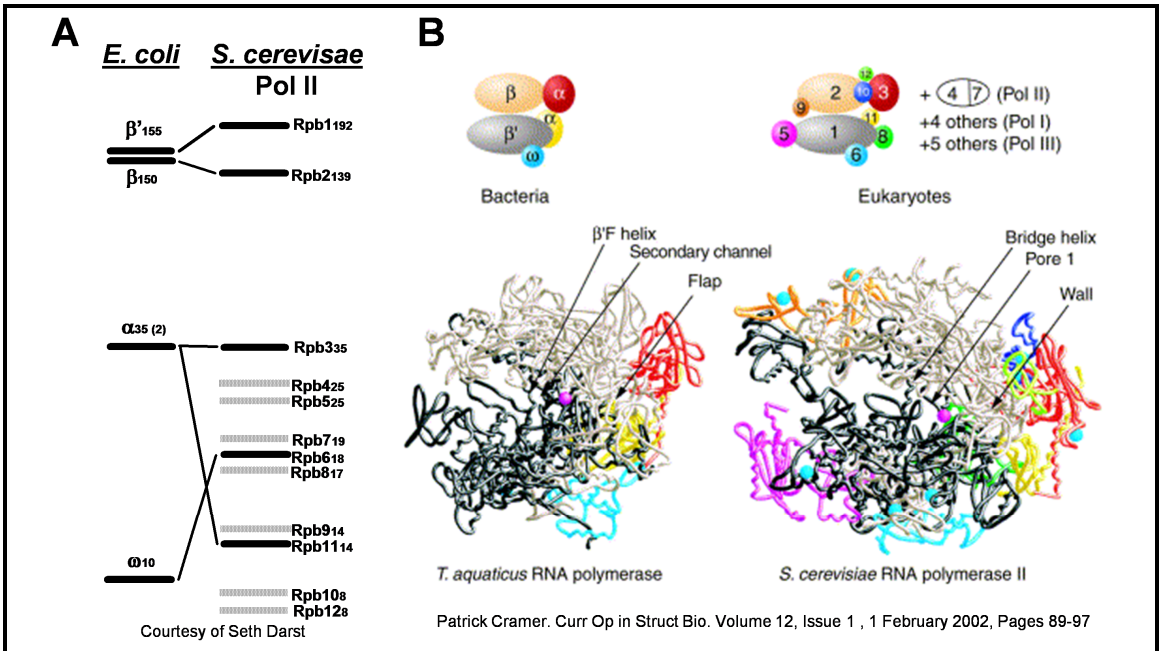


Figure 3.1 – Side by Side Comparison of Multi-Subunit RNAPs.

(A) Virtual gel indicating the molecular weight and identities of the subunits from the bacterial (*Ec*) and eukaryotic (*Sce*) multi-subunit RNAPs. Black lines indicate homologous subunits.

(B) Schematic and structural view of the subunit similarity between the bacterial (*Taq*) and eukaryotic (*Sce*) multi-subunit RNAPs. Subunit homologues are indicated by color.

Multi-Subunit RNAP Shared Sequence Regions

With the availability of the first large subunit sequences it became apparent that the bacterial and eukaryotic RNAPs shared several regions of sequence conservation connected by intervening gaps of non-conversation. In 1987, Sweetser *et al.* defined shared sequence regions (A-I) for the bacterial β subunit and its homologues by aligning the bacterial *Escherichia coli* (*Ec*) β subunit and

its pol II homologue (B150) from *Sce* [71]. Shortly followed by Jokerst *et al.*, which defined shared sequence regions (A-H) for the bacterial β' subunits and its homologues by aligning the *Ec* β' subunit and its pol II/III homologues (B150/C160) from *Sce*, along with the pol II homologues from mouse and *Drosophila melanogaster* [72]. Later these regions were slightly updated with the first x-ray structures of multi-subunit RNAPs [4]. When mapped to the bacterial and yeast RNAP structures the shared sequence regions encompass the inner core of the two large subunits surrounding the active site, presumably in regions that govern aspects of transcription common to all class of multi-subunit RNAPs.

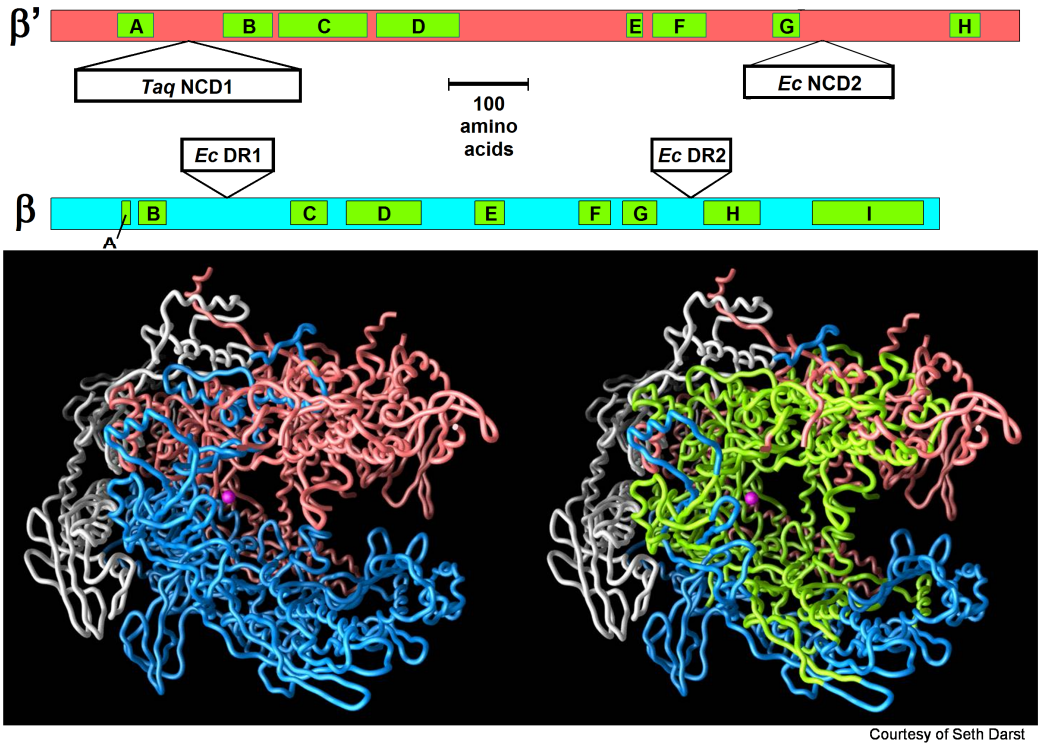


Figure 3.2 – RNAP Shared Sequence Regions and Insertions.

The *Taq* RNAP structure is shown as cartoon worms with β (blue), β' (salmon), α (white). The shared sequence regions (A-H and A-I) are colored green. The catalytic magnesium is shown a purple sphere. The location of some common lineage-specific domain insertions is shown on the protein schematics.

Bacterial RNAP Lineage Specific Domain Insertions

The multi-subunit RNAPs also contain lineage-specific domain insertions which in the case of the bacterial RNAP β and β' subunits can range from 50-500 amino acids. Using a small but diverse set of bacterial sequences, Iyer *et al.* was able to detect and characterize bacterial lineage-specific insertions [73]. They determined that β and β' both contain ubiquitous and lineage-specific insertion domains that fall into four identifiable categories: (1) Zn ribbons, (2) Sandwich Barrel Hybrid Motif (SBHM), (3) β-β' Module 1 (BBM1), (4) β-β' Module 2 (BBM2). The subsequent structures of the two lineage-specific domain insertions *Taq*β′NCD (non-conserved domain) and *Ec*β′GNCD (Region G non-conserved domain) confirmed that both were SBHM domain repeats involved in important protein-protein and/or protein-nucleic acid interactions [74]. The *Taq*β′NCD structure also showed that instead of the predicted 3.5x SBHM domains it actually contained 5x SBHM structural domains, some of which were split and not in sequential order [74].

Results and Discussion

Progress Timeline

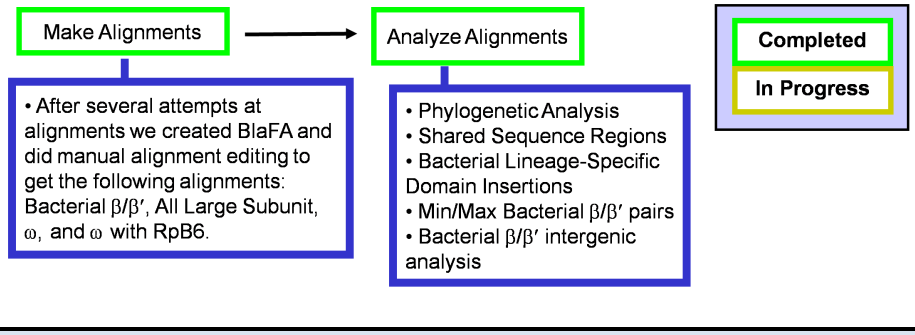


Figure 3.3 – Large Scale Sequence Analysis of the Multi-Subunit RNA Polymerases Progress Timeline.

Above is a schematic of this projects progress. Green boxes are completed parts. Yellow boxes are parts in progress or uncompleted. Black boxes are parts not started. Blue boxes important details for each step.

BLAST to FASTA File to Alignment (BlaFA)

Due to the inherent complexities associated with aligning the bacterial β/β' subunits and their homologues, the process of sequence selection required many steps and special considerations. For example, some sequences needed to be joined since β' is encoded by two gene products in cyanobacteria, archaea (β homologue also), and plastid RNAPs. Some sequences needed to be split since a small number of bacteria, including Helicobacter, have fused β and β' into a single protein product. In addition, there are hundreds of partial β and β' sequences in the NCBI database. Unfortunately, simple sequence gazing was not a practical approach for identifying sequences that need to be joined, split, or removed. The primary reasons for this difficulty are: i) the large number of sequences (~5000-7000 BLAST hits), ii) the intrinsically large size of the large subunits (~1000-2000 amino acids each), and iii) the numerous small and large lineage-specific inserts which can either displace or misalign whole regions.

Therefore, we created an automated approach (Figure 3.4) termed BlaFA, which allowed for custom processing using both taxonomy and sequence patterns as listed in Table 3.2 – Table 3.6.

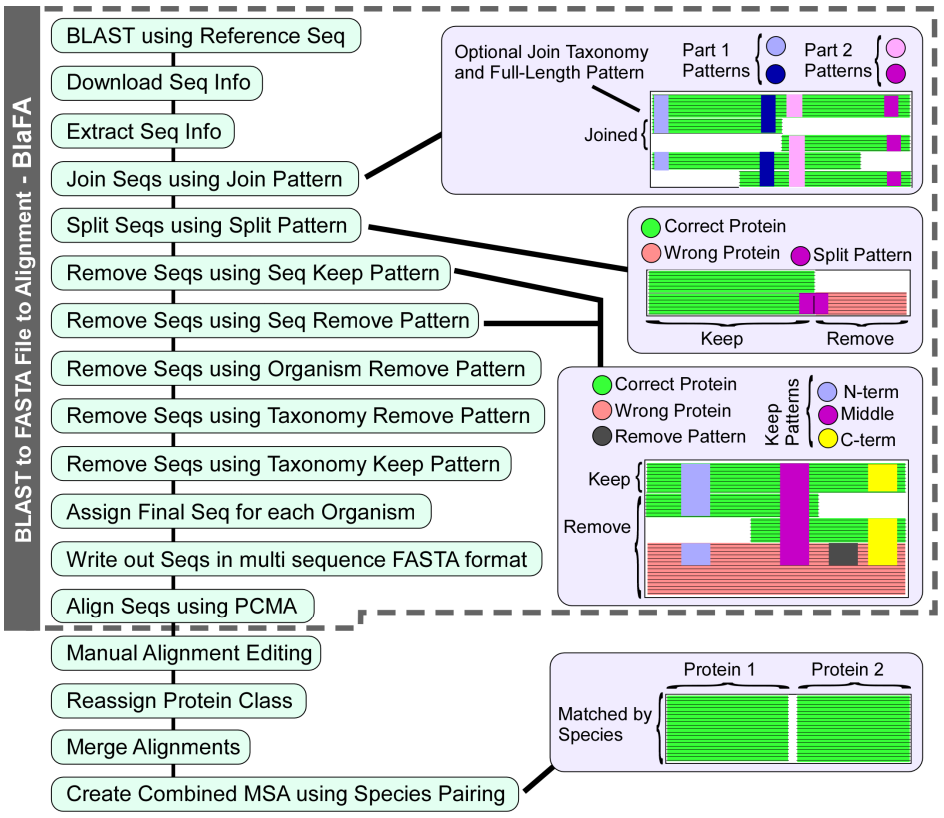


Figure 3.4 – Sequence Retrieval, Processing, and Alignment Methodology.
The creation of the bacterial β/β′ and All RNAP Large Subunit alignments required several steps. First BlaFA (gray dashed region) was used to retrieve and process the sequences, which were then aligned using PCMA, followed by manual alignment editing. In the case of the All RNAP Large Subunit the class of the RNAPs also had to be reassigned and merged together.

BlaFA first did a BLAST search to determine a list of the available NCBI sequences. This was followed by sequence selection in which the downloaded sequences were processed to join split gene products, fused gene products were split, and incorrect and partial sequences were removed. Sequences were initially aligned using the program PCMA (Profile Consistency Multiple sequence

Alignment) followed by manual alignment editing in PFAAT [75] to fix alignment errors as well as to remove the lineage-specific insertions.

Large Subunit Alignments

We used BlaFA plus manual alignment editing to create alignments for the bacterial β and β′ large subunits. In addition, since all multi-subunit RNAPs include homologues to the bacterial β and β′ subunits, we also created All RNAP Large Subunit alignments by extending our analysis to include the non-bacterial RNAPs from eukaryotic pol I/II/III RNAP, Nuclear-Cytoplasmic Large double-stranded DNA Viruses (NCLDV) pol like RNAP, archeal pol II like RNAP, and plant plastid RNAP. Table 3.1 shows the number of sequences for each subunit alone, as well as the number of species where both subunits are available.

Table 3.1 – Number of Sequences in the Large Subunit Alignments.

	Number of Large Subunit and Homologue Sequences		
RNAP Class	β	β'	β/β'
Bacterial RNAP	**958**	**842**	**814**
pol I RNAP	99	97	81
Eukaryota	**60**	**59**	**48**
Viruses	39	38	33
pol II RNAP	113	134	94
Eukaryota	63	77	45
Archaea	40	41	40
Viruses	10	16	9
pol III RNAP	58	64	51
Eukaryota	58	64	51
plastid RNAP	71	50	49
Eukaryota	71	50	49
Totals	**1299**	**1187**	**1089**

Phylogenetic Analysis of the All RNAP Large Subunits

Figure 3.5 shows the phylogenetic tree for the combined All RNAP Large Subunit alignment. As you can see from the tree each class of RNAP was clearly segregated indicating that our RNAP class assignments were accurate.

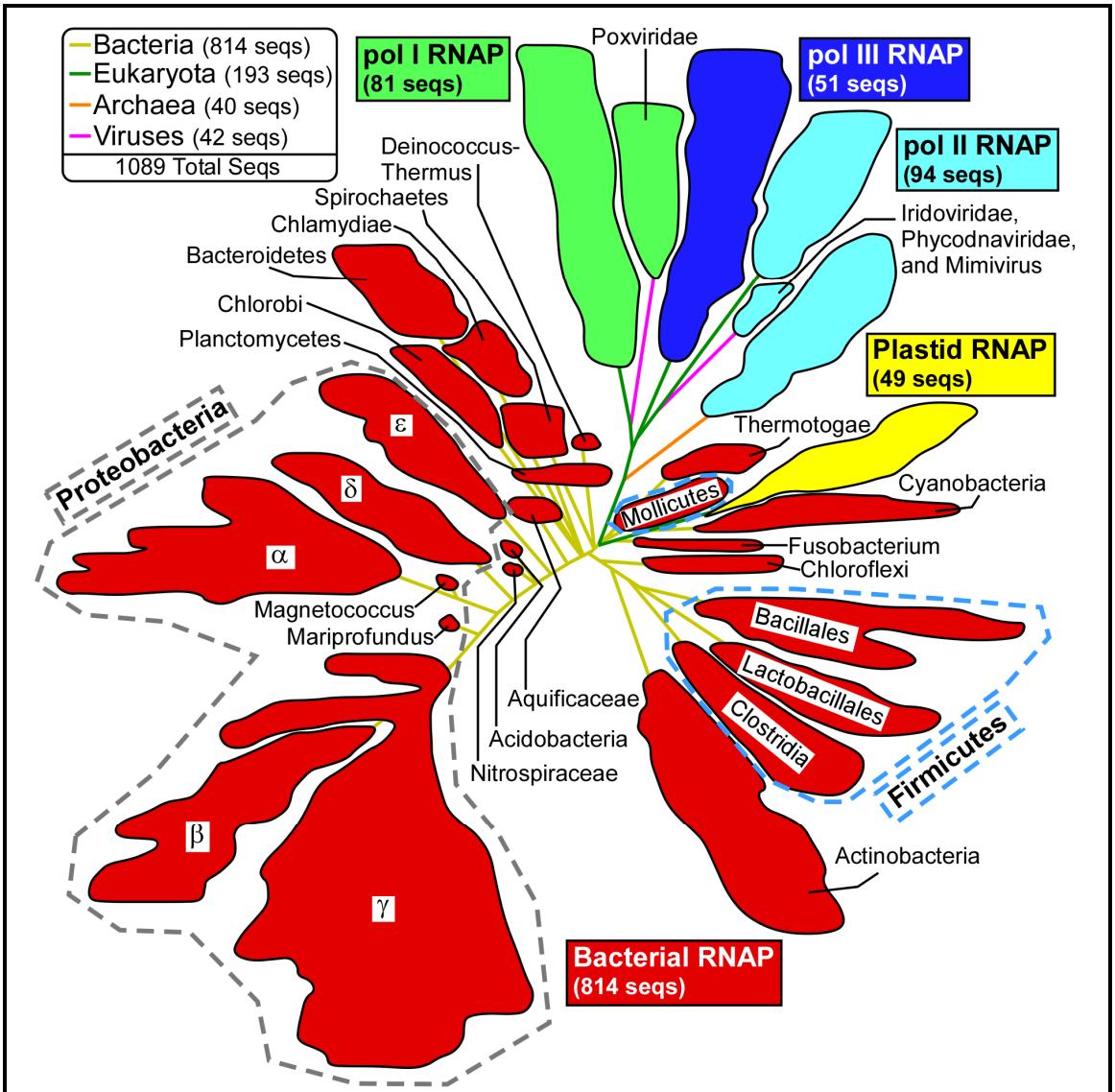


Figure 3.5 – Phylogenic Analysis of the All RNAP Large Subunit MSA.

The two All RNAP Large Subunit alignments were combined by species and the residues positions pruned to only keep those in the our shared sequence regions. The phylogenic tree was calculated using PHYLIP v3.66 with bootstrapping 100 replicates, followed by protein distance calculation using the JTT (Jones-Taylor-Thornton) method and neighbor joining to create the phylogenic tree which was analyzed using TreeDyn [76]. Due to the large number of sequences only the boundaries for each group of leaves are shown colored by RNAP class with bacterial RNAP (red), plastid RNAP (yellow), pol I RNAP (green), pol II RNAP (blue), and pol III RNAP (cyan). The branches for each leaf region are colored by taxonomy with bacteria (yellow), eukaryota (green), archeaea (orange), and viruses (magenta). Due to their diversity the proteobacteria (gray dashed region) and firmicutes (light blue dashed region) taxonomy subdivisions have been individually labeled.

In addition, the pol I/II/III RNAPs were located on one side and the bacterial RNAP and closely related plastid RNAPs were located on the other. As expected, the analysis also showed that although the archeal RNAPs clearly belong to the pol II class they also represent an intermediate between the eukaryotic and bacterial RNAPs. Furthermore, our analysis showed that although the viral RNAP from the NCLDVs are in fact related to eukaryotic RNAP, they have diverged since being acquired from their eukaryotic hosts. However, to our knowledge it has not been appreciated before that the NCLDV Iridoviridae, Phycodnaviridae, and Mimivirus families seem to have acquired a pol II RNAP, while the Poxviridae family seems to have acquired a pol I RNAP. In addition, close examination of the bacterial RNAP branch showed that the pattern of segregation correlated with bacterial taxonomy, demonstrating that our alignment contained sequences from a large set of diverse bacteria. Furthermore, it also highlighted the previously established close relationship between the cyanobacteria and plastid RNAPs.

Bacterial Large Subunit Fusions

Recently, the naturally occurring fusion of β and β' [77, 78] in the Helicobacter species has been implicated in the fitness of bacterial infection as well as the decreased sensitivity of Helicobacter RNAP to urea [79]. As expected, we found fused β and β' subunits in all of the examined Helicobacter family species including: *Helicobacter pylori 26695* (gi:15645812), *Helicobacter pylori HPAG1* (gi:108563562), *Helicobacter pylori J99* (gi:04155718), *Helicobacter hepaticus ATCC 51449* (gi:32261909), and *Helicobacter acinonychis str. Sheeba*

(gi:109948061). As expected, we also found a fused β and β' subunit [78] in the related Wolinella family species *Wolinella succinogenes DSM 1740* (gi:34556892). It is thought that all ε-proteobacteria of the Helicobacteraceae family (Helicobacter and Wolinella) contain a fused β and β' subunit [78]. However we found one assigned Helicobacteraceae family species *Thiomicrospira denitrificans ATCC 33889* which seems to have separately encoded β (gi:78497094) and β' (gi:78497095) subunits. On closer examination, we uncovered that in *Thiomicrospira denitrificans ATCC 33889* the genes encoding for β and β' share an unusual two codon overlap also found in the closely related Campylobacteraceae species that have non-fused β and β' subunits. In addition, based on our phylogenetic analysis *Thiomicrospira denitrificans ATCC 33889* segregates to its own branch located directly before the branch that contains the Helicobacteraceae and Campylobacteraceae branches. Given that RNAP large subunits have been used for taxonomy classifications, we believe that our results indicate that *Thiomicrospira denitrificans ATCC 33889* should not be considered part of the Helicobacteraceae family, but rather as its own family under the Campylobacterales (which also includes Campylobacteraceae and Helicobacteraceae), with the following proposed full taxonomy: Bacteria; Proteobacteria; Epsilonproteobacteria; Campylobacterales; Thiobacteraceae.

Surprisingly, we also discovered a previously uncharacterized (to our knowledge) β/β' fusion in 3 of 4 sequences from the parasitic intracellular α-proteobacteria and Rickettsiaceae member Wolbachia family including:

Wolbachia endosymbiont strain TRS of Brugia malayi (gi:58419220), *Wolbachia endosymbiont of Drosophila melanogaster* (gi:42409679), *Wolbachia sp. wMel* (gi:81652940), but not in *Wolbachia pipientis* (gi:15081478). However, it is important to note that on closer examination the one Wolbachia species exception, *Wolbachia pipientis, was* from a phylogenic study that only sequenced the β subunit [80]. Therefore, it would be reasonable to conclude that in all of the Wolbachia, including *Wolbachia pipientis*, contain fused β and β' subunits. This presence of fused β and β' in another distant branch of bacterial RNAP is very intriguing and possibly represents a convergent evolutionary event. Furthermore, the sequence of the Wolbachia fusion site is not similar to the Helicobacteraceae fusion which also contains 6 additional residues. Though the importance of the Wolbachia fusion is not known, similar to the Helicobacteraceae it might increase pathogenic fitness.

Bacterial rpoB/rpoC Intergenic Gap Analysis

Normally, the genes for the bacterial large subunits are transcribed as a single transcriptional unit with the gene for β (rpoB) preceding the gene for β' (rpoC), but they are translated separately with the rpoB stop codon and the rpoC start site separated by an untranslated 20-100 bp linker [78]. As mentioned above, the Campylobacteraceae species, which do not contain fused β and β' subunits, have a two codon overlap between rpoB and rpoC. It has been proposed that either a 1 bp addition or 2 bp deletion in a common ancestor could of have lead to a frame shift mutation capable of fusing β and β' in the related Helicobacteraceae [78]. Therefore, we decided to examine the rpoB/rpoC

intergenic gap in an effort to possibly understand the Wolbachia β and β' fusion as well as update our understanding of this gap across the known bacterial genomes. Figure 3.6 shows the rpoB/rpoC intergenic gap for 426 bacterial species.

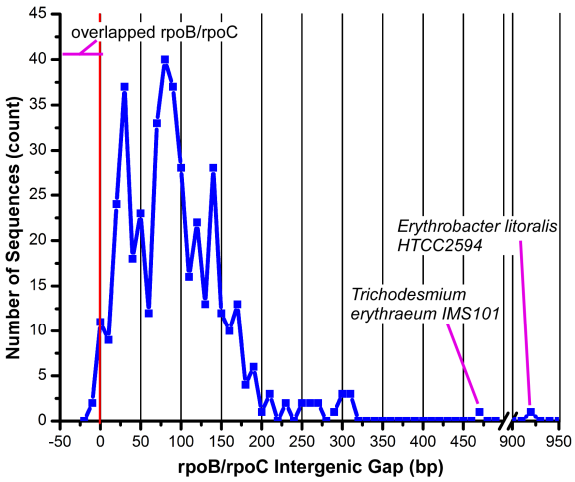


Figure 3.6 – Bacterial rpoB and rpoC Intergenic Gap Analysis.

The distance between the β gene (rpoB) stop codon and β′ gene (rpoC) start codon was analyzed. The number of sequences vs. intergenic gap was ploted as a blue line. The x-axis has been split between 500 and 900 bp. The red vertical line indicates an intergenic gap of zero with minus values indicting overlapping rpoB and rpoC genes. The species with fused β/β′ subunits are not shown since they do not have a true intergenic gap.

In agreement with the previously known results we find that the intergenic gap is usually between 10-200 bp. However, we also find a number of interesting exceptions. Regarding overlapped genes (negative intergenic gap), we find the previously known 8 bp overlap in Campylobacteraceae and the 4 bp overlap in Aquificae, plus additional overlaps of 14 bp in Chloroflexi, 4 bp in *Candidatus Carsonella* (γ-proteobacteria), 1 bp in Alcaligenaceae (β-proteobacteria), 1 bp in *Clostridium novyi NT* (Clostridia), and 1 bp in *Thiomicrospira crunogena XCL-2* (γ-proteobacteria). We also found unusually large intergenic gaps of 462 bp in

Trichodesmium erythraeum IMS101 (Cyanobacteria) and 916 bp in *Erythrobacter litoralis HTCC2594* (α-proteobacteria). In general Cyanobacteria, which have split β' subunits, contain a β gene followed by two sequential β' genes. We found that in most Cyanobacteria the intergenic gap between the β gene and the first β' gene is between 38-134 bp. The unusual *Trichodesmium erythraeum IMS101* contains the same β and β' gene organization, but for some unknown reason it has an extra long gap been the β gene and the first β' gene. The extremely large gap in *Erythrobacter litoralis HTCC2594* is the result of an unknown gene product (gi:85375720) encoded in the same direction between rpoB and rpoC. It should be noted that the *Erythrobacter litoralis HTCC2594* rpoB and rpoC genes both encode for full-length protein subunits. The unique rpoB/unknown/rpoC gene organization in *Erythrobacter litoralis HTCC2594* is extremely interesting since seems likely that the unknown gene is transcribed with the rpoB and rpoC and might even interact with RNAP.

Additionally, we found that the species most closely related to Wolbachia contained non-overlapped but short rpoB/rpoC intergenic gaps of 11-19 bp. Therefore, we believe that although the fusion of Wolbachia did not take place exactly as in Helicobacteraceae, it is certainly possible that this small gap could have been transformed into a β and β' fusion by the correct frame shift mutation or small deletion. It also supports the idea that the Wolbachia and Helicobacteraceae fusions were independent evolutionary events. In addition, it would seem that persistent β/β' fusions are rare events, since there are multiple species with small gaps and overlapping genes that to our knowledge have not

resulted in closely related species with β/β' fusions. Presumably, the persistent existence of β/β' fusions must also confer an evolutionary advantage, as in the Helicobacteraceae.

Bacterial Lineage-Specific Insertions

Iyer *et al.* have previously studied RNAP lineage-specific domain insertions in a small but diverse group of 42 bacterial species [73]. Using our alignments of the bacterial β (958 seqs) and β′ (842 seqs) subunits we located all of the previously identified insertions along with new lineage patterns and identified additional insertions. In β (Figure 3.7) we located 12 inserts (βIn1-βIn12) and in β′ (Figure 3.8) we located 7 (β′In1-β′In7). Mapping of the insert start locations onto the bacterial RNAP structure revealed that, without exception, the inserts are located on the outer surface of RNAP (Figure 3.10). Based on lineage-specific domain insertions (Figure 3.8) and phylogenetic analysis (Figure 3.5), the Acidobacteria and Nitrospirae bacterial species seem to belong to what Iyer *et al.* defined as the Group I bacteria, which also includes Proteobacteria, Aquificae, Spirochaetes, Chlamydiae, Planctomycetes, Chlorobi, Fusobacteria, and Bacteroidetes.

We also find that the Acidobacteria contain a Zn ribbon motif (βIn2) inserted 4 amino acids after the known Aquificae Zn ribbon insertion (βIn1), as shown in Figure 3.8 and Figure 3.10. Interestingly, although the Acidobacteria and Aquificae Zn ribbons align to each other, they appear different both by sequence gazing and via the BLOCKS motif finder (http://blocks.fhcrc.org/). Similar to Iyer et al, we find that the Aquificae Zn ribbon is related to the eukaryotic RNA

polymerase Rbp10 Zn ribbon, possibly representing inter-kingdom horizontal gene transfer [73]. However, using BLOCKS we find that the Acidobacteria insert is a FYVE type Zn ribbon. Therefore, we believe that the two insertions represent independent horizontal gene transfer events. However, it is very possible that both Zn ribbons play similar functional roles.

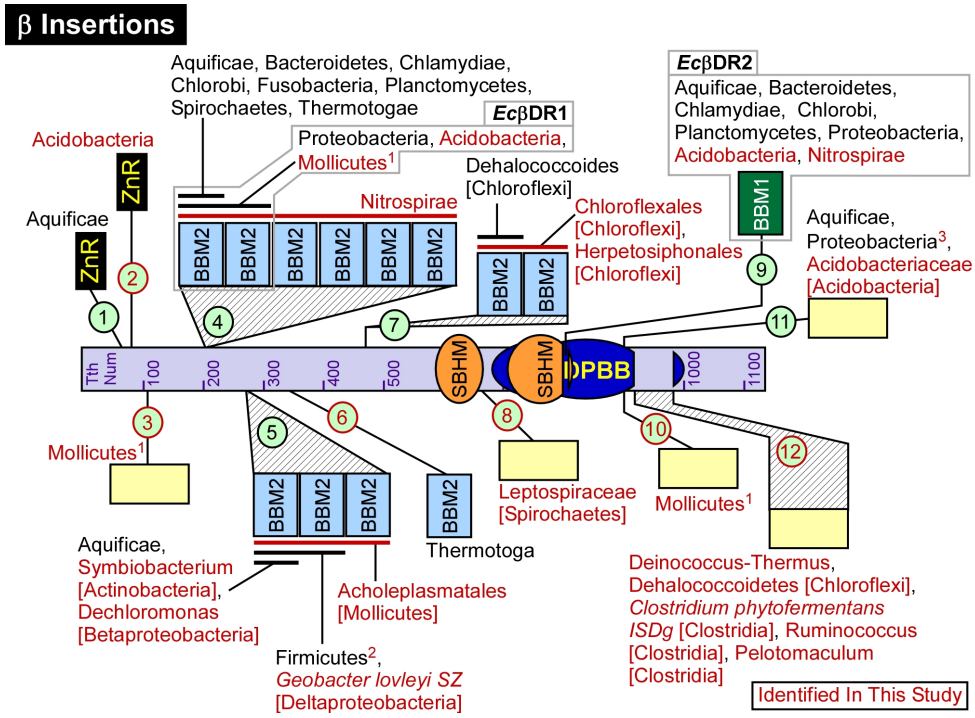


Figure 3.7 – Bacterial β Lineage-Specific Domain Insertions.

The locations of the β Inserts (βIn1-βIn12) are indicated using numbered light green circles. Red text or lines indicate inserts or lineage details identified in our study. The light gray boxes indicate the identities of previously well studied inserts. The taxonomy lineage details are as inclusively broad as possible. Where subfamily taxonomy is given the root taxonomy name to which it belongs is given in square brackets (Proteobacteria and Firmicutes are given taxonomy names one level more specific). The individual bacteria species name is given if it is the only member of a number of related bacteria to contain the insert. [1]Missing in some Mollicutes. βIn4 and βIn10 contain the same Mollicutes species and are mutually exclusive with βIn3 in terms of Mollicutes species. [2]Missing in some Firmicutes species. Some of the Firmicutes missing this insert represent the top 8 species with the smallest combined β/β′ sequence lengths. [3]The Wolbachia species, which also have fused β/β′, have an additional 69 amino acid extension at the N-term of this insert.

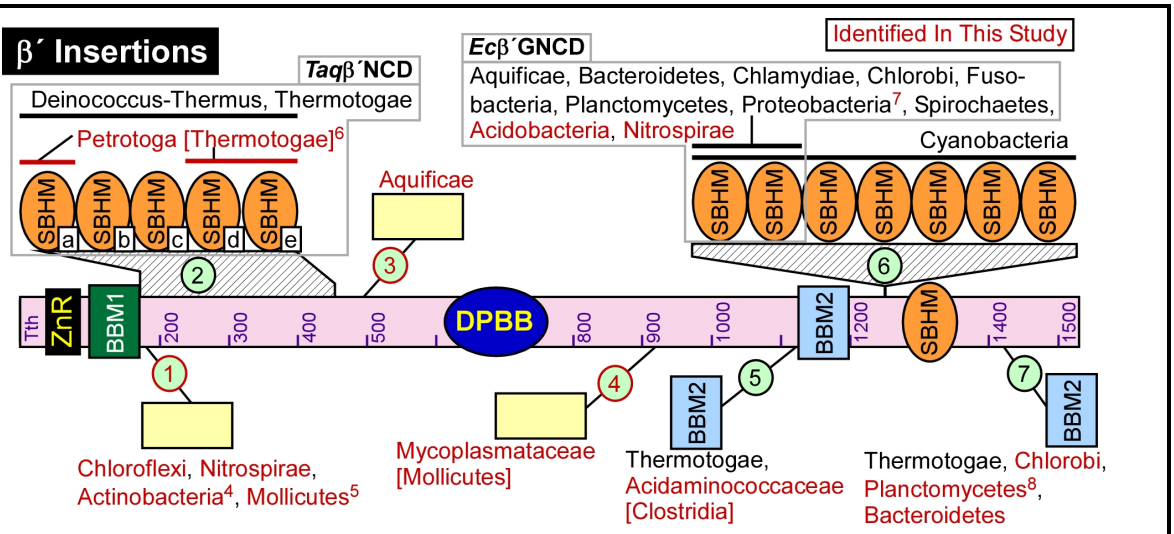


Figure 3.8 – Bacterial β′ Lineage-Specific Domain Insertions.

Same as Figure 3.7, but with the locations of the β′ Inserts (β′In1-β′In7). [4]Missing in Symbiobacterium subfamily species. [5]Missing in Acholeplasmatales subfamily species. [6]Missing a region of sequence in the middle of the insert that removes domains b and c, which interestingly both extend past the σ subunit and therefore lack interactions at the interface between this insert and σ. [7]The ε-Proteobacteria subfamily inserts contain ~150 additional amino acids. [8]Missing in *Candidatus Kuenenia stuttgartiensis*.

Additionally, there are several examples of lineage-specific modifications either representing the addition to or removal of only part of a pre-existing insertion. The Wolbachia species, which also have fused β/β′, have an additional 69 amino acid extension at the N-term of the shared Proteobacteria insertions βIn11. Given that a very large group of 523 related Proteobacteria contain βIn11 without the additional amino acids found in the Wolbachia species, it would be reasonable to assume that this represents an extension to a pre-existing insertion. Similarly, we find that β′In7 in the ε-Proteobacteria contains ~150 additional amino acids mostly inserted at two locations in the 1st SBHM domain. In contrast, we also find a possible partial loss of β′In2, whose SBHM structural

domains are split and not in sequential sequence order. *Petrotoga mobilis SJ95*, which is a member of the Thermotogae family and the only examined member of the Petrotoga subfamily, is missing a stretch of amino acids in the middle of the insert sequence, leading to the clean removal of β′In2 domains b and c (Figure 3.8). The removal of domains b and c is particularly interesting since in the context of holoenzyme they both extend into space beyond the interaction interface between the σ subunit and domains a, d, and e [74]. Furthermore, the discrete loss of domains b and c would be in agreement with the proposal that domains b and c have a role independent from the σ subunit interaction of a, d, and e domains. However, since only 12 of the available sequences contain β′In2 it would be difficult to definitively say if the case of domains b and c truly represent a loss and not an extension into the middle of a pre-existing insert.

We also determined the minimal and maximal bacterial large subunits (β/β′) using the total combined length of β and β′. In general, the shortest β/β′ came from the Firmicutes sub-families of Bacillales and Clostridia. The 8 shortest β/β′ were Clostridia sequences missing the ~90 amino acid 2x SBHM insertion (βIn5) found in every other Firmicutes including other Clostridia sequences. The shortest β/β′ was from the thermo and halophilic bacterium *Halothermothrix orenii H 168* [81], which is missing an additional 22-82 amino acids when compared to the other 7 shortest Clostridia. The maximal β/β′ was from the Nitrospirae and Cyanobacteria, which contain large repeated domain insertions in βIn4 and β′In6 respectively.

Despite our best attempts there are several insertions with unidentifiable domain motifs. Either these domains represent uncharacterized insertion motifs or they match the known motifs, but are not identifiable due to low sequence similarity or non-sequential sequence order. As in the case of β′In2 (*Taq*β′NCD), proper identification might require structural information in the form of either a complete RNAP or isolated insertion structure. When combined with the right experimental evidence, this structural information might allow for a more complete understanding of the lineage-specific domain insertions.

Shared Sequence Regions

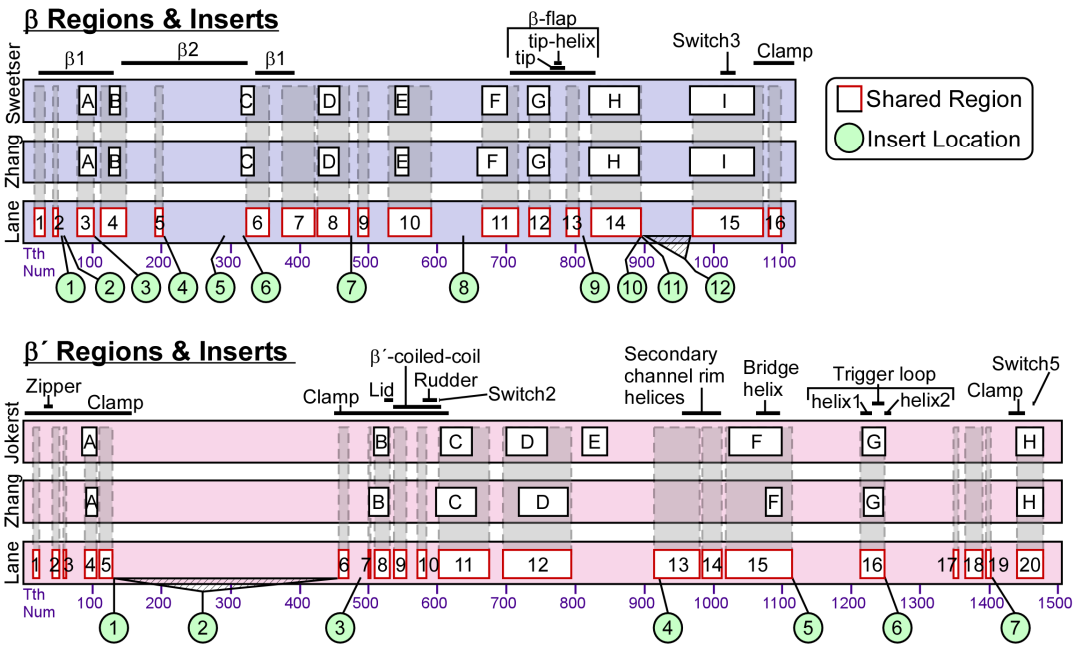


Figure 3.9 – Shared Sequence Regions Common to All Multi-Subunit RNAPs.
We used the residue positions alignable in all multi-subunit RNAPs to determine new shared sequence regions β (βR1-βR16) and β′ (β′R1-β′R20). Our new regions (red boxes) are shown in comparison (gray dashed line box) to the old regions (black boxes). Common structural features are indicated above each proteins schematic and bacterial lineage-specific domain insertion start locations (light green circles) bellow.

Previous studies have established regions within the two large subunits that are similar across all classes of RNAP [4, 71, 72]. However, the initial β and β′ regions were established in 1987 [71] and 1989 [72] using very few sequences. Although Zhang *et al.*, did update these regions in 1999 using the bacterial RNAP structure, they did not fundamentally alter the previous regions [4]. However, since our alignments contained many more sequences we decided to use them to define a completely new set of shared sequence regions, using the positions alignable in all large subunit sequences.

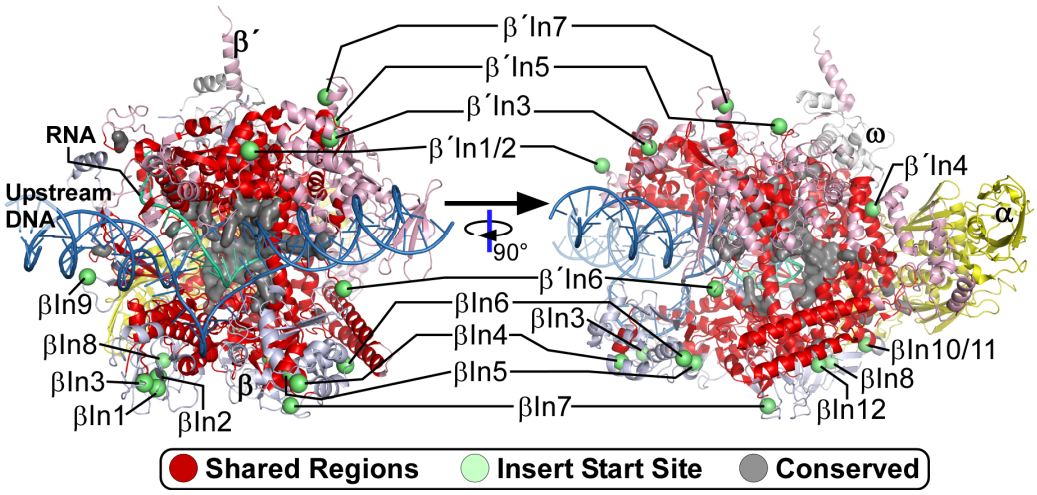


Figure 3.10 – Structural Mapping of Shared Sequence Regions and Bacterial Lineage-Specific Domain Insertion Start Sites.

The *Tth* RNAP structure is shown as cartoons with α (yellow), β (light blue), β′ (light pink), ω (white), shared sequence regions in β and β′ (red), and bacterial lineage-specific domain insertion start sites (light green spheres). The TEC model DNA (blue) and RNA (teal) are shown as phosphate-backbone ribbons, with bases indicated schematically as sticks. The residues conserved across all classes of RNAP are shown as grey molecular surfaces.

For β we defined 16 regions (βR1-βR16) and for β′ we defined 20 regions (β′R1-β′R20). In general we found most of the previously established regions and for some we were able to extend the boundaries. Also, similar to Zhang *et al.*, we

find that β′ region E as originally established by Jokerst *et al.* is not in a region shared by all classes of RNAP. In addition, we have added several new regions which were previously not identified. Figure 3.9 shows a comparison of our shared sequence regions and the previously established regions, along with the locations of the bacterial lineage-specific inserts and important structural features. Mapping of the shared sequence regions onto the bacterial RNAP structure revealed that they comprise the center as well as parts of the outer surface of RNAP (Figure 3.10).

Conclusions

We created comprehensive alignments of the multi-subunit RNAP large subunits. During this process we discovered an uncharacterized fusion of the β and β′ subunits in the parasitic intracellular Wolbachia bacteria. In addition to clarifying the shared sequence regions of RNAP common to all classes of multi-subunit RNAPs, the alignment also allowed us to gain additional insights into the bacterial lineage-specific domain insertions. We identified all of the previously characterized insertions (some with expanded lineage patterns) and a number of new insertions. We also uncovered several examples of possible horizontal gene transfer of intact or partial domain insertions. By creating a comprehensive list of the intergenic gap between the bacterial β and β′ genes we revealed important insights into the genesis of β and β′ subunit fusion. We also discovered a unique β and β′ gene organization in *Erythrobacter litoralis HTCC2594*, which has an unknown gene product encoded in the same direction between the β and β′ genes. Our extensive alignments will provide an extremely valuable resource for the study of multi-subunit RNAPs. In addition, we believe that our customizable sequence retrieval, processing, and alignment system, (BlaFA) along with our insertion detection methodology, will aid in the study of other large and complex protein families.

Materials and Methods

BLAST to FASTA File to Alignment (BlaFA)

Sequences were downloaded and aligned using a custom program called BlaFA which allowed for programmable automation. A representative sequence (ie from *Ec* K12) was used to BLAST the NCBI non-redundant (nr) dataset using NetBLAST (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/LATEST/). The BLAST result list was then used to extract the NCBI Genbank ID (gi) for each potential sequence. The sequences description page (INSDSeq XML Format) was used to extract the following information: organism name, strain name, sub-strain name, taxonomy, protein product description, and protein sequence. Next each protein sequence was evaluated in order to determine if it was the correct full-length sequence. These steps were necessary since in addition to partial sequences, some species have β and β' proteins that are naturally split in half or fused to each other.

In order to allow for a powerful and flexible system this process was automated using custom sequence and taxonomy patterns. Possibly split sequences were first identified from the non full-length sequences lacking either a N or C-term pattern (ie N-term: G.....T and C-term: KEN...G). If a sequence was identified as a potential split sequence the BLAST result list was then searched to find other sequences from the same organism. The potential halves were then identified using sequence patterns for the N-term and C-term of each half and joined by appending one to the next. In addition, since such splits usually correlate to certain taxonomies we often restricted the joins with

taxonomy patterns (ie cyanobacteria). Next the system identified fused proteins, in which two proteins that are normally expressed separately in most species are instead expressed as a single large protein. Each fused sequence was evaluated using a sequence pattern unique to the fusion site (ie 2,I.....F.....|ASP..I...S.GE where 1 or 2 specifies the half to keep and "|" indicates where to split). Once a fusion site was found the correct half was used in place of the originally fused sequence. Next, incorrect and partial sequences were removed using a list of sequence keep and remove patterns. In order, to be kept a sequence had to contain all of the specified sequence keep patterns (ie an N-term and a C-term pattern to remove partial seqs) and none of the sequence remove patterns (useful for removing unwanted proteins that might pass keep patterns, like when you only want a sub-set of proteins that are part of a much larger closely related protein family). Next, sequences were removed using taxonomy keep patterns (ie general ones like Bacteria or very specific ones like Enterobacteriaceae) and remove patterns (ie Not Bacteria or not Enterobacteriaceae). The remaining sequence with the best BLAST expect score was then assigned as the final sequence for each unique species and written to a multiple sequence FASTA file, which also contained the protein sequence extracted from a known structure.

In practice identifying the various sequence patterns *a priori* can be very challenging. Especially since choosing the best pattern often requires optimizing between stringency and effectiveness. Therefore, it was often necessary to do a series of pattern optimizing BlaFA runs. For example to determine the join patterns, it was best to do a pattern optimizing BlaFA run for the first part as well

as the second part of the protein, excluding irrelevant sequences using a taxonomy pattern. If possible we also created an additional alignment with both halves and some full-length sequences. Either way, our goal was to get an alignment where we could clearly identity a protein as either the first or the second half, allowing the generation of the sequence join patterns. For split patterns, it was best to a pattern optimizing BlaFA run to get the sequence pattern for the part of the fusion we wanted to split. We could then manually do a sub-alignment with only those sequences that look to be fused (ie those with a larger than expected sequence length). In addition, we could specify an appropriately large sequence size cutoff within BlaFA and the system will automatically prune the final sequence lists after BLAST to enrich for the potentially fused sequences. An alignment containing both fused proteins and known full-length sequences covering the first or second half of the fusion were then used to correctly determine the fusion site and generate a fusion specific sequence pattern for splitting. For generating the sequence keep and remove patterns it was often necessary to do a pattern optimizing BlaFA run from which we manually removed partial sequences. After determining the optimized patterns, we preformed a test BlaFA run, followed by an additional BlaFA run using only a species exclude pattern that excluded all of the species found in the test run. We used the second run to identify sequences that should have been included in the first test BlaFA run and adjusted our patterns accordingly. Once we were satisfied that the BlaFA patterns properly included and processed the target sequence while at the same time excluding unwanted sequences we

performed the final BlaFA runs (See Table 3.2 – Table 3.6 for BLAST dates and final BlaFA patterns).

Although this approach required a lot of upfront manual effort, once the patterns were established they could be used in concert to quickly identify the correct sequences without having to worry about the multitude of steps where human error could have resulted in a problem. In addition, well designed patterns based on a diverse set of sequences that are not too restrictive can be used in the future to quickly identify newly available sequences.

The sequences in the multiple sequence FASTA files were then aligned using the program PCMA [82] available at ftp://iole.swmed.edu/pub/PCMA/. PCMA was chosen since it uses a two-stage strategy in which it first quickly pre-aligns highly identical sequences (similar to ClustalW) followed by alignment of the divergent sequences using profile-profile comparison and consistency (similar to T-Coffee). In our experience, PCMA produced the most accurate alignments in a relatively short amount of time.

Nonetheless, the alignments still needed to be manually edited by hand, usually due to the presence of small (1-2 residue) and large (50-600 residue) lineage-specific insertions. Furthermore, the manual editing of the two large subunits of RNP were additionally complicated due to the large number of sequences, large size of the proteins, and many different lineage-specific insertions at various locations within the sequence. The program PFAAT [75], available at http://pfaat.sourceforge.net/, was used to manually edit the alignments in order to remove lineage-specific regions and fix any misaligned

positions. In general the manual alignment editing process consisted of iterative cycles in which an alignable region was identified between two conserved boundaries defined by stretches of positions that were identical or nearly identical in all of the sequences. The conserved boundaries were usually easy to spot since they were well aligned across all of the sequences by PCMA. In most cases all of the sequences contained the same number of intervening amino acids between the conserved boundaries, along with areas of high sequence similarity either between all or groups of sequences. In contrast, the positions outside of the conserved boundaries usually contained the lineage specific sequence insertions as well as stretches of low sequence similarity that PCMA tried to align by adding lots of gaps. In some cases, manual alignment editing was successfully in properly aligning the regions outside of the conserved boundaries, since often the presence of a lineage specific insertion on the outside edge of the conserved boundary simply caused PCMA to misalign positions that were otherwise alignable in all sequences. However, there were also many sequence regions outside of the conserved boundaries that possessed low sequence similarity that could not be aligned by PCMA or manual alignment editing. Therefore, the conserved boundaries were used to define the alignable regions which were kept, while positions outside of the conserved boundaries were either manually re-aligned or removed. In addition, we removed the lineage specific insertions due to their complicated patterns of insertion and the presence of stretches of low sequence similarly that were usually found before and after their sites of insertion.

Table 3.2 – Bacterial RNAP BlaFA Patterns

Bacterial β	
└ BLAST Date	6/5/2006 and updated on 5/3/2007
└ Input Sequence	Beta - RNA polymerase subunit [Escherichia coli K12]. GenBank ID: 1790419
└ Seq Split Pattern	# rpoB from fused RNAP with both rpoB and rpoC 1,I.....F.....|ASP..I...S.GE
└ Seq Keep Pattern	# rpoB Region B P.....G.....Q # rpoB start of Region I G.....KL.H....K # rpoB end of Region I GEME.WA.....T.KSD
└ Species Remove Pattern	# plamsa group ... they align as a separate group and contain many differences that mess everything up. Mycoplasma Ureaplasma Helicobacter pylori Mesoplasma Spiroplasma phytoplasma
└ Taxonomy Keep Pattern	#Keep only the Bacteria ^Bacteria
Bacterial β'	
└ BLAST Date	6/5/2006 and updated on 5/3/2007
└ Input Sequence	Beta Prime - RNA polymerase subunit [Escherichia coli K12]. GenBank ID: 2367335
└ Seq Join Pattern (alignment was done to figure out how the two halves split in comparison to the full-length protein)	# rpoC joining pattern for Cyanobacteria ... # All that need to be joined contain the following in its taxonomy taxonomy_contains,cyanobacteria # N and C-term of full-length protein full,G.....T.NY.....E..G full,S.....L.....KEN...G # N and C-term of part1 part1,FDY.K...ASP.R...W part1,G.....T..GR...N # N and C-term of part2 part2,F.N....K..L..L.....G.A part2,S.....L.....KEN...G
└ Seq Split Pattern	# rpoC from fused RNAP with both rpoB and rpoC 2,I.....F.....|ASP..I...S.GE
└ Seq Keep Pattern	# rpoC N-term G.....T.NY.....E..G # rpoC C-term S.....L.....KEN...G
└ Taxonomy Keep Pattern	#Keep only the Bacteria ^Bacteria

Table 3.3 – pol I RNAP BlaFA Patterns

pol I β Homologue	
└ BLAST Date	9/9/2006
└ Input Sequence	A135 - RNA polymerase I subunit [Saccharomyces cerevisiae]. GenBank ID: 6325267
└ Species Remove Pattern	#These are very long over 2000 residues ... not sure why. #They are removed to save time aligning etc ... Pseudendoclonium akinetum Gallus gallus
└ Taxonomy Remove Pattern	#remove bacteria ^Bacteria #only doing these in yrpoB2 ^Archaea ^Viruses #remove other junk ^Unclassified ^other sequences
pol I β' Homologue	
└ BLAST Date	9/9/2006
└ Input Sequence	A190 - RNA polymerase I subunit [Saccharomyces cerevisiae]. GenBank ID: 6324917
└ Species Remove Pattern	#These are very long over 2000 residues ... not sure why. #They are removed to save time aligning etc ... Pseudendoclonium akinetum Gallus gallus
└ Taxonomy Remove Pattern	#remove bacteria ^Bacteria #only doing these in yrpoC2 ^Archaea ^Viruses #remove other junk ^Unclassified ^other sequences

Table 3.4 – pol II RNAP BlaFA Patterns

pol II β Homologue	
⌊ BLAST Date	9/9/2006
⌊ Input Sequence	B150 - RNA polymerase II subunit [Saccharomyces cerevisiae]. GenBank ID: 6324725
⌊ Seq Join Pattern (alignment was done to figure out how the two halves split in comparison to the full-length protein)	# rpoC joining pattern for Archaea ... # All that need to be joined contain the following in its taxonomy taxonomy_contains,^Archaea # N and C-term of full-length protein full,R.R...Y..P.....P.M..S full,CP..G.....YAFKL...E # N and C-term of part1 part1,P...R.R...Y..(P|E|H) part1,E(A|T)R.(L|T)H.(T|S)..G.....PE....G..K # N and C-term of part2 part2,G...(G|D).....(R|K)..RR part2,YAFK(L|I)...E
⌊ Species Remove Pattern	#These are very long over 2000 residues ... not sure why. #They are removed to save time aligning etc ... Pseudendoclonium akinetum
⌊ Taxonomy Remove Pattern	#remove bacteria ^Bacteria #remove other junk ^Unclassified ^other sequences
pol II β' Homologue	
⌊ BLAST Date	9/9/2006
⌊ Input Sequence	B220 - RNA polymerase II subunit [Saccharomyces cerevisiae]. GenBank ID: 6320061
⌊ Seq Join Pattern (alignment was done to figure out how the two halves split in comparison to the full-length protein)	# rpoC joining pattern for Archaea ... # All that need to be joined contain the following in its taxonomy taxonomy_contains,^Archaea # N and C-term of full-length protein - could not find any this is the yrpoB2 one full,R.R...Y..P.....P.M..S full,CP..G.....YAFKL...E # N and C-term of part1 part1,P...R.....Y(D|E)..G.P.....D...G part1,T..SGY..RR.....V # N and C-term of part2 part2,G...AQS..EP part2,A.FE.....(L|I)..(A|T).....(G|A)..(E|V)(N|S)(V|I)..(G|N)
⌊ Species Remove Pattern	#These are very long over 2000 residues ... not sure why. #They are removed to save time aligning etc ... Pseudendoclonium akinetum
⌊ Taxonomy Remove Pattern	#remove bacteria ^Bacteria #remove other junk ^Unclassified ^other sequences

Table 3.5 – pol III RNAP BlaFA Patterns

pol III β Homologue	
⌊ BLAST Date	9/9/2006
⌊ Input Sequence	C128 - RNA polymerase III subunit [Saccharomyces cerevisiae]. GenBank ID: 6324781
⌊ Species Remove Pattern	#These are very long over 2000 residues ... not sure why. #They are removed to save time aligning etc ... Pseudendoclonium akinetum
⌊ Taxonomy Remove Pattern	#remove bacteria ^Bacteria #only doing these in yrpoB2 ^Archaea ^Viruses #remove other junk ^Unclassified ^other sequences
pol III β' Homologue	
⌊ BLAST Date	9/9/2006
⌊ Input Sequence	C160 - RNA polymerase III subunit [Saccharomyces cerevisiae]. GenBank ID: 6324690
⌊ Species Remove Pattern	#These are very long over 2000 residues ... not sure why. #They are removed to save time aligning etc ... Pseudendoclonium akinetum
⌊ Taxonomy Remove Pattern	#remove bacteria ^Bacteria #only doing these in yrpoC2 ^Archaea ^Viruses #remove other junk ^Unclassified ^other sequences

Table 3.6 – Plastid RNAP BlaFA Patterns	
plastid β Homologue	
└ BLAST Date	9/15/2006
└ Input Sequence	Beta - RNA polymerase subunit [Escherichia coli K12]. GenBank ID: 1790419
└ Taxonomy Keep Pattern (Might be too restrictive since we might also want some like Eukaryota; Rhodophyta; we later got some from these when doing the reassign annotation step)	#Keep the plants to get the cholorplast RNAP ^Eukaryota; Viridiplantae;
plastid β' Homologue	
└ BLAST Date	9/21/2006
└ Input Sequence	Beta Prime - RNA polymerase subunit [Escherichia coli K12]. GenBank ID: 2367335
└ Seq Join Pattern	# yrpoC4 joining pattern for Chroloplast RNAP ... # All that need to be joined contain the following in its taxonomy taxonomy_contains,^Eukaryota; Viridiplantae; # N and C-term of full-length protein - none since all are split full, # N and C-term of part1 part1,SP..I..W.....(P|S).....(G|N)E(V|I) part1,G.....P.QD...G....T # N and C-term of part2 part2,(K|R).....(G|D).....K..G....T part2,(T|L)...(L|I).....D...G(L|S).EN.....G.G

└ Seq Keep Pattern	# yrpoC4 N-term
	SP..I..W.....(P|S).....(G|N)E(V|I)
	# yrpoC5 C-term
	(T|L)...(L|I).....D...G(L|S).EN.....G.G

└ Taxonomy Keep Pattern (Might be too restrictive since we might also want some like Eukaryota; Rhodophyta; we later got some from these when doing the reassign annotation step)	#Keep the plants to get the cholorplast RNAP ^Eukaryota; Viridiplantae;

Creation of Bacterial Large Subunit Alignments

We used the *Ec* K12 β and β′ sequences as input reference sequences for BlaFA, along with the sequence from the bacterial *Tth* RNAP structure (pdb code 2BE5). We then manually edited the alignments and removed all regions not to all of the bacterial sequences. In addition to using the alignment editor PFAAT, we used a custom program (msa_util.pl) that allowed us to quickly manipulate various aspects of the alignment including removing or gapping regions of the alignment by specifying the inclusion of certain sequences and/or positions.

Creation of Alignments containing All RNAP Large Subunits

The creation of alignments containing the two large subunits from all classes of multi-subunit RNAP was a multistep process. Similar to the bacterial β/β′ alignments we first used BlaFA to determine the sequences for the two large subunits from the following classes of RNAP: pol I, pol II, pol III and plastid. However, the above sequence lists contained overlapping or incorrectly assigned RNAPs. To correct for this we used a custom program (create_reassigned_gene_and_comb_fas.pl) that read in the pol I, pol II, pol III and plastid RNAP sequences and then reassigned their class according to the class of the input sequence class to which it had the best BLAST score. We then created and PCMA aligned two multiple sequence files: (1) for pol I, pol II, and pol III sequences and (2) for plastid sequences.

To aid in the manual cleaning of the above two alignment alignments we created reference alignments with the sequence from the *Tth* structure (pdb code 2BE5), the sequence of the *Sce* yeast structure (pdb code 1TWF), and the sequence of *Ec* (gi:01790419/ 02367335) after it was manually edited and cleaned as per the previous bacteria alignment section. The *Tth* structural and *Ec* sequences were previously aligned in the bacterial alignment and the yeast structural sequence was structurally aligned to the *Tth* structural sequence. For the plastids we added the reference sequences for the bacterial *Tth* structural sequence and cleaned *Ec* K12 sequence. For pol I/II/III we added the reference sequences for the yeast structural sequence and the *Ec* K12 cleaned sequence.

We used the reference sequences as guides when manually cleaning the alignment in which we removed any sequence positions that were not in the

cleaned up *Ec* sequence, since we were only interested in creating alignments with regions shared by all classes of RNAP. We also used the reference alignment to aid in manually aligning the sequences.

We next used a custom program (msa_merge.pl) that used the reference alignments to merge two alignments together. We first merged the plastid and bacterial alignments, followed by the pol I/II/III alignments. We then removed the reference sequences leaving only the natural sequences and the *Tth* structural sequence, resulting in the All RNAP Large Subunit alignments.

Phylogenetic Analysis of the Combined All Large Subunit Alignment

Phylogenetic analysis was performed using only the shared sequence regions in a combined alignment created using a custom program (combined_msa_util.pl) that joined β and β′ or there homologues from the same species. Table 3.1 lists the number of sequences in each combined alignment. We used PHYLIP v3.66 (http://evolution.genetics.washington.edu/phylip.html) with bootstrapping 100 replicates, followed by protein distance calculation using the JTT (Jones-Taylor-Thornton) method and neighbor joining to create the phylogenetic tree. TreeDyn [76](http://www.treedyn.org/) was used to view and analyze the phylogenetic trees.

Intergenic Gap Analysis

We used a custom program (get_rpoB_rpoC_intergenic_gap.pl) that read in the NCBI record for each bacterial β and β′ species matched pair and extracted the gene start and stop locations for rpoB and rpoC if available. The program them verified that the two genes were going in the same direction and calculated

the bp distance between the stop codon of the rpoB gene and the start codon of the rpoC gene. Unusual distances were verified by visiting the NCBI gene link for the corresponding β or β′ subunit.

Detection of Bacterial Lineage-Specific Domain Insertions

In order to detect the bacterial lineage-specific insertions we first created individual alignments for each full-length protein sequence with its sequence from the final cleaned up alignment containing only the alignable sequence positions. In order to facilitate this we used a custom program (find_inserts.pl) to automate the creation of the individual alignments using PCMA, followed by manual correction of mismatched positions identified by the custom program. We then used a custom program (find_inserts.pl) to search through each alignment for large gaps (usually >50 residues) in the cleaned up sequence that would indicate where we removed a possible lineage-specific insertion or sequence region not contained in all of the bacterial sequences. In order to have a common frame of reference we also converted the insertion start and end positions to the *Tth* structure residue numbering. We then manually sorted the list of insertions to locate insertions with the same start and end points and extracted the insertion residues followed by alignment using MUSCLE [83], which proved to the best alignment program for this task. We then tried to identify the sequence motifs of the insertions by comparing our results to those obtained by Iyer *et al.* [73]. We also made use of the available structural information for the *Taq*β′NCD and *Ec*β′GNCD SBHM motif lineage-specific domain insertions [74].

Chapter 4 - The Molecular Evolution of Multi-Subunit RNA Polymerases*

Introduction

Protein Co-Evolution and Statistical Coupling Analysis (SCA)

Throughout evolution proteins maintain their structure and function by compensating for mutations at one position with co-evolving mutations at others. Statistical Coupling Analysis (SCA) allows for the pair-wise determination of co-evolution within large diverse multiple sequence alignments (MSA) [84-88]. SCA has previously shown that groups of co-evolving residues can form the basis of multi-residue networks that often map to contiguous stretches throughout the structure of a protein [84-88]. Furthermore, it has been experimentally shown that these networks generally govern core aspects of protein function including structural stability, regulation, and activity [84-88].

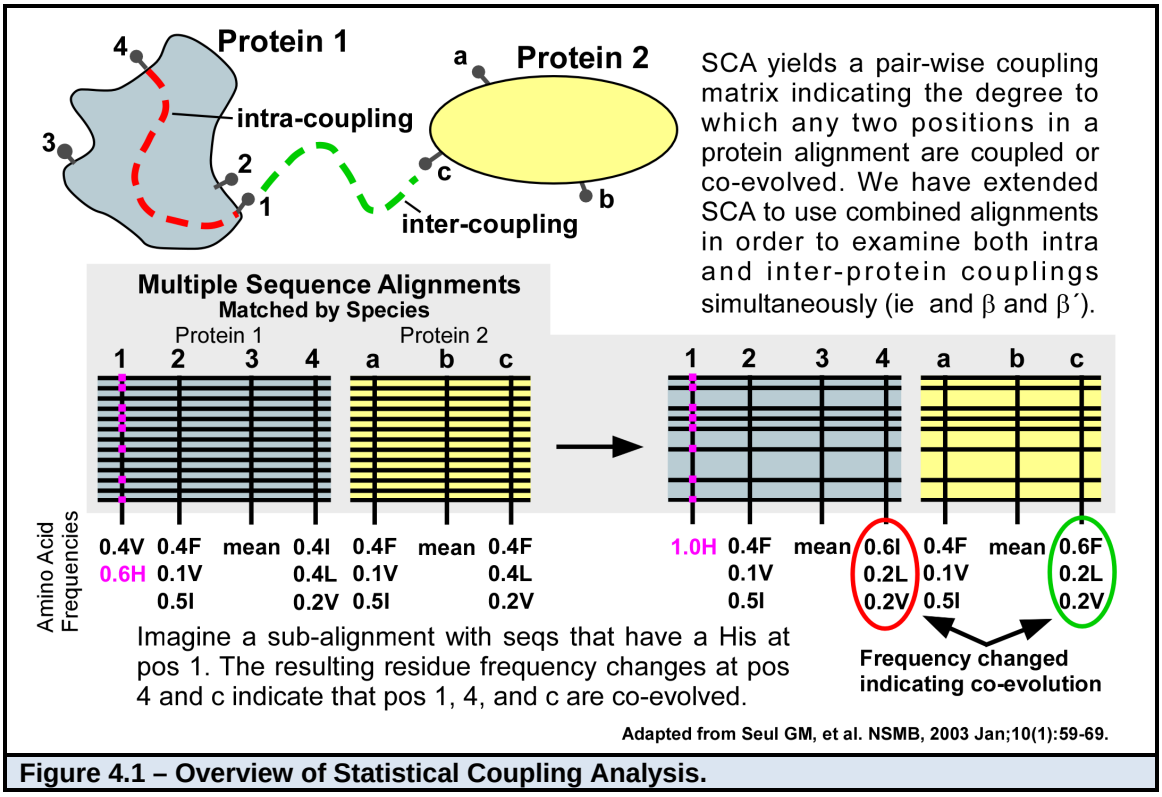


Figure 4.1 – Overview of Statistical Coupling Analysis.

*Paper in Preparation

Results and Discussion

Progress Timeline

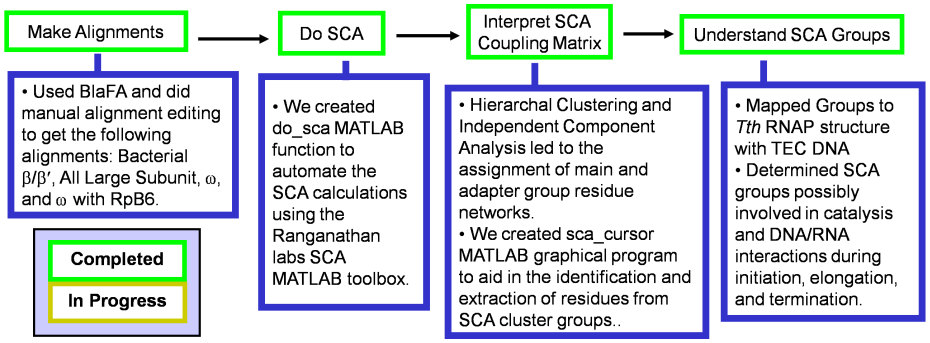


Figure 4.2 – The Molecular Evolution of Multi-Subunit RNAPs Progress Timeline.
Above is a schematic of this projects progress. Green boxes are completed parts. Yellow boxes are parts in progress or uncompleted. Black boxes are parts not started. Blue boxes important details for each step.

Multi-Subunit RNAP Large Sub-Unit Alignments

We created alignments of the bacterial large subunits (β and β′) and their non-bacterial homologues (see Chapter 3). Briefly, BlaFA was used to download and process the large subunit sequences, followed by manual alignment correction. The number of sequences in each alignment are indicated in Table 4.1.

Table 4.1 – Number of Sequences in the SCA Large Subunit Alignments.

	Number of Large Subunit and Homologue Sequences		
RNAP Class	β	β'	β/β'
Bacterial RNAP	**770**	**607**	**573 (196)**
pol I RNAP	102	98	84 (51)
Eukaryota	**62**	**60**	**50 (39)**
Viruses	40	38	34 (12)
pol II RNAP	117	136	95 (58)
Eukaryota	66	77	45 (23)
Archaea	41	42	41 (30)
Viruses	10	17	9 (5)
pol III RNAP	62	64	53 (36)
Eukaryota	62	64	53 (36)
plastid RNAP	72	50	49 (9)
Eukaryota	72	50	49 (9)
Totals	**1123**	**955**	**854 (350)**

Note: parentheses indicate the number of sequences after a 95% sequence identity cutoff.

Bacterial RNAP β/β′ SCA

In order understand the degree of co-evolution within and between the two large subunits of bacterial RNAP, we determined both the intra and inter-protein SCA Coupling Matrixes for β and β′ (Figure 4.3). We then clustered the intra-protein matrixes, followed by reordering of the inter-protein SCA Coupling Matrix using the intra-protein clustering order (Figure 4.4A). This result indicated that the various intra-protein clusters within one subunit couple to corresponding intra-protein clusters within the other subunit.

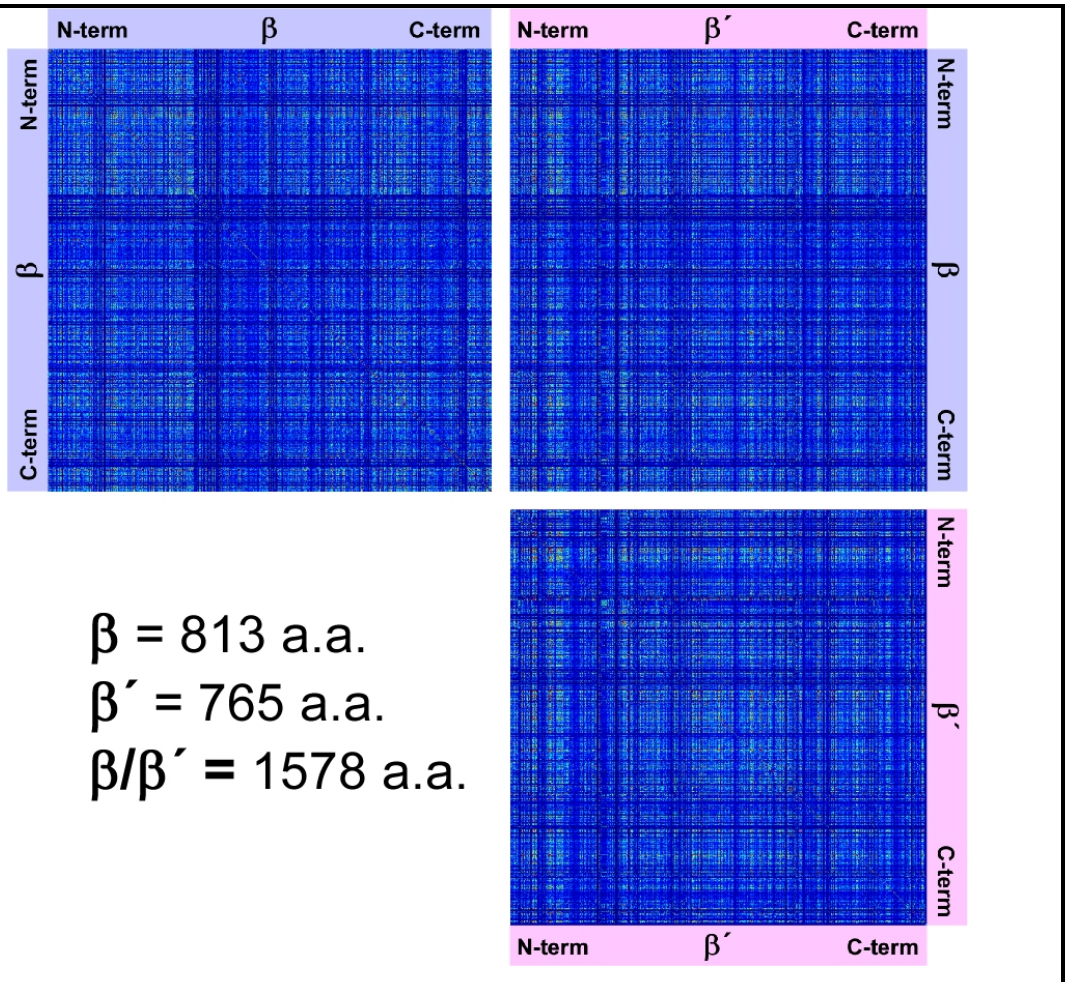


Figure 4.3 – β and β′ Intra and Inter-Protein SCA Coupling Matrixes.
β (light blue) and β′ (pink) intra (top left, bottom right) and inter-SCA (top right) Coupling Matrixes. The matrixes are unclustered and ordered from N to C-term as indicated.

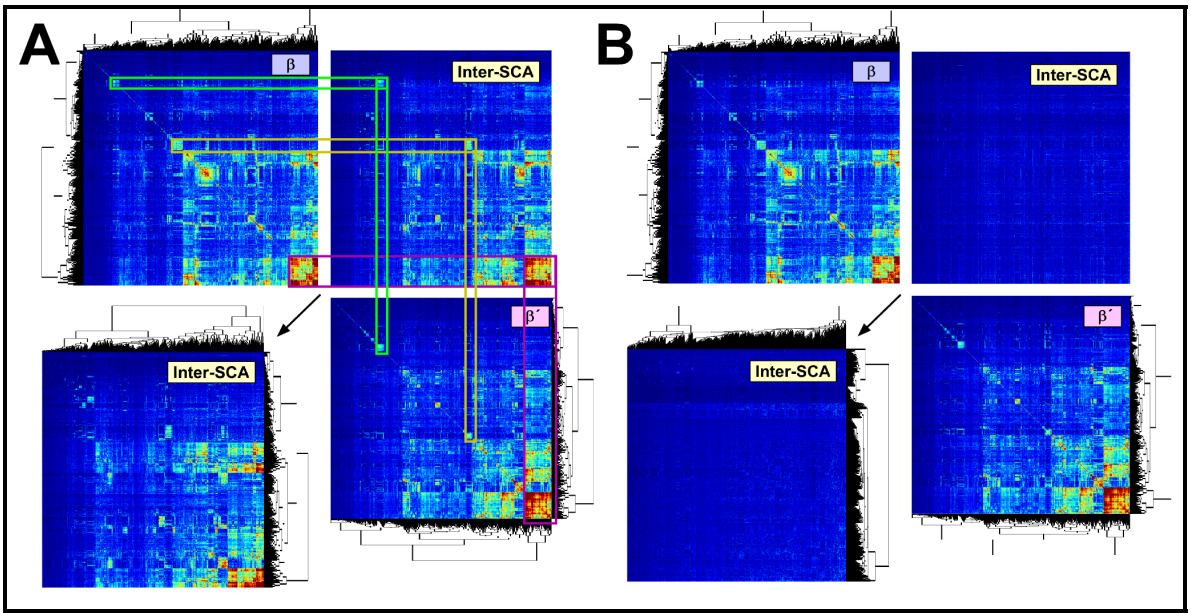


Figure 4.4 – β and β′ Clustered Intra and Inter-Protein SCA Coupling Matrixes.

(A) Clustered β and β′ intra-SCA Coupling Matrixes (top left, bottom right) with inter-SCA Coupling matrix re-ordered with β and β′ clustering. Clustered inter-SCA Coupling Matrix (bottom left). The three colored rectangles indicate intra-SCA clusters that couple through inter-SCA clusters.

(B) Same as (A), but done as an Inter-SCA coupling negative control where the β and β′ species pairing was shuffled. The inter-SCA Coupling Matrixes are empty (top right, bottom left), indicating that the inter-SCA coupling only occurs across species matched β and β′ sequences and is not simply an artifact of joining any two β and β′ sequences together.

As a result of this large degree of inter-protein co-evolution, we decided that the best approach was to simultaneously analyze β and β′ together by creating a combined β/β′ MSA (Figure 4.5A). Furthermore, as a negative control we mixed up the species pairing and, as expected, the inter-protein co-evolution disappeared (Figure 4.4B, Figure 4.5B). Figure 4.6 shows the hierarchal clustering for the combined bacterial β/β′ SCA Coupling Matrix (total of 1,578 residue positions), in which the protein positions have been clustered to reveal common patterns of co-evolution.

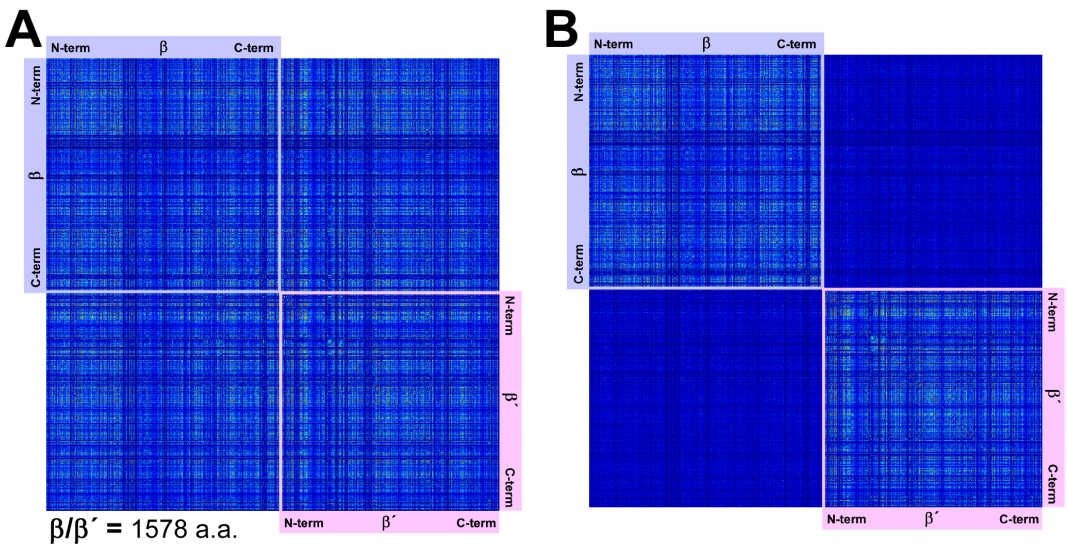


Figure 4.5 – Combined β/β′ Intra and Inter-Protein SCA Coupling Matrix.

(A) β (light blue) and β′ (pink) alignments were combined by species matching. This allowed for the simultaneous determination of intra and inter-protein SCA Coupling Values. The matrix is unclustered and ordered from N to C-term as indicated.

(B) Same as (A), but the β and β′ species pairing was shuffled. The inter-SCA, but not the intra-SCA Coupling Values empty (top right, bottom left).

By combining hierarchal clustering with the more mathematically rigorous Independent Component Analysis (ICA) (Figure 4.7), we delineated six distinct co-evolving groups (Group1 to Group 6). We also used the clustered SCA Coupling Matrix to determine the inter-group connectivity, allowing us to understand how the various groups co-evolved with each other (Figure 4.6 inset). Group 1 and Group 2 evolve completely independently of the large cluster Group 6. In contrast, Group 3, Group 4, and Group 5 all connect to each other as well as to Group 2 and all except Group 3 connect to Group 6. Therefore, despite being distinct units of co-evolution all of the six cluster groups do relate to each other either directly or indirectly.

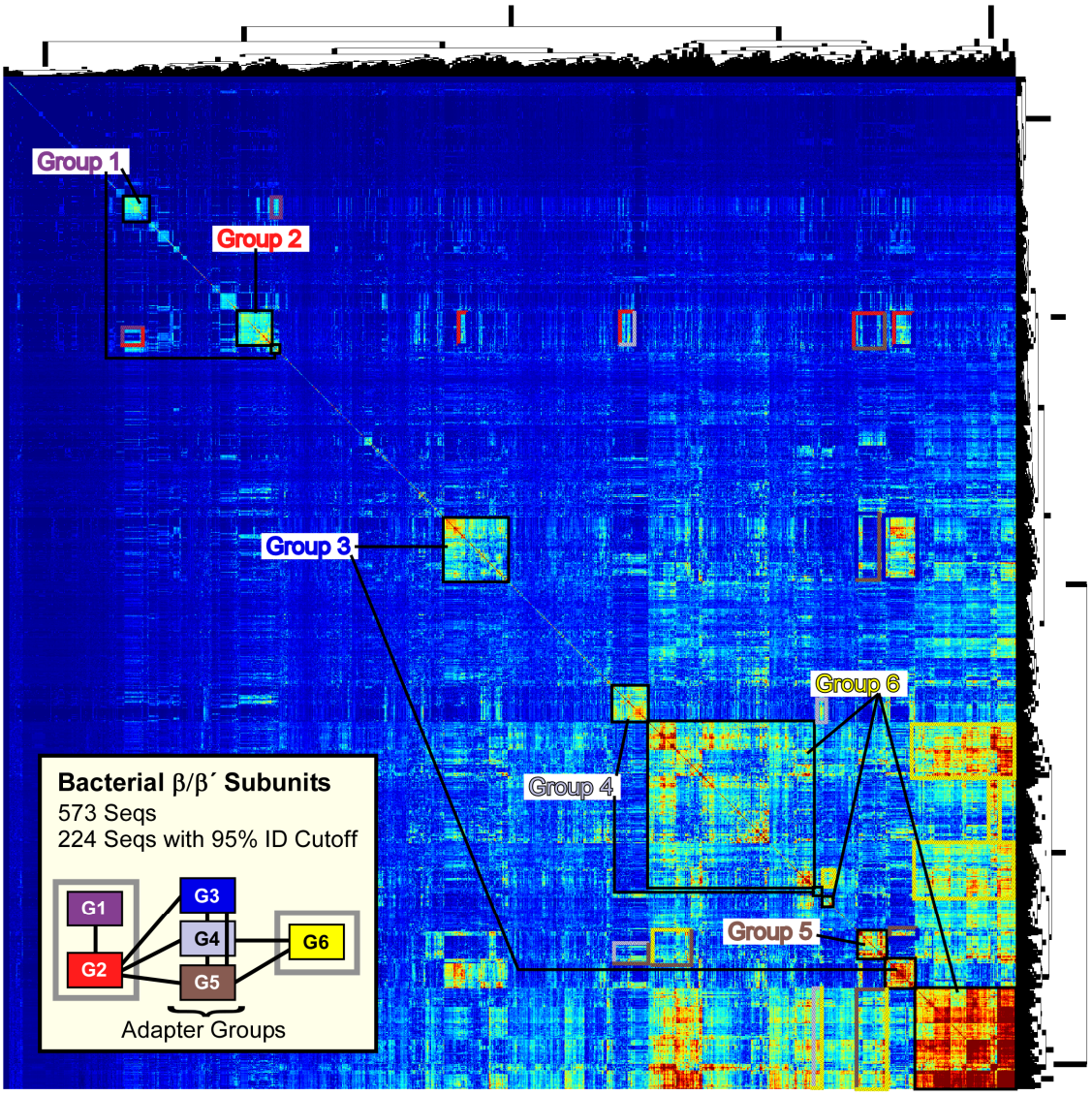


Figure 4.6 – Clustered Bacterial Combined β/β′ SCA Coupling Matrix.

The Bacterial pair-wise SCA Coupling values were re-ordered using hierarchical clustering. The SCA coupling value was represented by a pixel colored from blue (no coupling [0]) to red (high coupling [≥2]). Due to the nature of the clustering the main clusters are located on a diagonal line from the top left to the bottom right corner. The degree to which any two residues are coupled can be determined by looking at the intensity of the symmetric off-diagonal pixels that would join the residues by moving left/right and up/down. We identified six groups using a combination of clustering and ICA (Group 1–6 or G1–G6). We also used the off-diagonal positions to determine the level of coupling between the cluster groups. Single color stippled boxes indicate intra-group off-diagonal coupling and two colored stippled boxes indicate inter-group off-diagonal coupling. Using this information we created diagrams of how the various groups co-evolved to each other (inset).

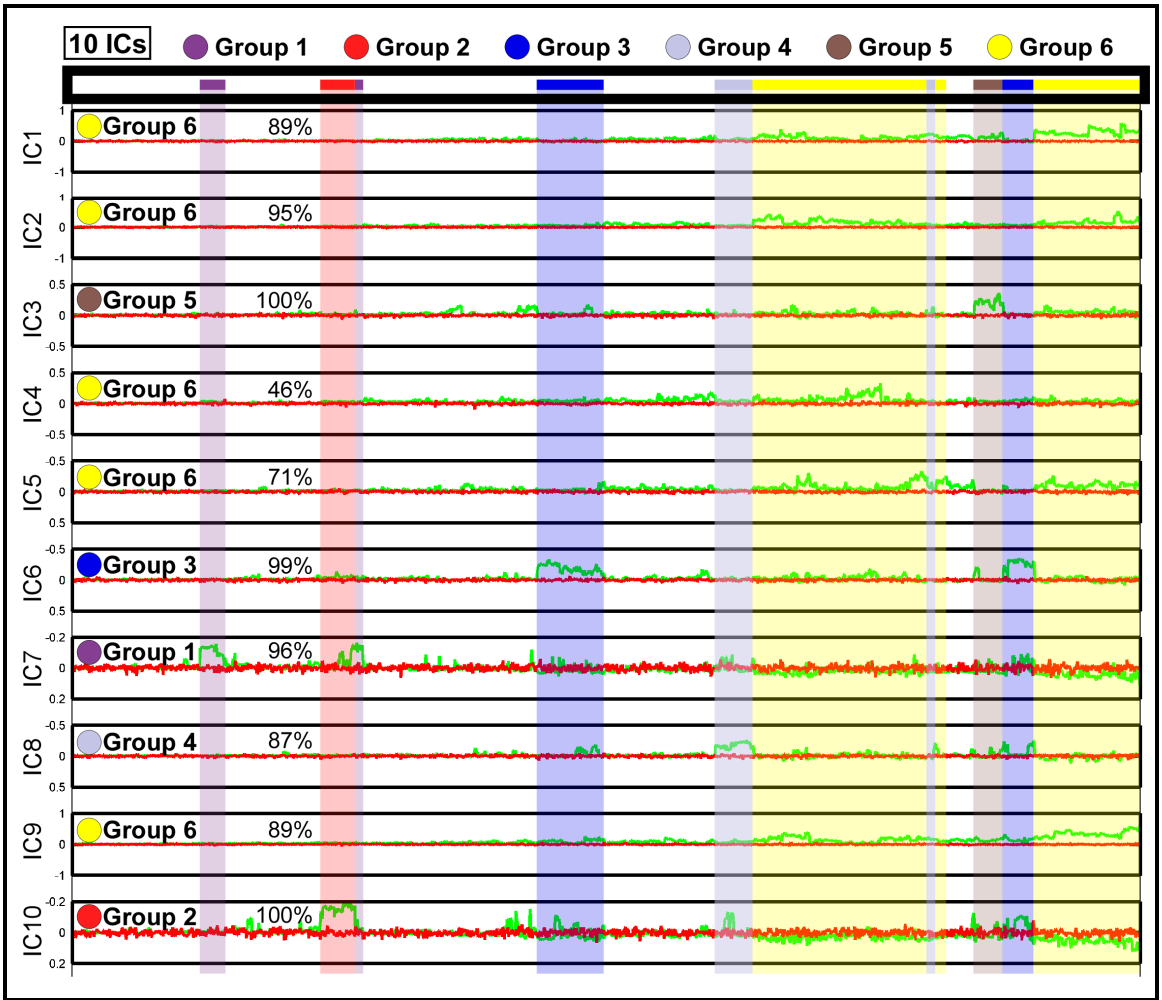


Figure 4.7 – Bacterial Combined β/β′ SCA Coupling Matrix ICA.

The Bacterial combined β/β′ SCA Coupling Matrix was analyzed by ICA. The above shows the results when ICA was told to expect 10 Independent Components (IC). We constructed plots of ICA signal value vs. residue position for each IC (green plot) and for a negative control from a randomized SCA coupling matrix (red plot). We next re-order these plots using the residue ordering given by the hierarchal clustering. This approach allowed us to directly compare if the hierarchal cluster groups corresponded to the ICA groups. The hierarchal clustering group boundaries are indicated by transparent vertical regions. The hierarchal clustering group identity of each IC is shown on each plot, with percentage of residues from the hierarchal clustering group found in the ICA group indicated.

The six groups were mapped to the *Tth* bacterial RNAP ternary elongation complex (TEC) structure. Based on the inter-group connectivity and the structural mapping there seems to be two kinds of groups: (1) main groups which evolve

independently and map to structurally contiguous or distinct regions, and (2) adapter groups which evolve between two main groups and map to sparsely separated regions within the structure. The main groups include the closely related Group 1 and Group 2 and the orthogonal or independently evolving Group 6. The adapter groups include Group 3, Group 4, and Group 5. Group 1 maps exclusively to the left side of RNAP. Group 2 maps to the internal surface of RNAP around the DNA and RNA in the TEC. Group 6 maps to an extensive area enclosing the core of RNAP. The adapter networks Groups 3, 4, and 5 sparsely across RNAP. Unfortunately, since SCA requires amino acid diversity to evaluate a given residue position, the extensive sequence conservation over a large portion of β/β′ rendered it inaccessible for study. There were 407 residue positions with an Information Score [89] >0.98 (0.00 for no conservation to 1.00 for complete conservation), which empirically was the highest Information Score among those residue positions that still showed some SCA coupling.

All RNAP Large Subunit SCA

In order, to overcome the limitations imposed by the extensive conservation within the Bacterial β/β′ MSA, we increased the diversity of the alignment by adding the β and β' homologues from non-bacterial RNAPs including: eukaryotic pol I/II/III RNAPs, Nuclear-Cytoplasmic Large double-stranded DNA Viruses (NCLDV) pol like RNAPs, the archeal pol II like RNAPs, and plant plastid RNAPs.

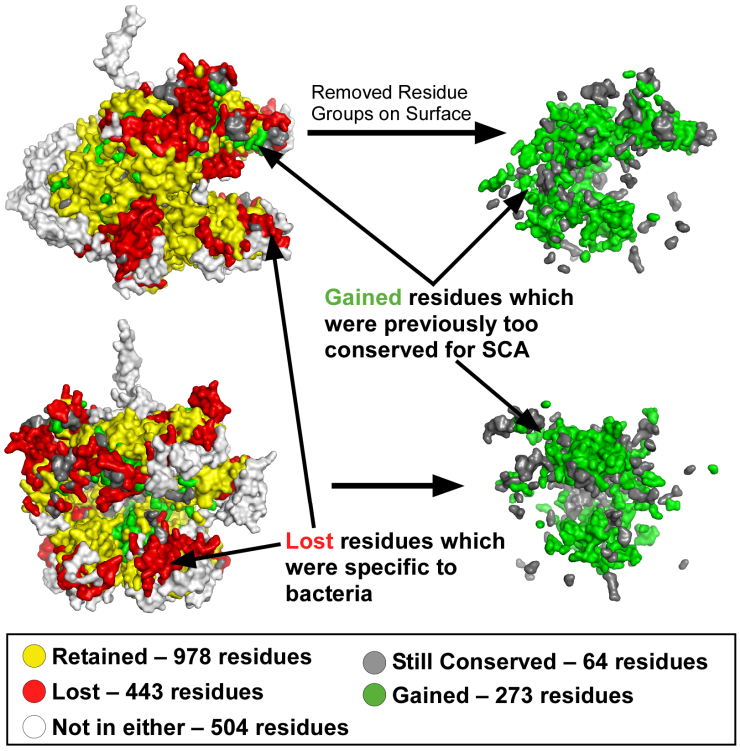


Figure 4.8 – Diversity Changes From Adding Non-Bacterial Large Subunits.
When the non-bacterial large subunits were added to the bacterial β/β′ alignments there were diversity changes at various positions within the alignment. Based on the diversity changes some residues were lost (red), gained (green), or retained (yellow) in terms being analyzable by SCA. There were some residues there were still not included (white) and too conserved (gray). These change are colored on the molecular surface of the bacterial β/β′ subunits.

The increased diversity of the resulting All Large Subunit alignment allowed 273 additional positions in the alignment to be examined by SCA, while 443 positions were lost since they were bacterial specific (i.e. they did not structurally

align with the yeast RNAP II structure). In addition, 978 positions from the bacterial SCA analysis were retained (for a total of 1,251 residue positions). 64 positions were still too conserved in the All Large Subunit alignment to be included in the SCA analysis (Figure 4.8). Importantly, the extra residues accessible to the SCA analysis for the All Large Subunit alignment (green in Figure 4.8) mostly cluster around the active site of the enzyme, and so are likely to be involved in fundamental, highly conserved aspects of RNAP function, which we are most interested in analyzing. The residues lost (due to the lack of structural alignment with yeast RNAP II, red in Figure 4.8) tend towards the periphery of the structure and are not likely to be of universal importance.

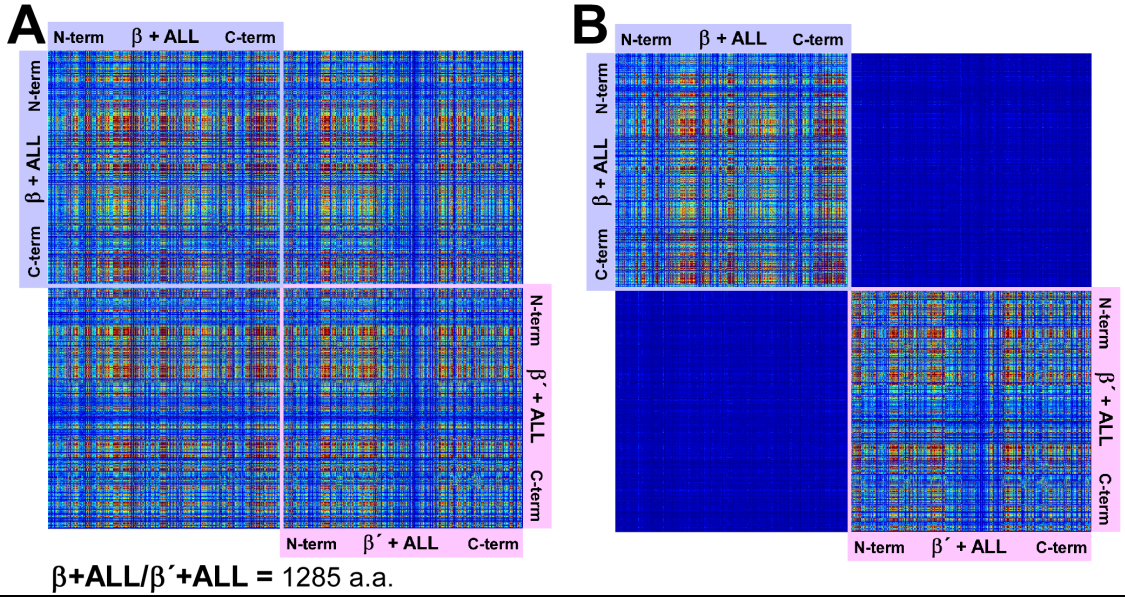


Figure 4.9 – Combined All Large Subunits Intra and Inter-Protein SCA Coupling Matrix.

(A) β+ALL (light blue) and β′+ALL (pink) alignments were combined by species matching. This allowed for the simultaneous determination of intra and inter-protein SCA Coupling Values. The matrix is unclustered and ordered from N to C-term as indicated.

(B) Same as (A), but the β and β′ species pairing was shuffled. The inter-SCA, but not the intra-SCA Coupling Values empty (top right, bottom left).

The All Large Subunit SCA allowed us to detect residue networks that were either invisible or only partially visible in the Bacterial β/β′ SCA. The All Large Subunit SCA Coupling Matrix (Figure 4.9) was analyzed by using a combination of clustering (Figure 4.10) and ICA (Figure 4.11) in order to define the co-evolving groups (aGroup 1 to aGroup 6) and their interconnections (Figure 4.10 inset). Similar to the Bacterial β/β′ SCA, there were two orthogonal independently evolving main groups (aGroup 1 and aGroup 5/6) connected by adapter groups (aGroup 2 and aGroup 4). aGroup 1 maps around the innermost conversed core of RNAP as well as to the trigger loop. aGroup 3 is a remnant of the Bacterial β/β′ SCA Group 6. aGroup 5 is an extension of the Bacterial β/β′ SCA Group 2, mapping mostly to the interior surface of RNAP. The additional positions contained within aGroup 5 were previously too conserved to be detected in the Bacterial β/β′ SCA.

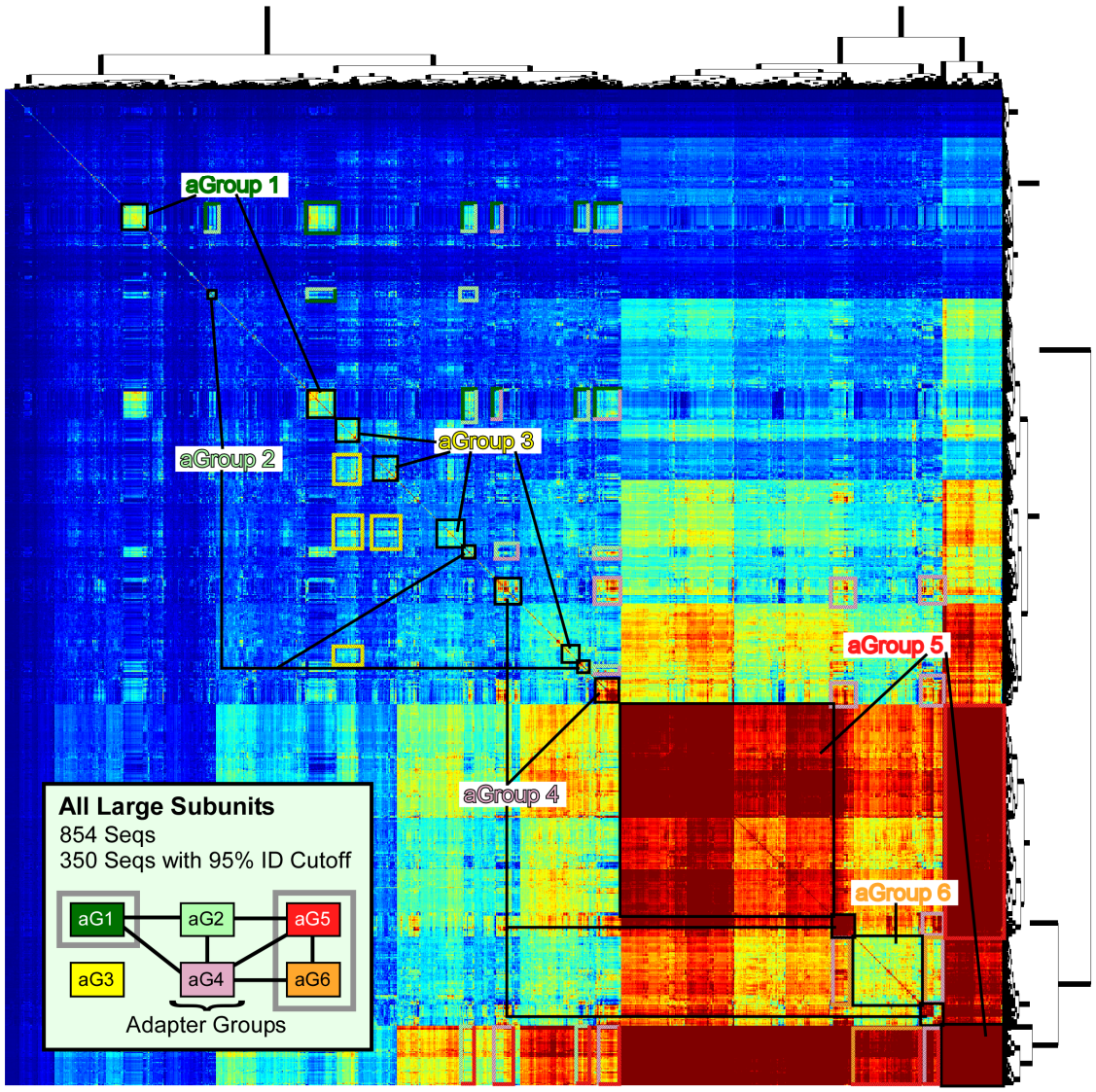


Figure 4.10 – Clustered Combined All Large Subunits SCA Coupling Matrix.

The Combined All Large Subunits SCA Coupling values were re-ordered using hierarchical clustering. We identified six groups using a combination of clustering and ICA (aGroup 1–6 or aG1–aG6). We also used the off-diagonal positions to determine the level of coupling between the cluster groups. Using this information we created diagrams of how the various groups co-evolved to each other (inset).

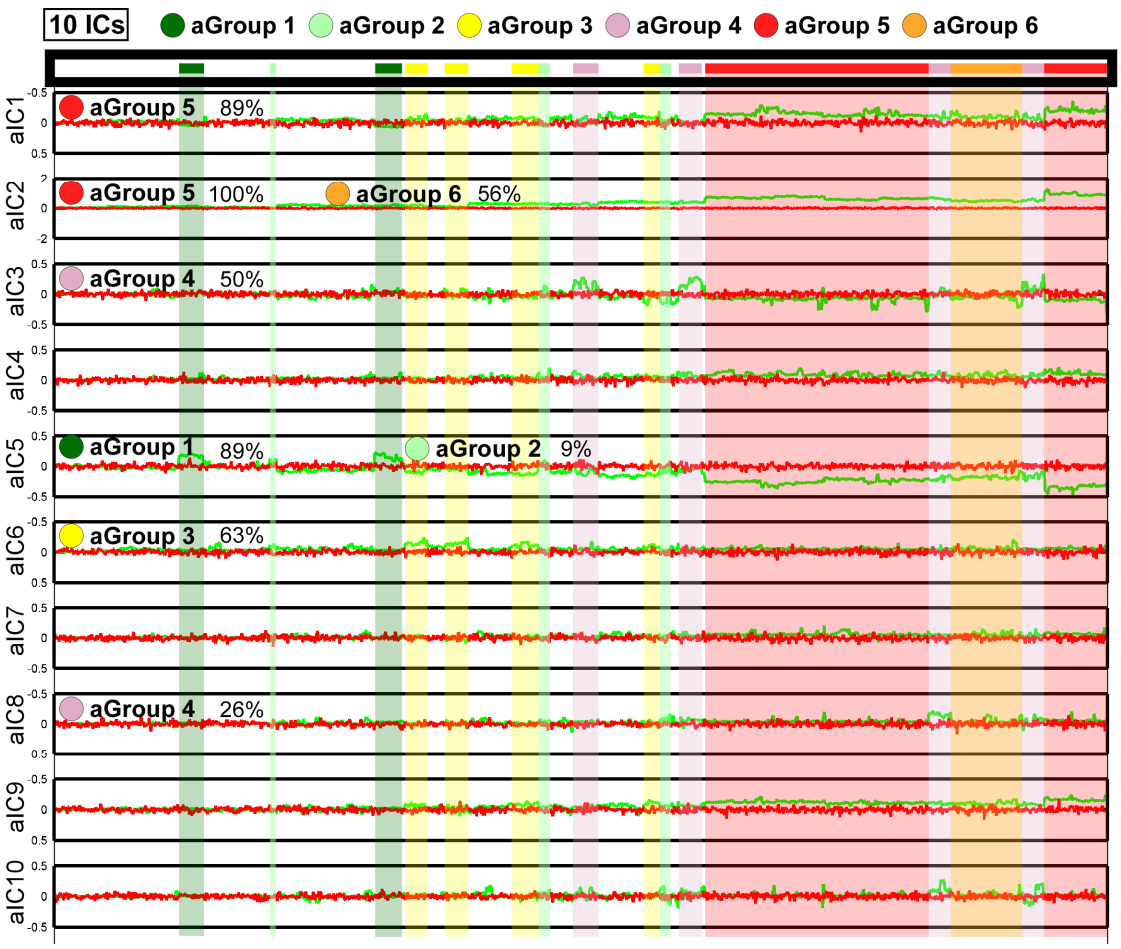


Figure 4.11 – All Large Subunit SCA Coupling Matrix ICA.

The Combined All Large Subunits SCA Coupling Matrix was analyzed by ICA similar to Figure 4.7

Furthermore, since aGroup 5 and Group 2 represent the same residue network we used them to combine the Bacterial SCA and the All Large Subunit SCA results into one large connectivity diagram (Figure 4.12). Using the structural mapping for the main groups (Figure 4.12), we propose three distinct functional roles: (1) Catalysis, (2) DNA/RNA interactions during initiation, elongation, and termination, and (3) structural stability/scaffolding.

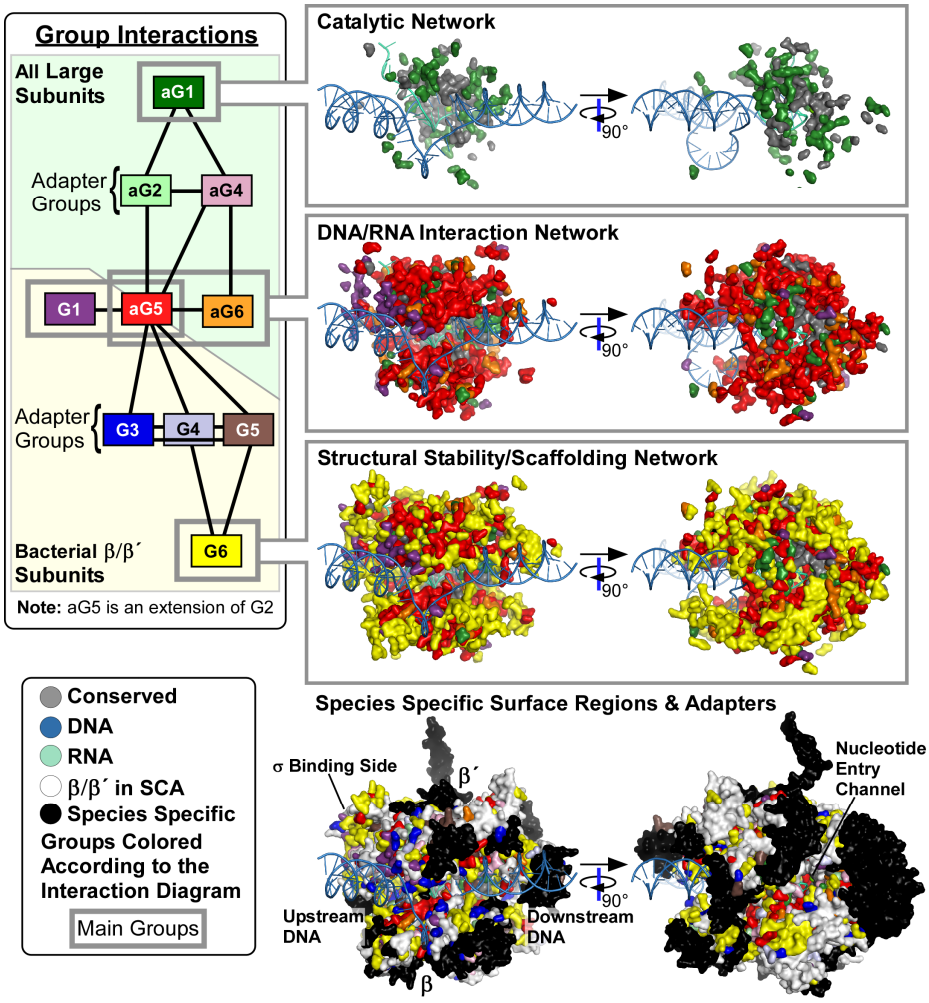


Figure 4.12 – Overview of the RNAP Large Subunit Co-evolution Networks.

The co-evolution group networks from the Bacterial and the All Large Subunit SCA were connected using the common DNA/RNA Interface Group found in both studies (Group 2 and aGroup 5). There are three sets of main groups with the following proposed functions: catalysis, DNA/RNA interactions, and structural stability/scaffolding. The three main groups are connected to each other via adapter groups. The structural mapping of the main groups on the *Tth* large subunits (β/β′) highlights the progression of the main groups from the inner most catalytic core to the outermost exterior of RNAP. The proteins are shown as molecular surfaces with the β and β′ subunits white. The TEC model DNA (blue) and RNA (teal) are shown as phosphate-backbone ribbons, with bases indicated schematically as sticks. The residues conserved across all classes of RNAP are shown as grey molecular surfaces. The main groups are shown as molecular surfaces: catalytic group (aGroup 1/green), DNA/RNA Interaction Groups (aGroup 5/red, aGroup 6/orange, Group 1/purple), and the Bacterial specific structural scaffolding group (Group 6/yellow). The *Tth* large subunit residues not aligned in the bacterial β/β′ MSA are shown as black molecular surfaces.

Catalysis

Given that aGroup 1 maps closely to the completely conserved catalytic core, including the DNA/RNA hybrid and the trigger loop, it might be involved in the catalytic mechanism by which RNAP produces RNA (Figure 4.13). Recently, it has been shown that conformational changes in the trigger loop are important for nucleotide recognition and catalysis [9, 90].

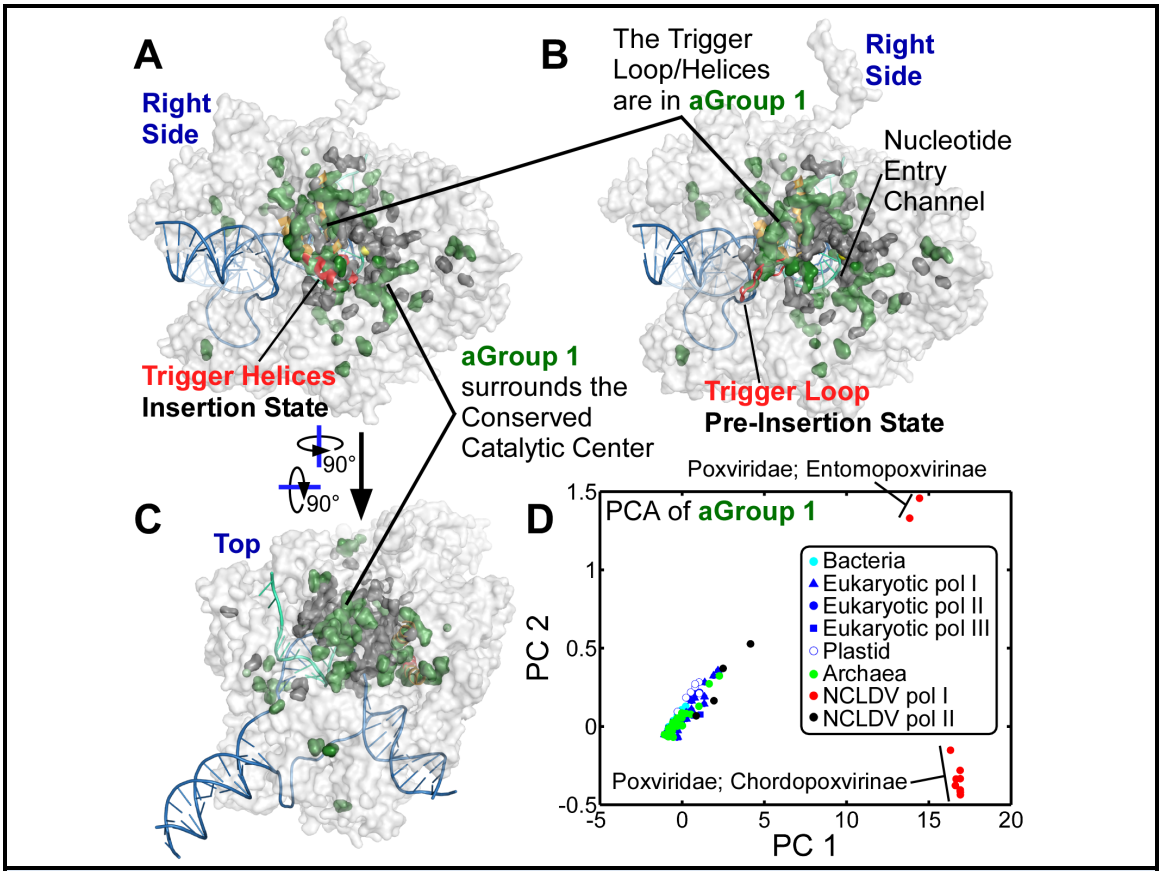


Figure 4.13 – Structural Mapping of the Catalytic Group.

(A/B/C) The Catalytic Group (aGroup 1/green) was mapped onto the *Tth* RNAP structure. The Catalytic Group is located in the center to RNAP surrounding the conserved catalytic core, the secondary channel, and the trigger loop/trigger helices (red). The proteins are shown as molecular surfaces with the β and β′ subunits transparent white. The TEC model DNA (blue) and RNA (teal) are shown as phosphate-backbone ribbons, with bases indicated schematically as sticks. The residues conserved across all classes of RNAP are shown as grey molecular surfaces. **(D)** PCA on the residues in aGroup1. A plot of the first two principle components (PC1/PC2) shows that sequences from the NCLDV pol I RNAPs segregate from the rest.

Principal Component Analysis (PCA) was used to determine groups of similar and dissimilar sequences in regards to the residue positions that make up aGroup 1. PCA showed that the NCLDV pol I RNAPs segregated from the rest of the RNAPs, indicating that the detection of aGroup 1 was the result of differences between the NCLDV pol I RNAPs and the rest (Figure 4.13D). Interestingly, without the NCLDV pol I sequences the positions in aGroup 1 are too conserved to be accessible to SCA. It is also important to note that this group is not simply an artifact of the NCLDV sequences being different from the other sequences in the alignment, since the NCLDVs also contains positions that are identical in all RNAPs as well as variable positions that vary across the RNAPs.

Similar to Group 2, aGroup 1 might represent the fringes of a larger group which in this case most likely includes highly conserved positions within the core of RNAP. Nonetheless, the NCLDVs altered these positions upon acquiring pol I from their eukaryotic hosts. It is possible that the modified aGroup 1 residues are biologically relevant, possibly allowing the NCLDVs to enhance or escape some intrinsic or regulated aspect of RNAP catalysis.

DNA/RNA Interactions: Initiation, Elongation, and Termination

Group 1, Group 2, Group 5, and aGroup 6 map to the interior surface of RNAP and are presumably involved in the RNAP interactions with the DNA and RNA. If RNAP is split in half and viewed from the inside of the active site channel, it can be seen that aGroup 5 follows the path of the DNA transcription bubble and interior wall of the RNA exit channel with exquisite precision (Figure 4.14). aGroup 5 also extends onto the floor of the nucleotide entry channel

possibly linking the entry of nucleotides with the movement of DNA/RNA in the transcription cycle or possibly highlighting its involvement in backtracking (when a 3'-fragment of the RNA transcript extrudes into the nucleotide entry channel (Figure 4.14 Bottom Inner Surface).

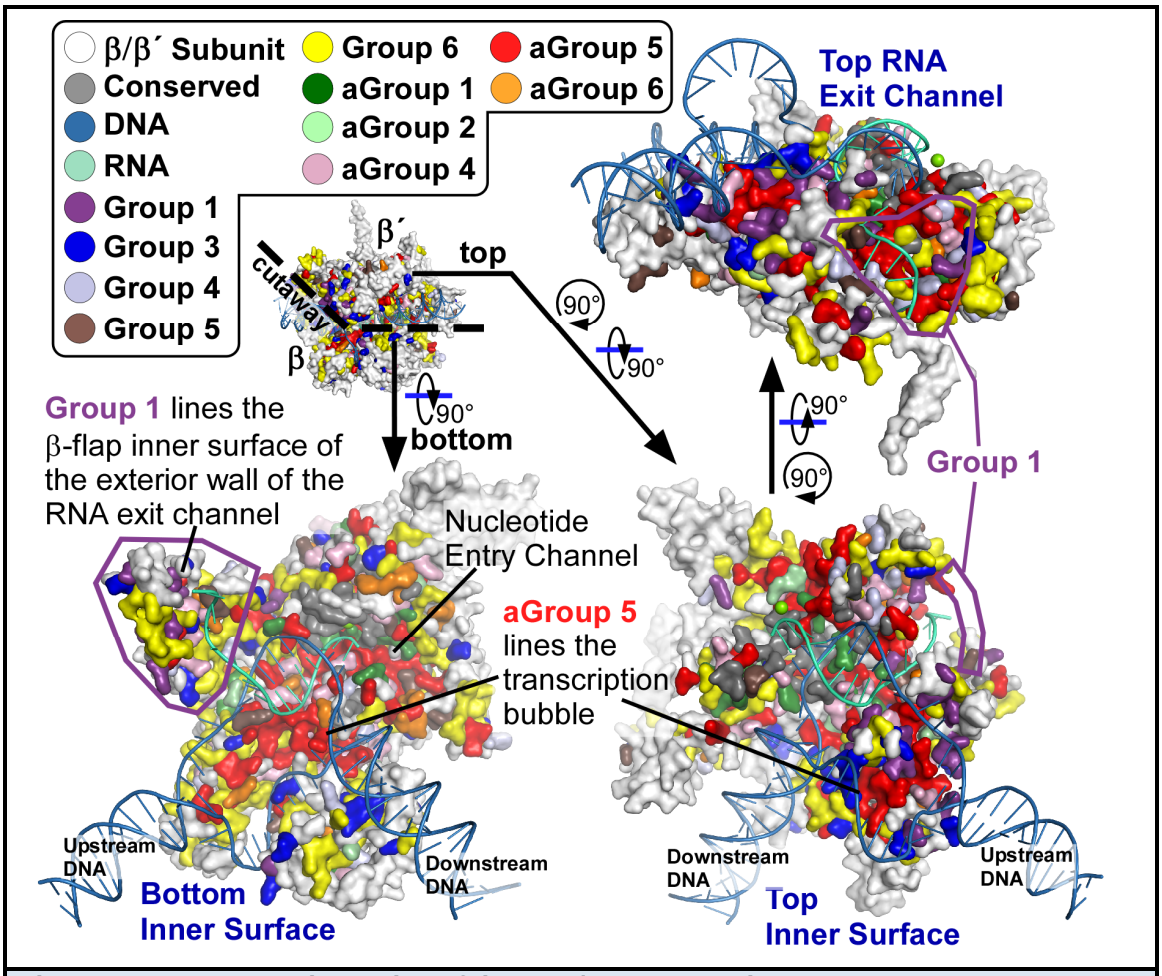


Figure 4.14 – Structural Mapping of the DNA/RNA Interaction Groups.

aGroup 1–6 were mapped onto the *Tth* RNAP structure. In order to better see the relationship between the DNA/RNA and aGroup 5 RNAP was cut in half in order to see interior surfaces of the bottom and top half as well as the RNA exit channel. aGroup 5 maps to the surfaces of β and β′ where the DNA and RNA are located. The proteins are shown as molecular surfaces with the β and β′ subunits white. The TEC model DNA (blue) and RNA (teal) are shown as phosphate-backbone ribbons, with bases indicated schematically as sticks. The residues conserved across all classes of RNAP are shown as grey molecular surfaces. The location of the Group 1 (purple) β-flap residues are shown by the purple region.

Furthermore, it is possible that the bacterial specific Group 1, which connects to the main aGroup 5 DNA/RNA Interaction Group, is involved in a separate but connected set of DNA interactions (Figure 4.15).

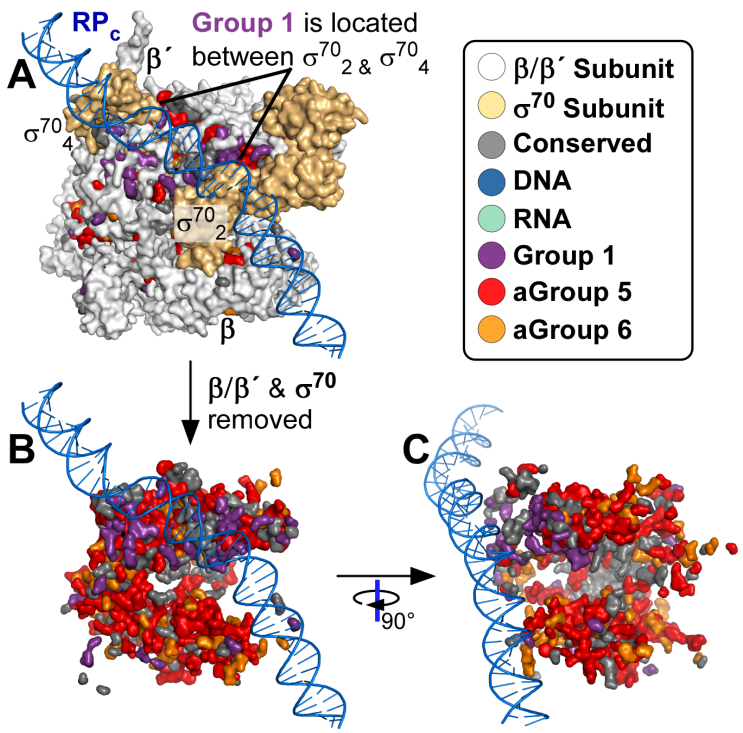


Figure 4.15 – Structural Mapping of the DNA Interaction Group Specific to Bacterial RNAP.

(A) The bacterial specific Group 1 (purple) was mapped onto the *Tth* RNAP structure. Group 1 exclusively maps to the σ side of RNAP holoenyzme, in particular between the DNA binding domains of σ^{70} (σ^{70}_2 and σ^{70}_4). The proteins are shown as molecular surfaces with the β and β′ subunits white and the σ subunit light orange. The TEC model DNA (blue) and RNA (teal) are shown as phosphate-backbone ribbons, with bases indicated schematically as sticks. The residues conserved across all classes of RNAP are shown as grey molecular surfaces. The main DNA/RNA Interaction Group (aGroup 5/red) as well as its closely related offshoots aGroup 6 (orange) and the bacterial specific Group 1 (purple) are shown as molecular surfaces.

(B/C) Same as A, but with σ (light orange) and the rest of β/β′ (white) has been removed. From this you can see the proximity of the main DNA/RNA Interaction Group (aGroup 5/red) as well as its closely related offshoots aGroup 6 (orange) to the bacterial specific Group 1 (purple).

Group 1 is located between the promoter binding sites of the closed complex RNAP holoenzyme such that the spacer DNA between the -35 and -10 promoter elements would lay directly over it (Figure 4.15A). It is therefore possible that Group 1 plays a role in initiation, such as in propagating structural changes during the isomerization steps in which the DNA template strand relocates to the catalytic center of RNAP.

Even more importantly, Group 1 might also play a role in destabilizing the transcription bubble during the process of intrinsic transcription termination. In fact, previous modeling of the intrinsic termination RNA hairpin has hinted that the unstable process of initiation and termination might involve similar structural rearrangements [8]. Therefore, it is possible that Group 1 might play a role in both initiation and termination. Group 1 stretches from the β-flap to the upstream edge of the transcription bubble (Figure 4.14A). Group 1 residues on the β-flap could sense the RNA hairpin formed during intrinsic termination. This signal could then propagate through Group 1 to the β′ coil-coil and rudder, thereby destabilizing the transcription bubble.

As shown in Figure 4.16B, PCA on the residue positions that make up Group 1 greatly separated obligate intracellular bacterial pathogens belonging to the Anaplasmataceae and Wolbachieae families in the order Rickettsiales. In addition other obligate pathogic bacteria from the Spirochaetes, Tropheryma, and Chlamydiae families, along with radiation resistant and hyperthermophlic bacteria from the Deinococcus-Thermus and Aquificae families were segregated (PC1 value > 0) from the rest of the bacteria (PC1 value < 0).

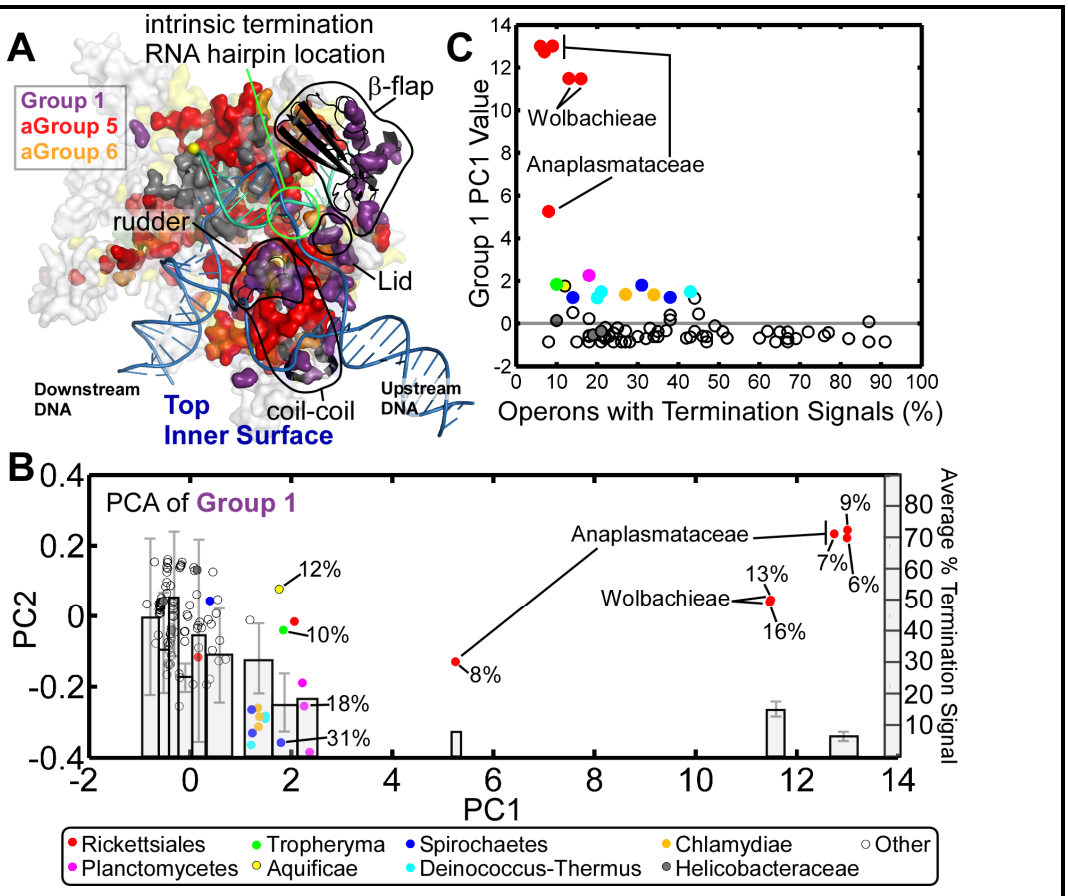


Figure 4.16 – Bacterial Intrinsic Termination Group.

(A) The Termination Group (Group 1/purple) was mapped onto the *Tth* RNAP structure. The Termination Group is located around upstream edge of the transcription bubble in the β-flap, lid, and coil-coil (rudder) structural domains. The view is the same as the Top Inner Surface representation in Figure 4.14, except that the molecular surfaces of the β and β′ subunits and non-DNA/RNA interaction groups are transparent. **(B)** PCA on the residues in Group1. A scatter plot of the first two principle components (PC1/PC2), with the location of sequences belonging to important bacterial families indicated by colored dots (including hyperthermophiles, obligate pathogens and endosymbionts). Sequences from the bacterial species with a low percentage of operons with intrinsic termination signals segregate from the rest. The overlaid bar graphs show the average value of predicted % of operons with termination signals for the species within the width of each bar, with error bars indicating standard deviation. **(C)** Scatter plot of PC1 value according to the percentage of operons with intrinsic termination signals, with important sequences colored coded as in (B). A transparent gray line indicates a PC1 value cutoff of 0. Sequences above this line have a lower percentage of operons with termination signals and essentially all species with >50% of operons with termination signals are bellow this line.

It has been previously indicated that obligate pathogens and hyperthermophlic bacteria have a lower reliance on intrinsic termination [91]. In fact, as shown in Figure 4.16C, the PC1 values show a meaningful relationship with results from a recent study that predicated the percentage of operons with termination signals for 343 bacteria [92]. The obligate pathogens, hyperthermophlic, and radiation resistant bacteria which were segregated by the PC1 values were also predicated to have a low % of operons with termination signals. Interestingly, the Anaplasmataceae posses some of the smallest predicated % of operons with termination singles. In addition, almost none of the bacteria with % of operons with termination signals >50% have PC1 values > 0. Therefore, we believe that Group 1 was detectible since it represents a network of residues that co-evolved to facilitate intrinsic termination in those species that rely on intrinsic termination signals.

Structural Stability/Scaffolding

The exact role of Group 6 (aGroup 3) is less clear. Based on its structural mapping it encloses the other main Groups. It is possible that this group serves a role in structural stability of the bacterial RNAP. Though not detectable by our analysis, it might also connect possibly via adapters to the outer surface of RNAP including the various species specific insertions. Unfortunately, due to the necessity for a large and diverse set of sequences the species specific insertions were not included in our analysis. Nonetheless, this would make sense given the paradigm by which the rest of RNAP is pieced together.

Adapter Groups

SCA has previously been performed on single domain proteins, only revealing networks similar to our main groups. However, we have shown that RNAP also contains adapter groups which appear to coordinate the independently evolving main groups. The absence of adapter groups in previous SCA work is suggestive that in general, relatively small, single-domain proteins do not require the presence of accessory adapter group networks. Perhaps as a protein acquires multiple functions while evolving into a large, multi-domain assembly, it requires the coordinating activity of adapter groups. It is important to note that the adapter groups are capable of co-evolving despite being structurally sparse (non-contiguous), since the distant adapter group positions are, in a sense, coupled with each other through their interactions with the main groups.

Omega the Inter-Protein Molecular Adaptor

Although β and β' account for the majority of the bacterial RNAP mass and include all the elements of the active site, it might be informative to examine the other RNAP subunits including α, ω, and σ^{70}. Unfortunately, the large degree of conservation within the bacterial sequences would severely limit the results. However, given these limitations, ω still represents an ideal candidate for analysis since it also has a pol II homologue known as Rpb6.

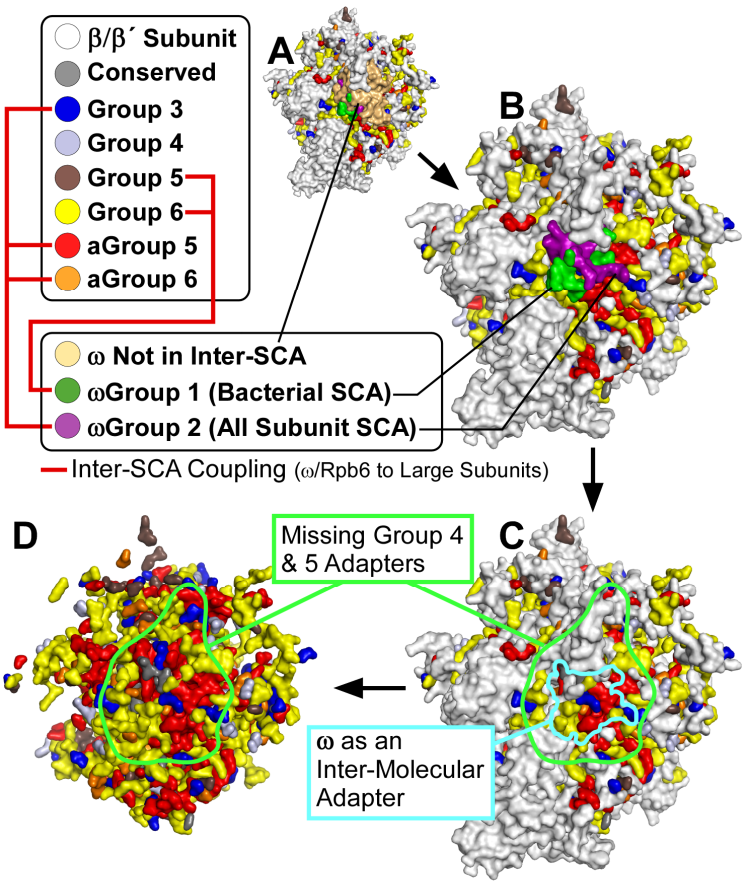


Figure 4.17 – Structural Mapping of ω the Inter-Molecular Adapter.

ω has two main intra-protein SCA Groups (ωGroup 1/green and ωGroup2/purple) each of which has separate inter-protein couplings to distinct β/β′ groups found previously (red lines).

(A) The proteins are shown as molecular surfaces with the β and β′ subunits white and the ω subunit in light orange. The residues conserved across all classes of RNAP are shown as grey molecular surfaces.

(B) Same as (A), but the ω residues not involved in any inter-protein coupling removed.

(C) Same as (B), but all of the ω residues have been removed. The region on β′ where ω binds is highlighted by a cyan border. Interestingly the surface of β′ where ω binds agrees with the predicted inter-protein group interactions.

(D) Same as (C), but the rest of β/β′ (white) has been removed. The green border indicated a large local region that is the only area on β/β′ where the intra-molecular Adapter Groups 4 and 5 are missing. As seen in (C), ω binds directly in the center of this region suggesting that ω serves as a local inter-subunit molecular adapter taking the place of the locally missing intrinsic adapters Group 4 and 5.

In order to examine the intra-protein SCA we made two single protein alignments: (1) ω alone and (2) ω plus Rpb6. In order to examine the inter-protein SCA we made two combined tri-subunit alignments: (1) bacterial RNAP β/β′ with ω and (2) a larger alignment including the bacterial β/β′ with ω plus the pol II large subunits with their corresponding Rpb6 sequences. The intra-protein ω SCA indicated that ω has two main intra-protein SCA Groups, ωGroup 1 (from the Bacterial ω SCA) and ωGroup 2 (from the Bacterial ω plus pol II Rpb6 SCA). In addition, despite being derived from separate SCA results, it is apparent that ωGroup 1 and ωGroup 2 are not evolutionarily independent of each other, since they share six common residues (33/38% of their total residues respectively) at their structural interface. This level of overlap indicates that they would presumably be connected to each other if it was possible to otherwise overcome the conservation limits of SCA and fully detect both groups in the same SCA analysis.

The tri-subunit SCA results indicated that ωGroup 1 and ωGroup 2 from the intra-protein SCA also display inter-protein co-evolution with the two large subunits of RNAP (Figure 4.17). ωGroup 1 has inter-protein co-evolution to the Bacterial β/β′ SCA Structural Scaffolding Group (Group 6) and ωGroup 2 has inter-protein co-evolution to the All Large Subunit SCA DNA/RNA Interaction Group (aGroup 5) and the Adapter Groups 3, 4, 5. Furthermore, ωGroup 1 and ωGroup 2 seem to directly interact with corresponding β/β′ surface patches from the Bacterial Structural Scaffolding Group (Group 6) and the DNA/RNA Interaction Group (aGroup 5) and the Adapter Group 3.

Furthermore, the region of β/β′ where ω binds is conspicuously missing the Adapter Groups 4 and 5, therefore locally there would be no intrinsic β/β' connection between the DNA/RNA Interaction Group (aGroup 5) and the Bacterial Structural Scaffolding (Group 6). However, given that ω co-evolves with these locally unconnected main groups, it seems that ω acts as an inter-subunit molecular adapter capable of taking the place of the locally missing Group 4 and Group 5 intra-β/β′ molecular adapters (Figure 4.17C, D). In fact, the only adapter group present is Group 3, which needs Group 4 or Group 5 to fully connect aGroup 5 and Group 6. The idea that ω might serve as an adapter in order to tie together two of the main RNAP groups might explain why ω is necessary for the activity of ppGpp, which likely acts at sites in RNAP distant from ω.

Conclusions

Our analysis has elucidated fundamental concepts regarding the evolution of the multi-subunit RNAPs (Figure 4.12). Importantly we have determined that evolution uses two general types of co-evolving residue networks, which we have termed main and adapter groups. The main groups are functionally important for key aspects of transcription, including catalysis and RNAP interactions with DNA and RNA during initiation, elongation, and termination. The main groups evolve independently of each other, allowing them to become optimized for their distinct roles. In order to allow the main groups to be uncoupled from each other, evolution uses adapter groups to maintain the necessary coordination between the independently evolving main groups. In addition, we have shown that evolution also employs accessory subunits to act as surrogate inter-subunit molecular adapters. Furthermore, it also seems that the three main networks build up, one around the other, like layers of an onion: with the innermost layer being the catalytic network, surrounded by the DNA/RNA interaction network, and then enclosed by the structural stability/scaffolding network. Interestingly, this same idea can be extended to include an outermost layer composed of the lineage-specific insertions. It could also be further generalized by considering the layering of additional subunits during the progression from the typically single subunit viral RNAP to the complex multi-subunit eukaryotic pol I/II/III RNAP assemblies (Figure 4.12). Furthermore, although additional work will be needed, it is reasonable to speculate that these same concepts might apply to the evolution of other large multi-subunit complexes.

Materials and Methods

Alignments for SCA

The alignments were created according to Chapter 3. Briefly, an automated sequence retrieval and processing system (BlaFA) was used to prepare accurate and complete lists of full-length protein. We then created initial alignments using the program PCMA [82] available at ftp://iole.swmed.edu/pub/PCMA/. The program PFAAT [75], available at http://pfaat.sourceforge.net/, was used to manually edit the alignments in order to remove species specific regions and fix misaligned positions.

In order to create the combined alignments for the inter-protein SCA we appended the corresponding proteins sequences from the same species. We then performed SCA normally. The results were then analyzed either as a single unit as in the case of β/β′ or just the inter-protein SCA Coupling Matrix as in the case of β/β′ with ω.

Statistical Coupling Analysis (SCA)

SCA was used to determine which alignment positions contained co-evolving residues. We performed the SCA calculations using the Ranganathan lab's SCA MATLAB functions (http://hhmi.swmed.edu/Labs/rr/). Prior to performing SCA we removed all residue positions which were 80% gapped or 100% conserved, followed by redundant sequence removal using a 95% sequence identity cutoff.

SCA yielded a pair-wise coupling matrix indicating the degree to which any two positions in a protein alignment were coupled or co-evolved. The SCA Coupling Matrixes were plotted with the x and y axes indicating the protein

residue positions ordered from the N to C-terminus. The SCA coupling value between any two positions was indicated by pixel colored from blue (no coupling [0]) to red (high coupling [≥2]). Previously, SCA has only been done on the intra-protein co-evolution of singular proteins. In order to extend SCA for the study of inter-protein coupling we preformed SCA on combined alignments in which we joined several proteins from the same species. As a result, the SCA Coupling Matrix from the combined alignments contained both the intra and the inter-protein couplings.

Hierarchal clustering and Independent Component Analysis

One of the primary methods used to understand SCA Coupling Matrixes is to look for residues that share common patterns of coupling or co-evolution. In order to determine these residue networks we employed two methods: (1) MATLAB hierarchal clustering and (2) Independent Component Analysis (ICA) using the FastICA MATLAB function available at http://www.cis.hut.fi/projects/ica/fastica/.

Hierarchal clustering has traditionally been the method of choice for identifying groups of similarly co-evolved residues within SCA Coupling Matrixes. However, hierarchal clustering is not without its limitations, since in addition to not being mathematically rigorous it also requires a great deal of subjective user input and experience to determine the cluster groups. In order to overcome these limitations, ICA has recently been used to analyze SCA results. In addition, to being mathematically rigorous ICA does not rely on as much user input to determine which residues belong to each group of co-evolving residues. Rather,

given the number of expected groups or independent components (IC) ICA assigns each residue position an ICA signal value indicating how strongly it belongs to a given group.

We compared the assignment of groups by hierarchal clustering and ICA in order to gain a deeper understanding of the SCA results. To do this we constructed plots of ICA signal value vs. residue position for each IC. We next re-order these plots using the residue ordering given by the hierarchal clustering. This approach allowed us to directly compare if the hierarchal cluster groups corresponded to the ICA groups. We found that if ICA was told to expect a number of ICs equal to or greater than the number of hierarchal cluster groups it would find an IC that directly corresponded to each hierarchal cluster group. Therefore, in general we have found that both the hierarchal clustering and ICA give the same results. However, only the hierarchal clustering allowed us to determine the amount by which any two groups were coupled or co-evolved to each other. As such, we only used ICA to help us identify mistakes with our assignment of hierarchal cluster groups. Thereby, we were able to use ICA to independently and rigorously validate the hierarchal cluster groups.

Due to the unprecedented large size of the resulting SCA Coupling Matrixes we developed SCACursor, a MATLAB graphical tool (scacursor.m) which allowed us to interactively examine the SCA Coupling Matrix plots. SCACursor allowed us to easily determine the residue numbering of each pixel in the clustered SCA result Matrix. In addition, SCACursor allowed us to draw around a cluster of interest in order to get a list of the enclosed residue positions.

We used PyMOL (available at http://pymol.org) to analyze the mapping of the residue to the bacterial *Tth* TEC RNAP structure (pdb code 2PPB) [9]. Since the *Tth* TEC RNAP structure is missing the transcription bubble we modeled it using the TEC cross-linking model [24].

Principle Component Analysis

In order to investigate the sequence differences that were responsible for the detection of a particular group, we first calculated the pairwise sequence identity between each sequence using only those residue positions contained in the group of interest. We then perform PCA by subtracting the mean of each position in order to center the values, followed by calculating eigenvectors using single value decomposition, resulting in the first two principle components (PC1 and PC2) which we used to create a scatter plot where each sequence was represented by a point. In addition, for Group 1, we correlated the PC1 scores with the percentage of operons containing termination signals, as predicted by TransTermHP (confidence value >75) available at http://transterm.cbcb.umd.edu/ [92].

Converting SCA Sequence Numbers to Other Species

We used reference alignments and a custom program (seq_pos_conversion.pl) to convert the residue numbering in each SCA cluster group to the residue numbering for other species, RNAP classes, the Tth TEC RNAP structure (pdb code 2PPB), and the *Sce* pol II RNAP structure (pdb code 1TWF).

Chapter 5 - Structure Based Modeling of Promoter Recognition in the RNA Polymerase Closed Complex

Introduction

Twist and Melt Model of DNA Melting

As mentioned previously, DNA melting is thought to occur as a result of the natural propensity of the -10 element sequence for thermal breathing, resulting in transiently flipped out bases that can be captured and stabilized by binding to aromatic residues on σ_2 [3, 23]. However, DNA melting may also involve the twist and melt model (Figure 5.1C), in which promoter recognition requires DNA twisting to properly orientate the -10 and -35 elements, thereby creating torsional strain that helps to unwind or melt the DNA [93-99]. Unfortunately, the transient nature of the RP_c makes direct observation difficult, allowing only indirect observations of RP_c kinetics from the rate of RP_o formation. Nevertheless, as seen Figure 5.1B, although promoters with 16 and 17 bp spacers (between the -10 and -35 elements) show the same RP_c binding stability (K_B), promoters with 17 bp spacers form RP_o at a rate (k_f) several fold higher *in vitro* [98] and are more active *in vivo* [100]. The effect of spacer length on transcription involves an isomerization step after RP_c formation, in which the nucleation of strand separation occurs [98]. In agreement, the majority of *Ec* σ^{70} promoters have 17 bp spacers (Figure 5.1A). Recently, the introduction of σ^{70} mutants deficient at melting [101] has facilitated the study of RP_c kinetics using DNA footprinting assays on promoters with 16 and 17 bp spacers, producing results consistent with the twist and melt model [99]. It should be noted that the twist and melt

model is not mutually exclusive with the thermal breathing and σ capture model described above.

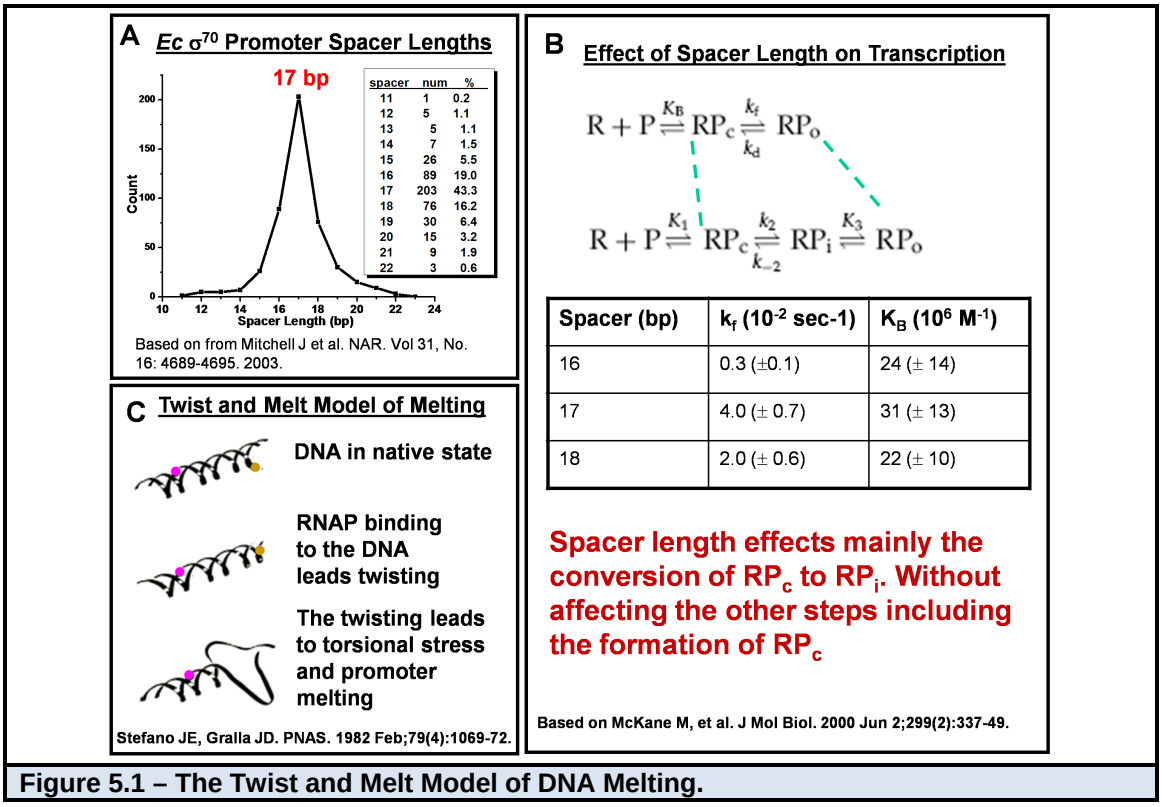
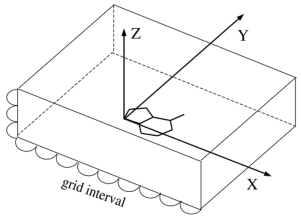


Figure 5.1 – The Twist and Melt Model of DNA Melting.

Prediction of Protein DNA Recognition

Due to the complexity and diversity of protein DNA recognition, most computational methods are either protein specific and/or rely on pre-existing libraries of recognized DNA sequences. However, as the number of DNA bound protein structures has increased there has been a renewed interest in developing more generalized structure based predictive algorithms. To this end several groups have extracted various parameters from experimentally known protein DNA structures in an attempt to fundamentally understand and evaluate the underlying interactions.



$$\Delta E^{ab}(s) = -RT \ln\left[\frac{f^{ab}(s)}{f(s)}\right]$$

$$f^{ab}(s) = \frac{1}{1+m_{ab}w} f(s) + \frac{m_{ab}w}{1+m_{ab}w} g^{ab}(s)$$

$$X = \sum_{s=1}^{n} \Delta E^{ab}(s)$$

$$Z = \frac{X - mean}{\sigma}$$

Based on Kono, H. and A. Sarai. Proteins, 1999. 35(1)

Briefly, a collection of diverse DNA protein structures is used to generate a database of parameters that encompass the structural preferences of DNA protein interactions. To accomplish this a 3-D grid space is generated around each DNA base using the N9 atom for A and G or the N1 for C and T and the point of origin. Then using the proteins Cα the frequency of protein residues in the grid space (voxel) around each base is determined (x,y,z grid of ±9 Å, ± 9 Å, ± 6 Å with a grid increment of 3 Å). Calculations for each amino acid type (a) and DNA base type (b) are done using a grid space of s voxels (1 .. n) around each DNA base. With m_{ab} = number of times a particular amino acid and base combination is observed, w is the weight of each observation (w = 1/20), $f(s)$ is the relative frequency of any amino acid with any base type in voxel s, $g^{ab}(s)$ is the relative frequency of that particular amino acid and base type base interaction in voxel s, RT is the gas constant and absolute temperature (T = 293 K, RT = 0.582 kcal/mol), X is the energy value for protein DNA sequence interaction, mean is the mean X value for random DNA sequences, σ is the standard deviation of X value for random DNA sequences, Z is the Z-score which indicated the statistical fitness for a particular DNA-protein interaction usually between -4 (favorable) and -4 (unfavorible).

Figure 5.2 – Computational Evaluation of DNA-Protein Interactions.

Using one such approach (Figure 5.2) it is possible to estimate the energy (X) of a protein DNA structure by applying knowledge of base and residue pairing frequencies from a catalog of non-redundant protein/DNA structures [102-104]. By threading random DNA sequences it is possible to calculate Z scores that indicate the statistical fitness of each DNA sequence in the evaluated protein-DNA interaction.

Results and Discussion

Progress Timeline

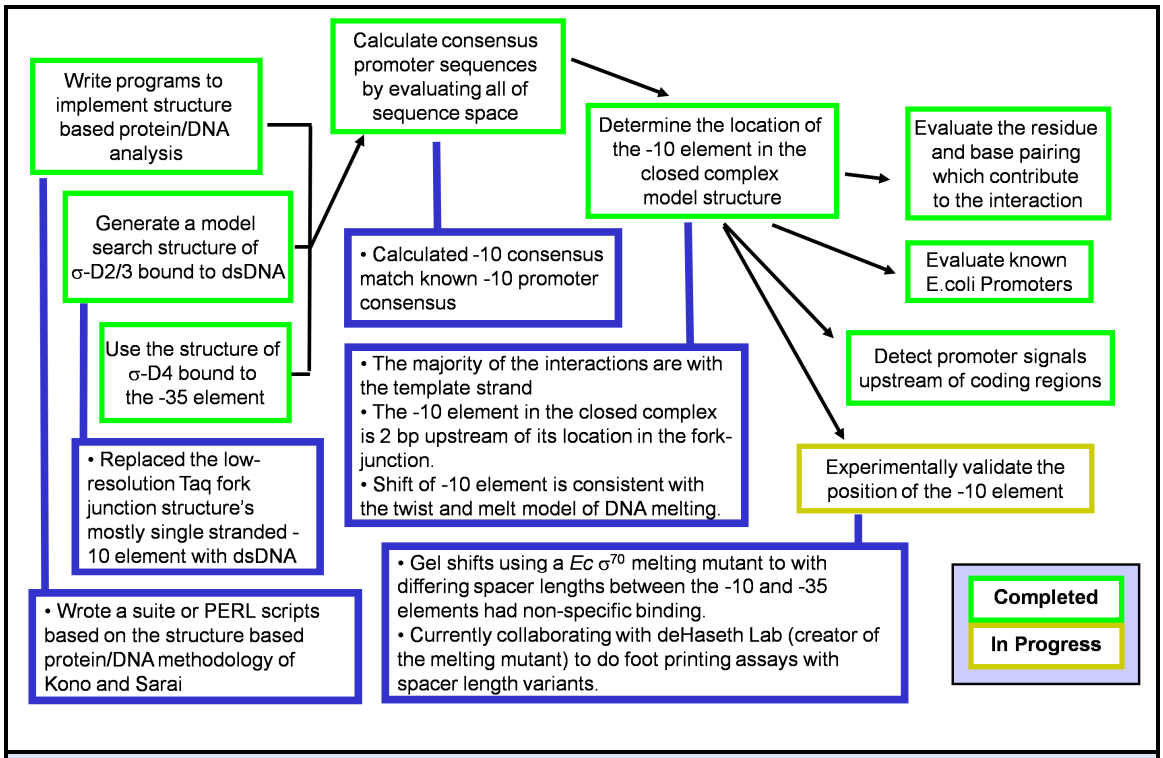


Figure 5.3 – Structure Based Modeling of the Closed Complex Progress Timeline.
Above is a schematic of this projects progress. Green boxes are completed parts. Yellow boxes are parts in progress or uncompleted. Black boxes are parts not started. Blue boxes important details for each step.

Search Models

In order to computationally evaluate the recognition of promoter DNA by the RP_c, we used a structure based predictive strategy that only relies on the position of Cα's relative to the DNA [102, 103]. Typically, using this approach an energy value (X) is calculated for a target protein-DNA structure, followed by normalization using energy values calculated using the same structure, but with the native DNA sequence replaced by random DNA sequences. However, the evaluation of the RP_c presents an immediate problem, since the structure based prediction method requires pre-existing structural information. Since, RP_c only

exists transiently due to the rapid and irreversible conversion to RP_o, crystallographic attempts at determining the structure of the RP_c have been unsuccessful. Nonetheless, two structures do offer some limited structural insight into RP_c recognition by holoenzyme containing housekeeping Group 1 σ factors. First the high-resolution structure of the Thermus aquaticus (*Taq*) σ_4 in complex with a -35 promoter element (*Taq* σ_4/-35) demonstrates that part of the RP_c protein DNA interactions occur through σ_4 binding to the major groove of the -35 element [21]. Second, the 6.5 Å low resolution x-ray structure of the *Taq* fork-junction complex (*Taq* RF), which is thought to mimic the RP_o, has been crystallized using fork-junction DNA containing a mostly single stranded -10 promoter element [5]. By extending a double stranded DNA helix through the mostly single stranded RF -10 element it is possible to propose a model of the closed complex (*Taq* model-RP_c). However, it is not possible to know the definitive location or phase of the -10 promoter element along the modeled double-stranded DNA.

Consensus Promoter Prediction

Using the part of the *Taq* model-RP_c that contained $\sigma_{2.2\text{-}3.0}$ bound to DNA (*Taq* model-$RP_{\text{-}10c}$), we replaced the DNA in the original structure systemically covering the entire possible sequence space in order to calculate a consensus -10 binding sequence from the best scoring DNA sequences (Figure 5.4). Remarkably, our calculated consensus (Figure 5.4B), which did not rely on previously known sequence information, contained the known -10 element and extended -10 consensus sequences, allowing us to place the -10 and extended -

10 promoters in *Taq* model-RP$_c$ (Figure 5.4A). Relative to *Taq* RF, the -10 element in *Taq* model-RP$_c$ has been shifted upstream by 2 bp.

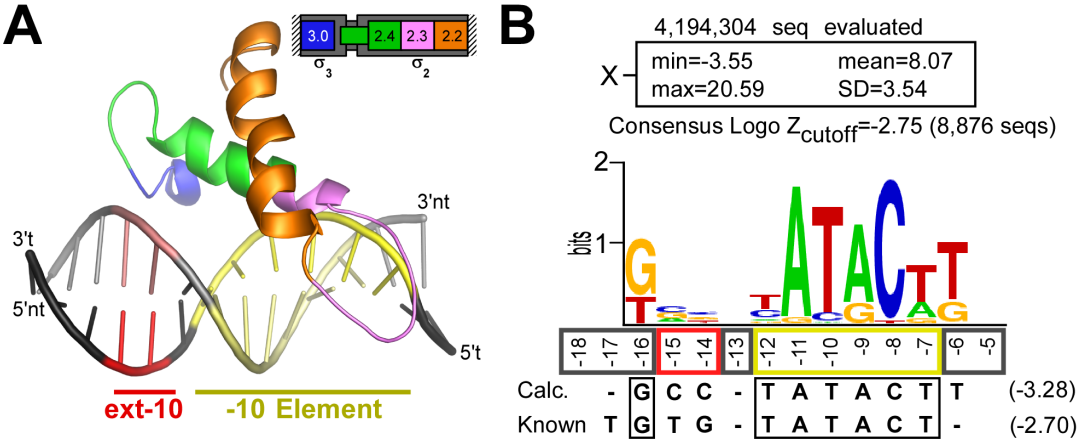


Figure 5.4 – Computational Prediction of a Consensus -10 Element Promoter.
Structure based computational modeling was used to *de novo* predicate consensus sequences for σ protein recognition of the -10 element. The σ subunit is shown as cartoons with $\sigma_{2.2}$ (orange), $\sigma_{2.3}$ (violet), $\sigma_{2.4}$ (green), $\sigma_{3.0}$ (blue). The DNA is shown as a cartoon ladder with the template strand (dark color) and non-template strand (light color) colored to highlight the -10 element (yellow), extended -10 element (red), and non-promoter DNA (gray). **(A)** *Taq* model-RP$_{-10c}$ search model and **(B)** predicted consensus sequence logo. The predicted and known consensus sequence are shown under the sequence logo along with the consensus sequence Z score in parentheses. Details from the computational prediction are displayed about the sequence logo.

In order to evaluate the interactions between σ_4 and the -35 element DNA, we created *Taq* model-RP$_{-35c}$ model using the previously solved *Taq* σ_4-DNA structure (Figure 5.5A). Figure 5.5B shows the sequence logo of the best scoring sequences for the *Taq* model-RP$_{-35c}$ structure. However, the calculated consensus derived using the *Taq* model-RP$_{-35c}$ search structure did not match the known consensus sequence. This discrepancy could be due to a failure of the computational technique. In fact, previous structures have shown that -35

element recognition relies on DNA bending [21] and/or sequence specific DNA geometry [105], two things our computational technique ignores.

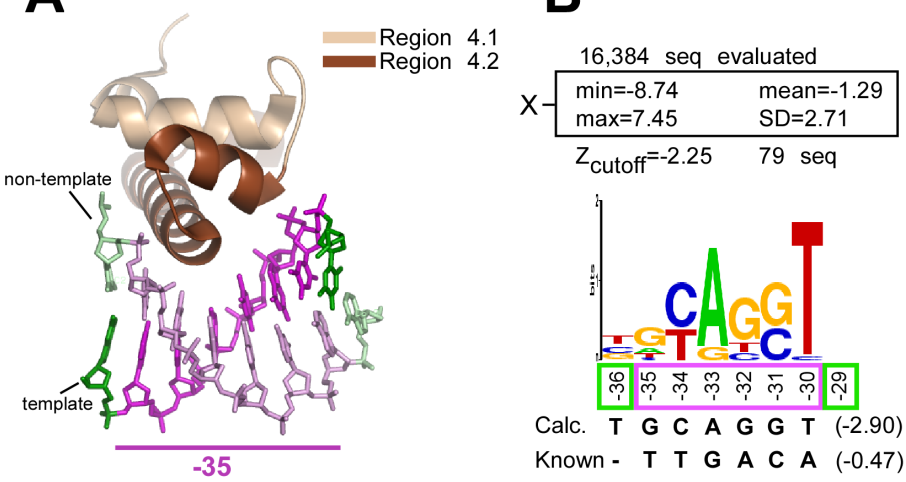


Figure 5.5 – Computational Prediction of Consensus -35 Element Promoters.
Structure based computational modeling was used to *de novo* predicate consensus sequences for σ protein recognition of the -10 element. The σ subunit is shown as cartoons with $\sigma_{4.1}$ (pale yellow) and $\sigma_{4.2}$ (tan). The DNA is shown as a cartoon ladder with the template strand (dark color) and non-template strand (light color) colored to highlight the -35 element (magenta) and non-promoter DNA (gray). **(A)** *Taq* model-RP_{-35c} search model and **(B)** predicted consensus sequence logo. The predicted and known consensus sequence are shown under the sequence logo along with the consensus sequence Z score in parentheses. Details from the computational prediction are displayed about the sequence logo.

Analysis of Group 1 σ -10 Element Recognition

When the combined energy score for each residue was mapped to the *Taq* model-RP_{-10c} structure it revealed pockets of favorable energy values around the regions of protein DNA interaction (Figure 5.6). Furthermore, the majority of the σ -10 element interactions were with the template strand, while the extended -10 element interactions were with the non-template strand (Figure 5.6B) [throughout this manuscript, DNA bases will be numbered as in Figure 5.6B, where negative numbers denote base pairs upstream of the transcription start site. Unprimed

numbers denote the nontemplate (top) DNA strand, while primes denote the template (bottom) strand].

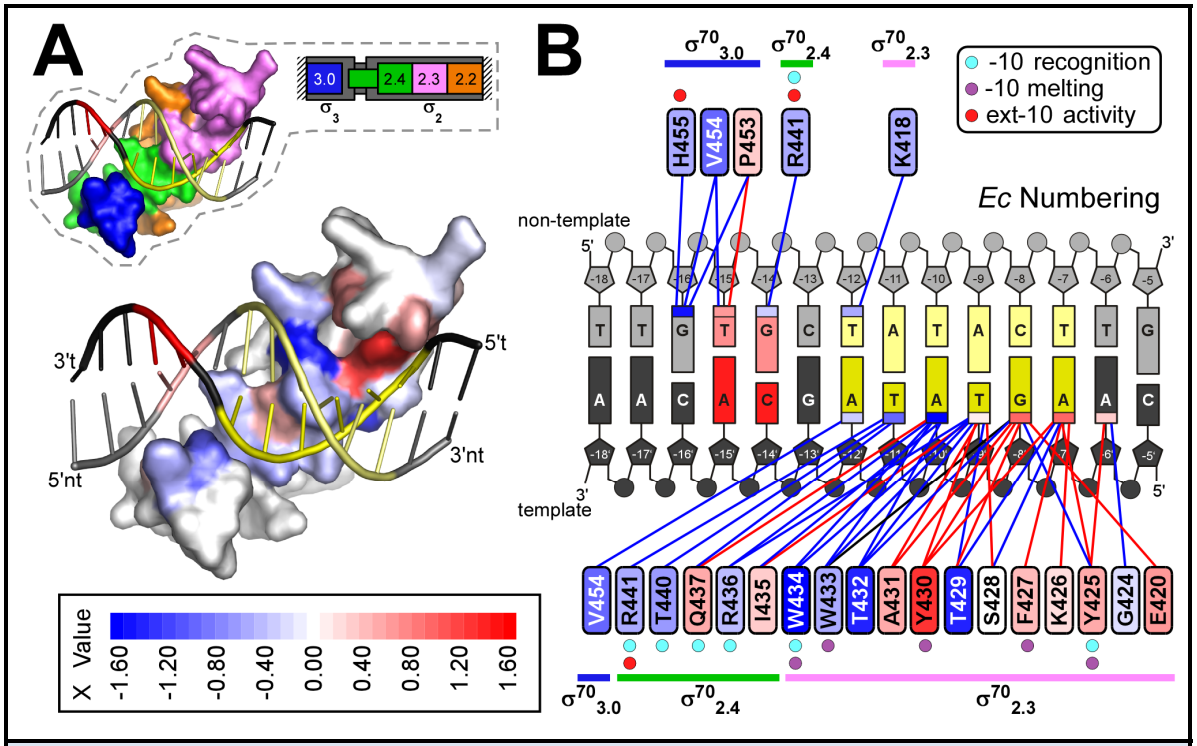


Figure 5.6 – Structural Mapping and Schematic Identification of Energy Interactions.

(A) Shows a surface rendering of *Taq* model-RP$_{-10c}$, in which σ residues were colored from blue (favorable) to red (unfavorable) depending on their individual contribution to the energy score for the DNA protein interactions. The DNA is shown as a cartoon ladder with the template strand (dark color) and non-template strand (light color) colored to highlight the -10 element (yellow), extended -10 element (red), and non-promoter DNA (gray). The gray dashed regions show the same surface view, but with the σ subunit structural domains $\sigma_{2.2}$ (orange), $\sigma_{2.3}$ (violet), $\sigma_{2.4}$ (green), $\sigma_{3.0}$ (blue).

(B) Shows a schematic of the energy interactions between σ (with *Ec* residue numbering) and the DNA (positions -18 to -5) in *Taq* model-RP$_{-10c}$. As in (A) σ residues have been colored from blue to red depending on their individual contribution to the energy score with colored horizontal lines indicating the σ structural domains. The DNA was colored as in (A) with small blue to red colored rectangular region at edge of each DNA base indicating their individual contribution to the energy score. Small circles next to each residue indicate known mutations which affect -10 element recognition (cyan), -10 element melting (purple), and ext-10 element activity (red).

The three upstream bases at positions -12′, -11′, and -10′ each have better combined X values than the three downstream bases at positions -9′, -8′, -7′. This difference is most likely the result of a small steric clash between σ and the DNA at the downstream bases. However, given that our models effective resolution is limited as a result of only using the position of the Cα within a 3 Å spaced grid, slight movements that might better orient σ with respect to the three downstream bases would be accommodated, or the downstream DNA might have a slightly different geometry. In truth, the effective resolution limits of the modeling were actually advantageous since they allowed the use of our search model (*Taq* model-RP_{-10c}) which was generated from the 6.5 Å low resolution *Taq* RF crystal structure.

The location of σ^{70} mutations known to effect -10 recognition [106, 107], -10 melting [101, 107, 108], ext-10 activity [109], and fork-junction binding are indicated in Figure 5.6B and Figure 5.7. As can be seen, the known σ^{70} mutations span the length of the ext-10 and -10 elements. In particular, most of the -10 recognition mutants are located on $\sigma^{70}_{2.4}$ and have favorable energy interactions with the two upstream template bases at positions -11′ and -10′. Our results extend this to the upstream edge of the -10 element, with favorable energy interactions between *Ec* V454 with the -12′ position and *Ec* K418 with the -12 position. In contrast, most of the -10 melting mutants are located in $\sigma^{70}_{2.3}$ and have unfavorable energy interactions with the four downstream template bases at position -10′, -9′, -8′, and -7′. Interestingly, if one assumed that the location of the -10 element was the same as in the *Taq* RF structure, most of the mutations

involved in -10 element recognition would be too far upstream of the -10 element. However, the 2 bp upstream shift revealed in *Taq* model-RP$_{-10c}$ results in a good overall alignment of the known σ^{70} mutations with the -10 element (Figure 5.7). Therefore, we believe that *Taq* model-RP$_{-10c}$ represents the actual state of σ^{70} -10 element recognition.

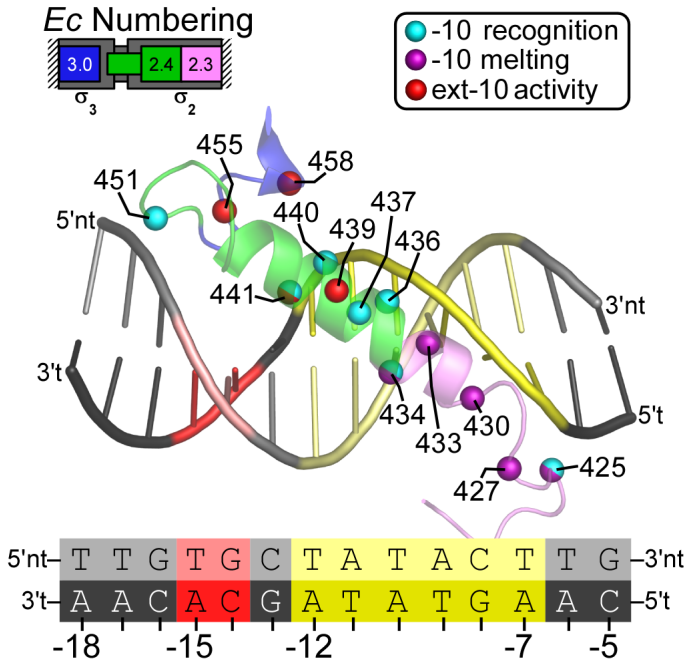


Figure 5.7 – Structural Mapping of Known Mutations.
The DNA is shown as a cartoon ladder with the template strand (dark color) and non-template strand (light color) colored to highlight the -10 element (yellow), extended -10 element (red), and non-promoter DNA (gray). The σ subunit is shown as a cartoon with $\sigma_{2.2}$ (removed for clarity), $\sigma_{2.3}$ (violet), $\sigma_{2.4}$ (green), $\sigma_{3.0}$ (blue). Small spheres indicate σ^{70} mutations (*Ec* residue numbering) which affect -10 element recognition (cyan), -10 element melting (purple), and ext-10 element activity (red). The DNA sequence is shown below the structure.

Studies of promoter recognition have implicated *Ec* σ residues R436, R441, R451 and W434, Y425 (*Taq* R259, R264, R274, W257, F248) as key residues in the recognition of the -10 promoter, independent of promoter melting [107]. In close agreement, our calculations showed that residues *Taq* R246, R259, and

W257 have very favorable interactions with the DNA (low X values). In addition, our results showed interactions with *Ec* residues W433 and Y430 (*Taq* W256 and Y253) both of which are experimentally implicated in -10 element DNA melting [101]. Furthermore, in agreement with experimental evidence [109], our calculations showed favorable interactions for *Ec* R441 and H455 (*Taq* R264 and H278) near the extended -10 element.

Recent experiments by Schroeder *et al.*, have called into question the long held assumption that σ^{70} residues *Ec* Y430 and *Ec* W433 interact with the highly conserved -11 position, since both residues affect RP_o formation even when the -11 position is not conserved [110]. Our *Taq* model-RP_{-10c} results support this notion since we find energy interactions downstream of the -11 position for *Ec* Y430 (-9′, -8′, and -7′ positions) and *Ec* W433 (-10, -9, and -8 positions). As Schroeder *et al.* conclude, another residue might interact with the conserved -11 position, which is thought to be the start site for strand separation. Based on the *Taq* model-RP_{-10c} model three residues (*Ec* R441, T440, and Q473) all of which have been experimentally implicated in -10 recognition show energy interactions with the -11′ position. Furthermore, the *Taq* model-RP_{-10c} results indicated that *Ec* W434, which has energy interactions with the template strand at positions -10′ and -9′, might be capable of capturing a flipped out non-template strand nucleotide base at the -11 position. Since, the capture of the non-template -11 position might be the initiating event in strand separation it makes sense that the interactions we see in the *Taq* model-RP_{-10c} might have a role.

Promoter -10 Element DNA Melting

The 2 bp upstream shift of the -10 element location in the RF and *Ec* model-RP_{-10c} models suggests that such a movement might play an important role in the transition from RP_c to RP_o. In fact, this movement would be in agreement with the the twist and melt model of DNA melting, in which RP_c formation requires the -10 and -35 elements to re-orientate and twist, thus straining the DNA and facilitating melting. In the twist and melt model the RP_c is thought to optimally recognize -10 and -35 elements separated by a distance equivalent to a 16 bp spacer. Therefore, since the majority of promoters in *Ec* have 17 bp spacers most promoters would be twisted upon formation of RP_c facilitating the formation of RP_o. Furthermore, 16 bp promoters would be deficient in transcription despite RP_c formation since they lack the additional torsional energy to melt the DNA. However, since the lag assays traditionally used to study RPc are not suitable for studying promoters with slow rates of RP_o formation, the errors associated with the slow 16 bp spacers are high and experiments are not possible when studying 15 bp spacers.

As a consequence of the DNA twist the total distance between the two DNA binding elements would decrease. Therefore, it was important to determine the spacer length that would result from our modeling. Based on the RF structure which contains a 17 bp spacer (Figure 5.8B), our *Taq* model-RP_{-10c} indicates that a complete model of RP_c (1^{st}-model-RP_c) would twist a 17 bp promoter by 2 bp creating a distance between the -10 and -35 elements equivalent to a spacer length of 15 bp (Figure 5.8A). However, when the high resolution structure of the

Taq σ_4/-35 element was aligned onto the low resolution RF structure using σ_4 (2^{nd}-model-RP_c) the curvature of the -35 element causes a 1 bp loss in the RF spacer, thus creating an 16 bp equivalent spacer length (Figure 5.8C). Either way, it is clear that the model-RP_c would compress the equivalent distance between the -10 and -35 elements such that a 17 bp spacer would be twisted and strained.

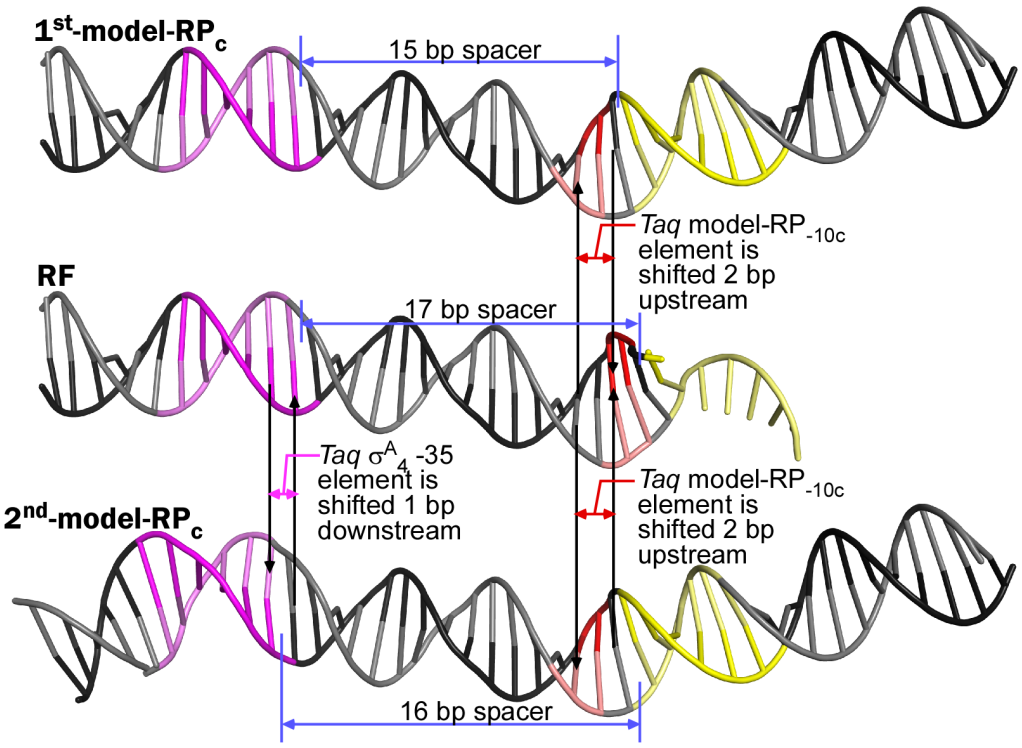


Figure 5.8 – Structural Modeling of RP_c Spacer Length.
The DNA is shown as a cartoon ladder with the template strand (dark color) and non-template strand (light color) colored to highlight the -10 element (yellow), extended -10 element (red), and non-promoter DNA (gray). 1^{st}-model-RP_c was constructed using *Taq* model-RP_{-10c} and the downstream DNA from the RF structure, resulting in a 15 bp spacer length due a 2 bp upstream shift of the -10 element and while the -35 element from the low resolution *Taq* RF structure. The 2^{nd}-model-RP_c was constructed using *Taq* model-RP_{-10c}, spacer DNA from the *Taq* RF structure, and the -35 element from the high resolution *Taq* σ^A/-35 element structure, resulting in a 16 bp spacer length due to a 2 bp upstream shift of the -10 element and a 1 bp downstream shift of the -35 element.

Therefore, our results are in good agreement with the experimental twist and melt model which indicates that RP_c recognizes the -10 and -35 element when they are separated by a distance equivalent to a 16 bp spacer. It is important to note that unlike the 16 bp spacer length, 15 bp spacers have not been adequately tested in the twist and melt model since the lag assays were hindered by its complete lack of RP_o complex formation. Therefore, an equivalent distance of 15 bp or a distance in between the 15 and 16 bp spacers could represent the actual limits of RP_c -10 and -35 element compression. We propose that *Taq* model-RP_{-10c} is representative of the -10 element recognition state after the DNA has been twisted and prior to strand-separation (Figure 5.9).

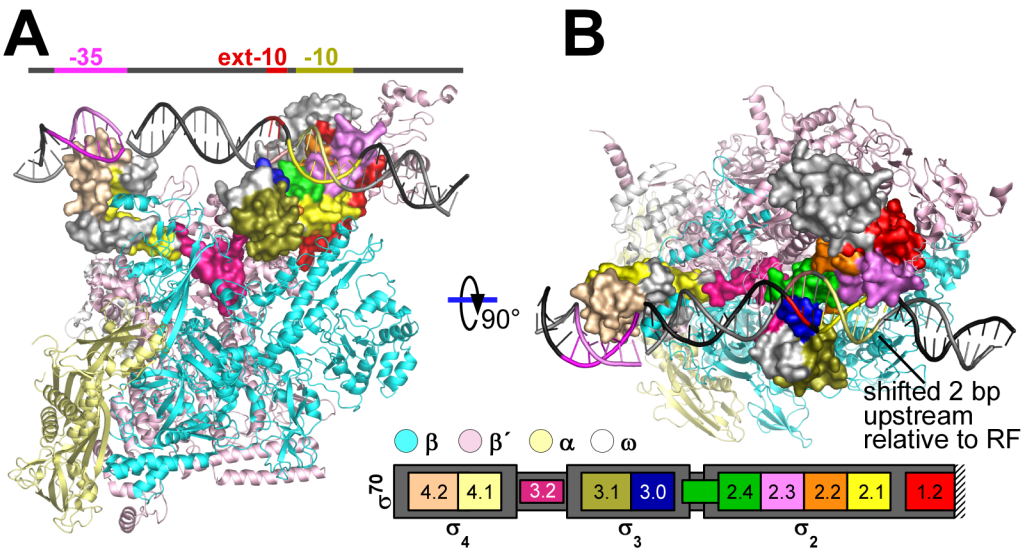


Figure 5.9 – Structural Model of the RP_c.

(A) The DNA is shown as a cartoon ladder with the template strand (dark color) and non-template strand (light color) colored to highlight the -10 element (yellow), extended -10 element (red), and non-promoter DNA (gray). The RNAP core is shown a cartoon with β (cyan), β′ (light pink), α (pale yellow), and ω (white). The σ subunit is shown a molecular surface with $\sigma_{1.2}$ (red), $\sigma_{2.1}$ (yellow), $\sigma_{2.2}$ (orange), $\sigma_{2.3}$ (violet), $\sigma_{2.4}$ (green), $\sigma_{3.0}$ (blue), $\sigma_{3.1}$ (olive), $\sigma_{3.2}$ (dark pink), $\sigma_{4.1}$ (pale yellow), $\sigma_{4.2}$ (tan). **(B)** Same as (A), but rotated 90°.

Ec RP_c Modeling

We generated an *Ec* model for -10 (*Ec* model-RP_{-10c}) and -35 (*Ec* model-RP_{-35c}) element recognition. Since the algorithm we employed only relied on the Cα position relative to the DNA, we simply replaced the amino acid identities in the model structure with those from *Ec*. This model was used to generate calculated consensus sequences for the -35 and -10/ext-10 element (Figure 5.10A). The calculated *Ec* consensus sequences were similar to those calculated for *Taq*. The calculated *Ec* -35 consensus showed little agreement with a consensus determined using 453 known *Ec* σ^{70} -35 element sequences (Figure 5.10B). However, the calculated *Ec* -10/ext-10 consensus (Figure 5.10A) showed close agreement with a consensus determined using 449 known *Ec* σ^{70} -10 element sequences (Figure 5.10B). In fact at several positions the calculated consensus results match the known experimental ones not only for most abundant base, but also agree on the second most abundant base at the -8 position.

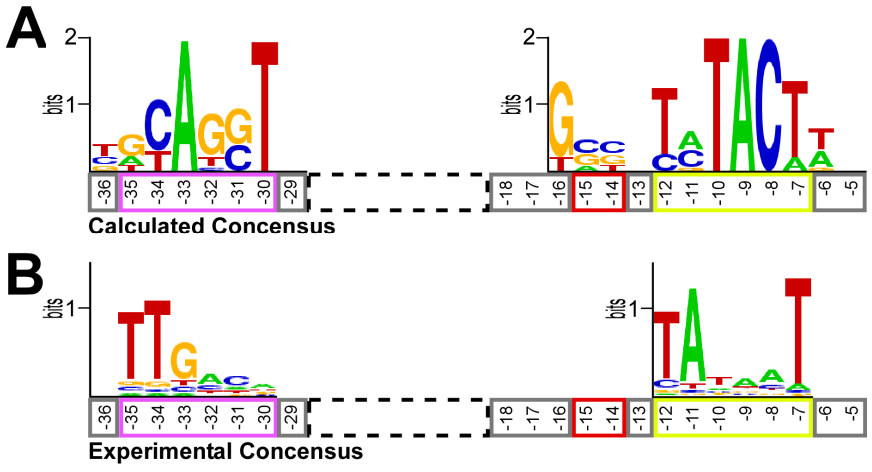


Figure 5.10 – Computational Prediction of *Ec* Consensus Promoters.

(A) *De novo Ec* consensus sequence logo predicated using structure based modeling.

(B) *Ec* consensus sequence logo created from known *Ec* promoters. The *de novo* calculated consensus closely matches the consensus created from known promoters.

Evaluation of Known *Ec* σ^{70} Promoters

In order to understand the diversity of *Ec* σ^{70} promoter sequences we used our models to evaluate the fitness of known Ec promoters (Figure 5.11). A total of 435 known *Ec* σ^{70} -35 elements were evaluated (Figure 5.11A, C). The best Z score was -2.07, the worst Z score was 2.24, the average Z score was -0.34, and the standard deviation of the Z scores was 0.71. Figure 5.11A shows a plot of the frequency distribution of Z scores for the entirety of sequence space and for the -35 promoter sequences. The Z scores from the -35 promoter sequence closely matches that of the entirety of sequence space at low Z values (favorable interaction), indicting very little enrichment of highly favorable interactions. However, there is a lack of promoter sequences with very high Z scores (unfavorable interactions) indicating that overall the -35 promoter sequences score more favorable then a random sequence would.

A total of 449 known *Ec* σ^{70} -10 elements (some with ext-10 elements) were evaluated (Figure 5.11B, C). The best Z score was -2.89, the worst Z score was 2.15, the average Z score was -0.94, and the standard deviation of the Z scores was 0.88. Figure 5.11B shows a plot of the frequency distribution of Z scores for the entirety of sequence space and for the -10/ext-10 promoter sequences. Unlike the -35 sequences, the Z scores for the -10 promoter sequence were significantly shifted to the left (more favorable) by almost one standard deviation (1 Z = 1 SD) with the -10/ext-10 promoter sequences scoring the best. This indicates that the -10 promoter elements are enriched for highly favorable interactions.

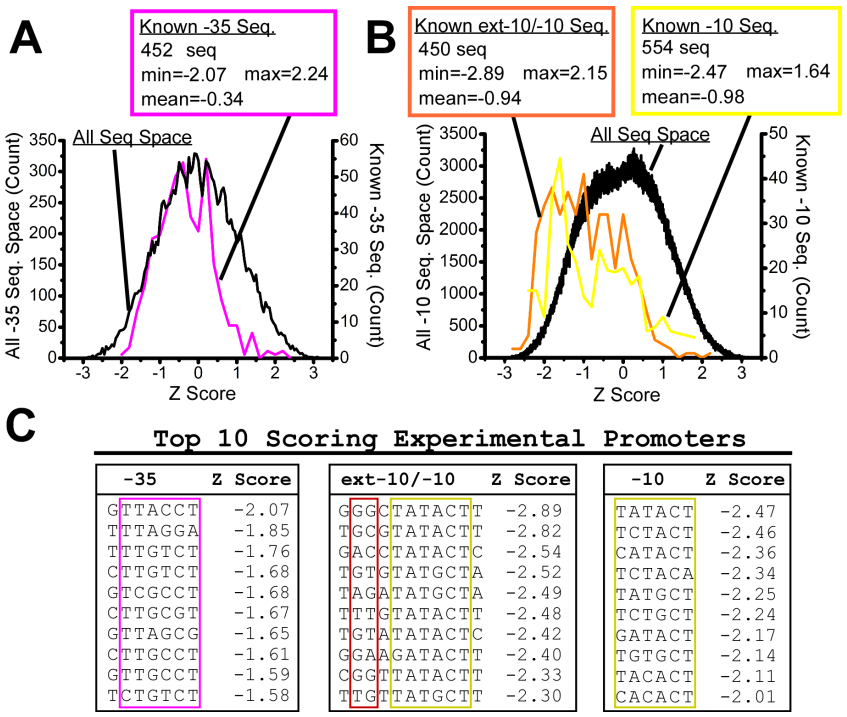


Figure 5.11 – Computational Evaluation of Known *Ec* Promoters.

(A) The *Ec* model-RP$_{-35c}$ model was used to evaluate known -35 element sequences. Plots of Z score frequency with known -35 element sequences (purple) and all possible sequence combinations (black) showed that the known -35 element sequences score better than random sequences. Details for the known -35 promoter Z score are shown the purple box.

(B) The *Ec* model-RP$_{-10c}$ model was used to evaluate known -10 element sequences. Plots of Z score frequency with known -10 element sequences with (orange) and without (yellow) extended -10 elements and all possible sequence combinations (black). The known ext-10 and -10 element promoter sequences clearly score better than random sequences. Details for the known promoter Z score are shown the orange and yellow boxes.

(C) Table of the top ten scoring known -35, ext-10/-10 and -10 elements.

Promoter Prediction

As a true test of our system, we sought to identify possible promoters from stretches of genomic DNA upstream of transcription start sites. To do this we used the *Ec* model-RP$_{-10c}$ and *Ec* model-RP$_{-35c}$ model structures to scan the 250 bases upstream of the start site. We then summed the Z score for the two interactions along with an additional term to weight the length between the two

promoter elements. These combined scores were then used to locate promoter signals based on a cut-off of -3.5. When we plotted our promoter signals along the promoter positions we were able to detect the known location for several genes (Figure 5.12).

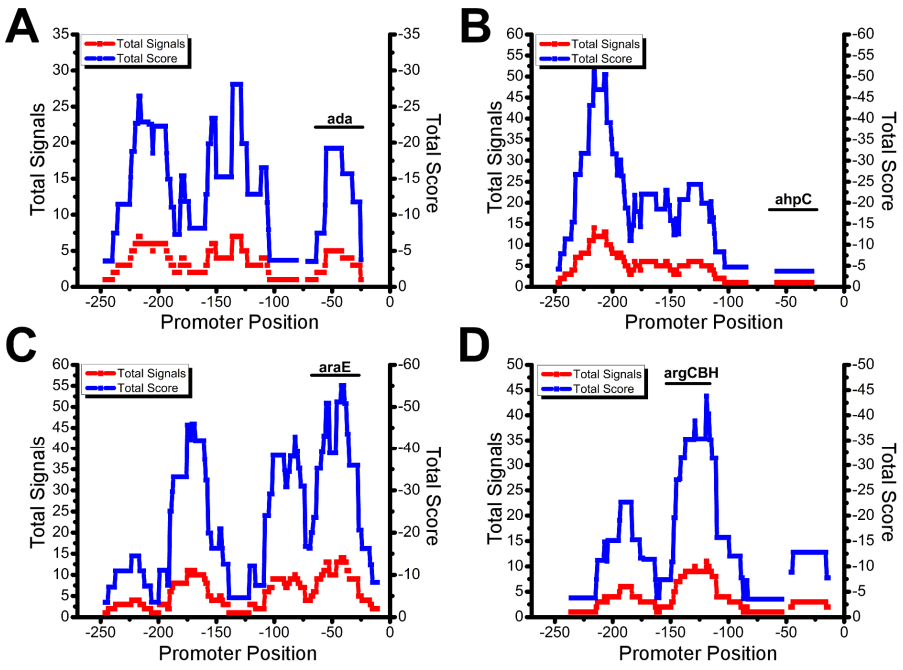


Figure 5.12 – *Ec* Promoter Detection.
Frequency plots of promoter total promoter detection signals and score along the 250 bp upstream of the promoter start site for: **(A)** ada, **(B)** ahpC, **(C)** araE, and **(D)** argCBH. The black bars indicate the actual known promoter location.

Our results are similar with previous results calculated using a position weight-matrix approach [111], including the identification of multiple regions with high signal that might help localize holoenyme to the promoter.

Conclusions

By computationally trapping the closed promoter complex we were able to: (1) determine a fairly specific model of the closed complex including the location of the -10 element, (2) make testable predictions in regards to the twist and melt model of DNA melting that should allow our model to be experimentally validated, and (3) establish a novel method of bacterial promoter identification. Future studies might focus on experimentally validating our model, by expanding the previous σ^{70} melting mutant studies with the determination of the RP_c affinity to promoters with 15 bp spacers. Additionally, we might improve and expand our methodology to model other σ factor groups, including those with less characterized regulons. We believe that this could be possible using customized search models along with algorithms that incorporate indirect protein recognition using *ab initio* DNA structure modeling and water mediated interactions [104].

Materials and Methods

Structure Based Modeling of Protein DNA Structures

The evaluation of protein DNA structures was based on the methods of Kono and Sarai [102, 103]. Briefly, a collection of 62 known DNA protein structures was used to generate a database of parameters that encompass the structural preferences of DNA protein interactions. To accomplish this a 3-D grid space was generated around each DNA base using the N9 atom for A and G or the N1 for C and T and the point of origin. Then using the proteins Cα the frequency of protein residues in the grid space around each base was determined. For our calculations we evaluated a total box size around each base of (x,y,z) = (+/- 9 Å, +/- 9 Å, +/- 6 Å) with a grid increment of 3 Å [103]. This database was then used to evaluate other protein DNA structures, by determining the agreement of parameters from the current structure with the parameters from the other protein DNA structures stored in the database. This evaluation results in an X value calculated using the energy function eqn. The more negative or positive the X value the closer the search structure agrees or disagrees with the database parameters. The X values were used to determine Z scores according the equation (X-m)/SD, where X is the X value, m is the mean and SD the standard deviation of the random/all DNA combinations X values. The Z score is a statistically relevant measurement of the fitness of a given interaction. In addition, the Z score is not dependent on the particulars of the protein DNA interaction, allowing direct comparison between difference protein DNA interactions. Z

scores generally fall between -4 to +4 with a Gaussian distribution in which 1 unit equals 1 standard deviation off of the mean X value.

We developed a suite of PERL scripts (dPro) in order to facilitate the calculations (Figure 5.13). The script get_pdb.pl downloaded the pdb files for the non-redundant set of DNA/protein structures. The script pdex_zem.pl was used to calculate the statistical parameters needed to calculate X. The script gdna.pl was used to generate DNA input files with specified overhangs either using input non-template strand sequences, input non-template strand genomic DNA stretches, random DNA, or from all possible sequence combinations. The script X.pl calculated the energy function for each protein DNA structure using a particular target sequence or from a list of sequences generated by gdna.pl. The script Z.pl normalized the energy calculations for the desired target sequence.

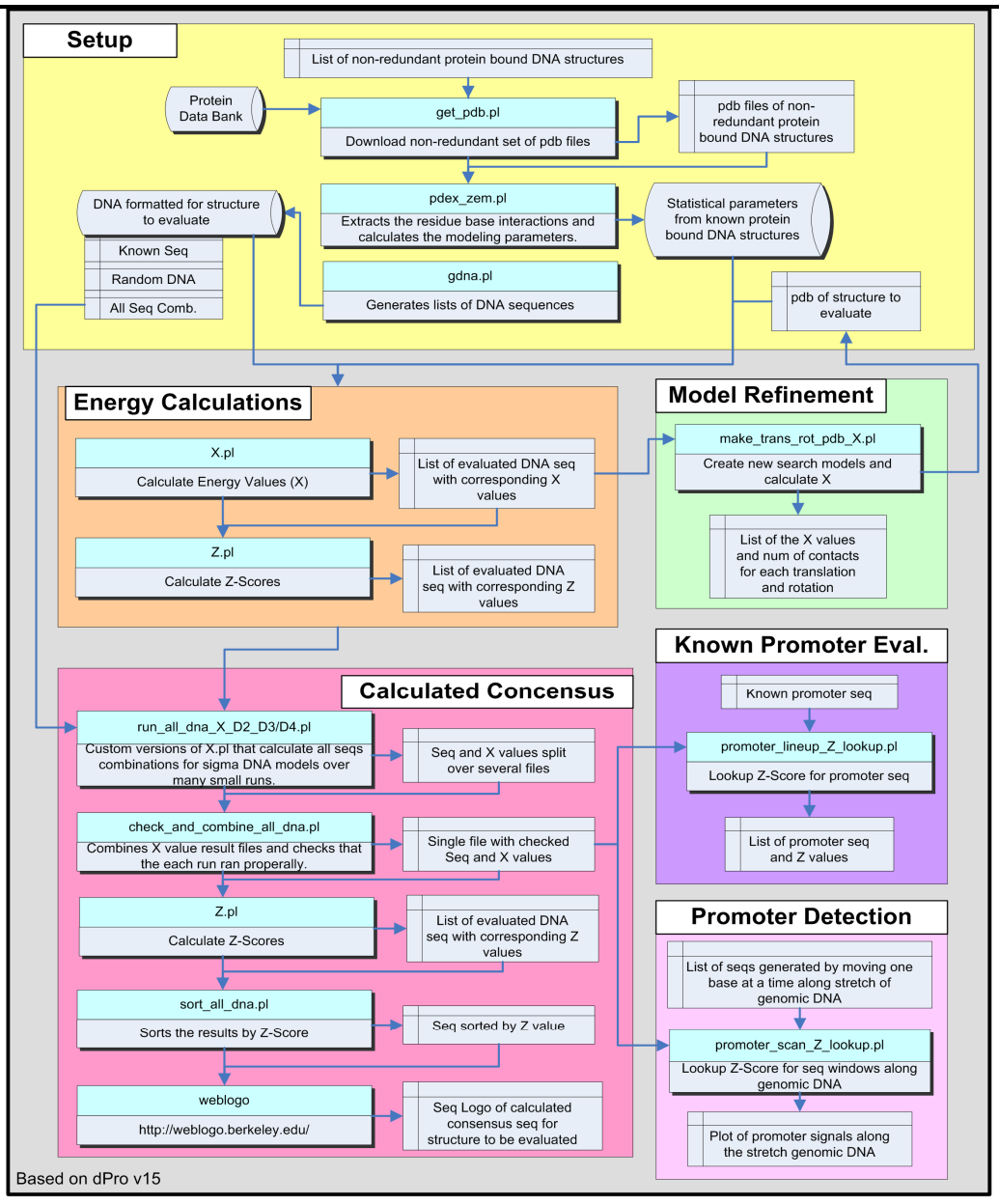


Figure 5.13 – Flowchart of dPro programs.

Even though we used employed a previously developed algorithm for the evaluation of protein/DNA structures. We programmed our own implementation allowing us to flexibility to extend the technique for consensus promoter prediction, evaluation of known promoter, the detection of promoters in a genomic context. We also briefly explored the concept of model refinement in which the orientation of the DNA and protein in the search model was altered in order to discover better models. However, this was never full pursued due to the computation requirements of modeling a large library of random DNA sequences for evaluated orientation. However, model refinement along with *ab initio* DNA structure prediction might someday be more feasible.

Structure Based Modeling of Known Promoters

In order to characterize promoter specificity, we independently evaluated the interactions between $\sigma_{2\text{-}3}$ and the -10 and extended -10 (model-$RP_{\text{-}10c}$) and σ_4 and the -35 (σ_4-DNA). The model-$RP_{\text{-}10c}$ structure was derived from the RF structure (pdb code: 1L9Z) by modeling a double stranded DNA element in place of the fork junction. The σ_4-DNA interaction was modeled using the previously known structure (pdb code: 1KU7).

The model-$RP_{\text{-}10c}$ structure was derived from the RF structure (pdb code: 1L9Z) by modeling a double stranded DNA element in place of the fork junction. The σ_4-DNA interaction was modeled using the previously known structure (pdb code: 1KU7).

Initial Z scores for each DNA protein interaction were calculated using 50,000 random DNA. We later evaluated all possible DNA sequence combinations in order to more faithfully calculate consensus promoters. However, doing so drastically increased the computational load. For example a double stranded 14 base pair DNA like that in the *Taq* model-$RP_{\text{-}10c}$ required evaluating 4,194,304 (2^14) sequences. In order to speed up the calculations, X.pl was used to determine which bases and residues were not contributing to the calculations, allowing their removal from the search structures. Furthermore, modified versions of X.pl called run_all_dna_X_D2_D3.pl and run_all_dna_X_D4.pl performed the calculation over all possible sequence combinations by breaking the number of sequence into smaller manageable sequence subsets. An individual Z score was calculated for each sequence by using the X values from the other sequences.

The sequences with the best scores were then used to generate calculated consensus sequences which were displayed as sequence logos [37].

Ec search models were generated from the *Taq* structures by simply replacing the protein residues with those from *Taq*. No further changes were necessary since the algorithm only relies of the position of the proteins Cα relative to the DNA base. The script promoter_lineup_Z_lookup.pl was used to calculate the Z scores of experimentally known *Ec* σ^{70} promoters.

Structure Based Modeling for *Ec* Promoter Detection

In order to detect *Ec* promoters in an actual genomic context we analyzed the 250 bases upstream of transcriptional start sites. Z scores for the *Ec* model-RP$_{10c}$ and σ_4-DNA were calculated independently at each location along the DNA by moving one base at a time. Given that the distance between the -10 and -35 is between 12 and 22 we separately added the -10 and -35 scores for each spacer length. Each of these summed scores was weighted according to the log of the frequency of that particular separation. Combined scores less than -3.5 were considered promoter signals. These signals were then used to calculate the frequency of promoter signals along stretches of genomic DNA.

The script promoter_scan_Z_lookup.pl scanned the entire *Ec* genome with the help of the Z scores previously calculated when covering all possible DNA sequence combinations when determining the calculated consensus. Rather then calculating the X value and Z score for each sequence we instead used the previous results to efficiently look up the sequences Z score as we slide one base at a time across the entire genome.

Chapter 6 - Structure of a Group IV Sigma Holoenzyme

Results and Discussion

Progress Timeline

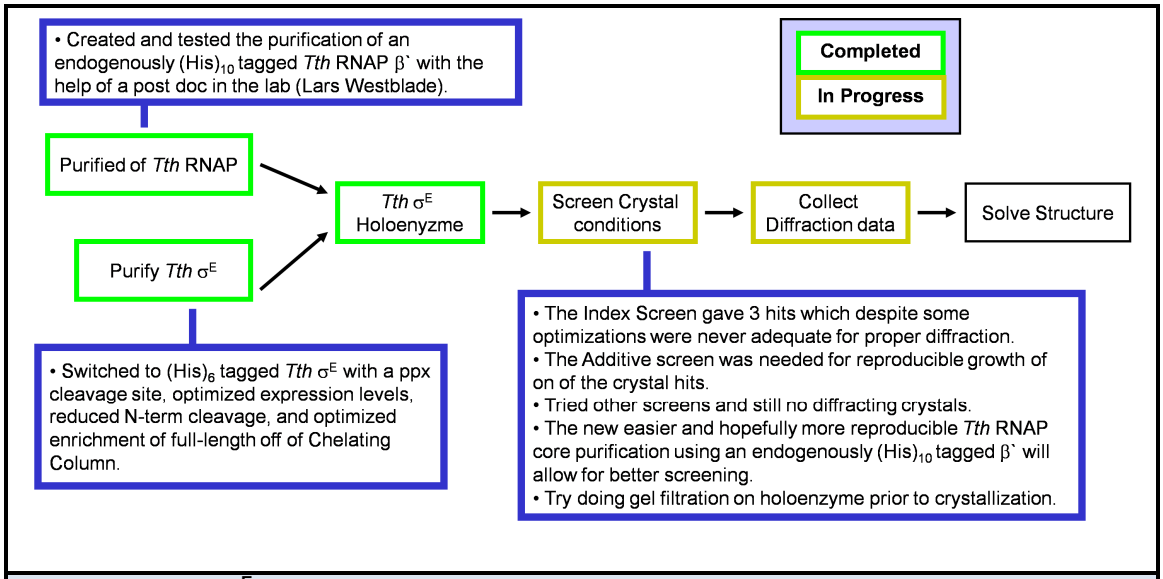


Figure 6.1 – *Tth* σ^E Holoenzyme Structure Progress Timeline.
Above is a schematic of this projects progress. Green boxes are completed parts. Yellow boxes are parts in progress or uncompleted. Black boxes are parts not started. Blue boxes important details for each step.

Identification of *Tth* σ^E

The high temperature stability of thermophilic proteins makes them prime candidates for protein crystallization. However, since there were no known thermophilic Group IV σ factors, we cloned a Group IV σ^E from the non-thermophilic bacteria *Deinococcus radiodurans* (*Dr*), known to be closely related to the thermophilic *Tth*. However, *Dr* σ^E proved to be highly insoluble during expression. As a result we looked for other candidates. With the availability of the *Tth* Group I holoenzyme structure [7] and *Tth* HB27 genome (working draft genome at Göttingen Genomics Laboratory at that time), we wished to identify a

Tth Group IV σ. Therefore, we performed a BLAST search against the *Tth* working draft genome using the known *Dr* σ^E sequence. Though the resulting sequence hit only shared 28% similarity with the *Dr* σ^E sequence, sequence analysis using the Clusters of Orthologous Groups of proteins (COG) database [33] showed it contained a rpoE/σ^E domain. This identification was later verified upon the publication of the finished *Tth* genome [112].

Cloning, Expression, and Purification of $(His)_6$-Thrombin-*Tth* σ^E

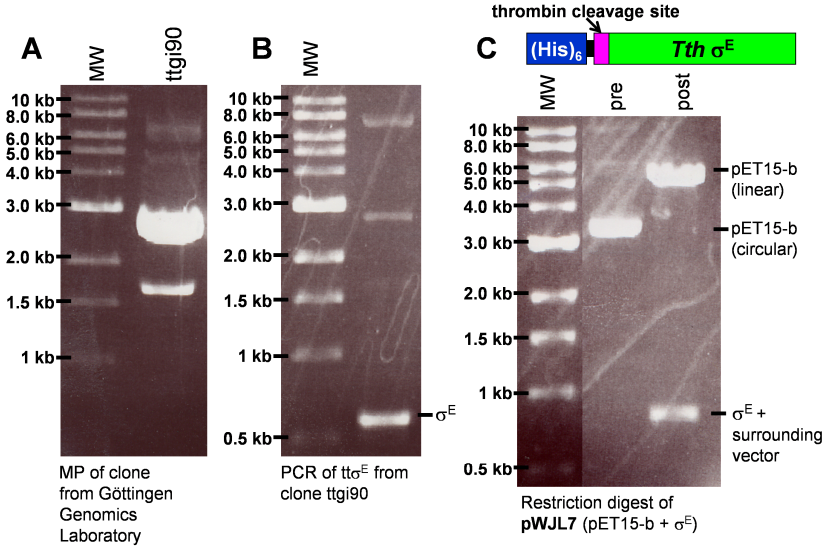


Figure 6.2 – $(His)_6$-Thrombin-*Tth* σ^E Cloning.

(A) Mimi-prep DNA from the ttgi90 clone received from the *Tth* genome sequencing group.

(B) PCR for for the *Tth* σ^E gene from the ttgi90 DNA from A.

(C) Restriction digest of mini-prep DNA from a clone containing the *Tth* σ^E gene ligated into a pET-15b vector, creating pWJL7. The post digestion sample shows that *Tth* σ^E gene was incorporated in to the pET15-b vector carried by the selected clone. DNA sequencing was used to confirm the sequence of the *Tth* σ^E gene insert.

We used PCR to generate *Tth* σ^E fragments from a sequencing clone (ttgi90) used by the *Tth* genome sequencing project. We cloned the *Tth* σ^E fragments into a pET15-b expression vector for expression of *Tth* σ^E with an N-term $(His)_6$

tagged removable by thrombin cleavage creating $(His)_6$-Thrombin-*Tth* σ^E (pWJL7).

IPTG induced protein expression of pWJL7 was performed in: BL21(DE3), BL21(DE3)pLysS, and Rosetta(DE3)pLysS (Novagen) expression systems. The BL21(DE3) cells were not viable after transformation. The BL21(DE3)pLysS cells were viable after transformation, but there was no visible expression of $(His)_6$-Thrombin-*Tth* σ^E and no enrichment using a Chelating column following cell lysis (Figure 6.3A). As shown in Figure 6.3B, analysis of the *Tth* σ^E DNA sequence revealed the presence of codons for tRNAs missing or expressed at low levels in *Ec* (rare codons).

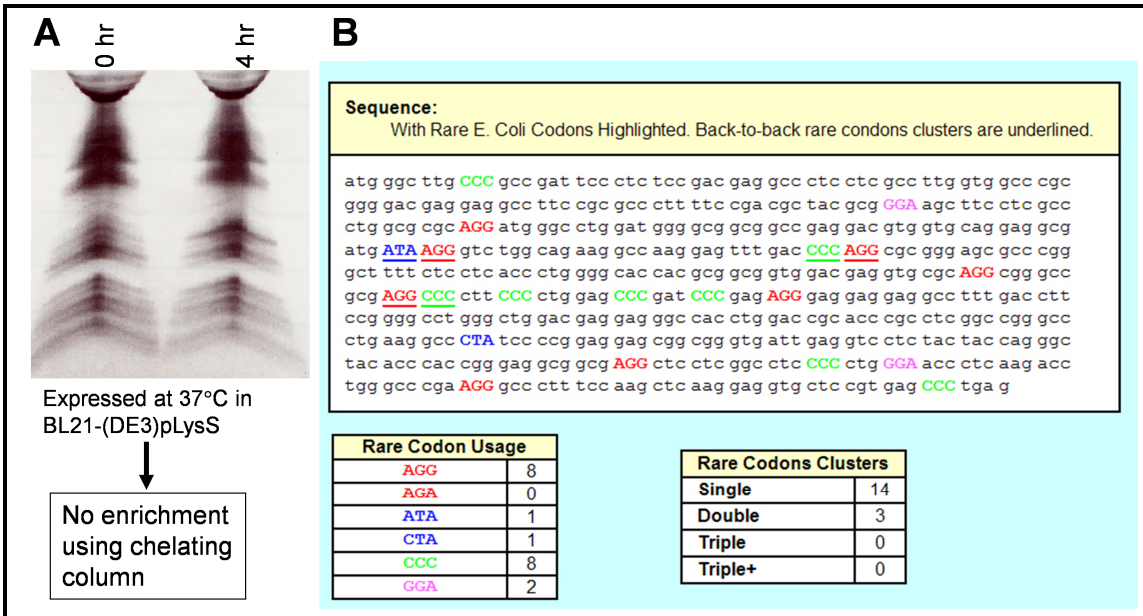


Figure 6.3 – $(His)_6$-Thrombin-*Tth* σ^E Expression and Rare Codon Usage.

(A) Expression of $(His)_6$-Thrombin-*Tth* σ^E in BL21-(DE3)pLysS cells resulted in no visible expression on a commasse stained 8-25% phast gel. This was confirmed by trying to enrich the $(His)_6$ tagged protein with a chelating column.

(B) Analysis of the *Tth* σ^E DNA sequence revealed the presence of rare codons. (Lane, WJ - http://lucare.com/cgi-bin/lab/rarc/rarc.pl).

Due to the presence of rare codons, we tried expression in Rosetta(DE3)pLysS cells which contain a plasmid that expresses six rare codon tRNAs. The approach proved promising, since expression in Rosetta(DE3)pLysS cells resulted in a prominent band around the correct MW that was soluble after cell lysis (Figure 6.4).

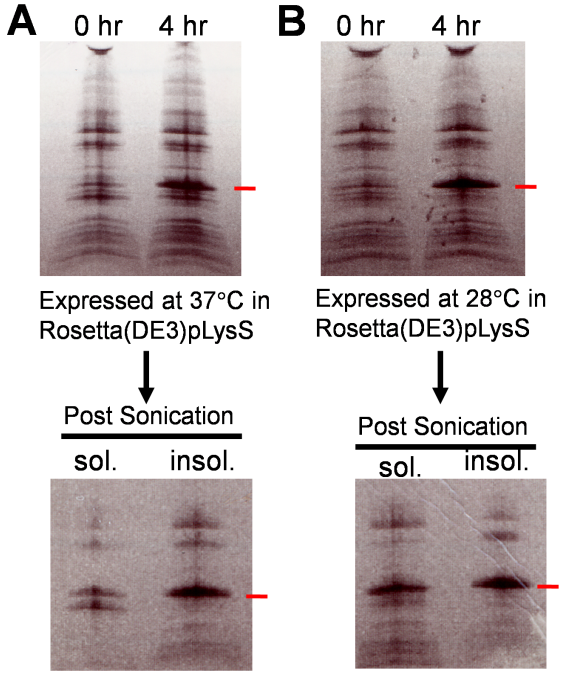


Figure 6.4 – $(His)_6$-Thrombin-*Tth* σ^E Expression in Rare Codon Supplemented Strain.

(A) Expression at 37 °C of $(His)_6$-Thrombin-Tth σ^E in Rosetta(DE3)pLysS cells which contain tRNAs for rare codons (AGA/AGG, AUA, CUA, CCC, GGA) resulted in a visible band of expression on a commasse stained 8-25% phast gel. The presumed expression product was also in the soluble fraction after cell lysis using sonication. The expected size of the expressed protein product is indicated by a red line.

(B) Same as (A), but expressed at 28 °C.

The purification *Tth* σ^E proved to be very problematic. One of the primary problems was degradation of the expressed protein. The degraded product was most evident after the chelating column step, which enriched two prominent bands (Figure 6.5). Since *Tth* σ^E is thermostable we tried heat treatment to

remove the *Ec* expression host proteases in an effort to reduce this degradation. However, heat treatment at 65 °C for 45 min after cell lysis, did not reduce the degradation. Interestingly, though the heat treatment did enrich three bands: the two bands that were previously enriched by the chelating column and one band that later went into the chelating column flow through. The protein that did not bind to the chelating column was identified as our antibiotic resistance protein chloramphenicol acetyl transferase (CAT) using N-term amino acid sequencing and its molecular weight. The two bands that bound to the chelating column were shown by N-term sequencing to contain the start of our expression vectors N-term $(His)_6$ tag. Electrospray mass spec on a heterogeneous sample containing the two bands tentatively indicated that the top band was full-length $(His)_6$-Thrombin-*Tth* σ^E and the bottom was a C-term truncated version of $(His)_6$-Thrombin-*Tth* σ^E.

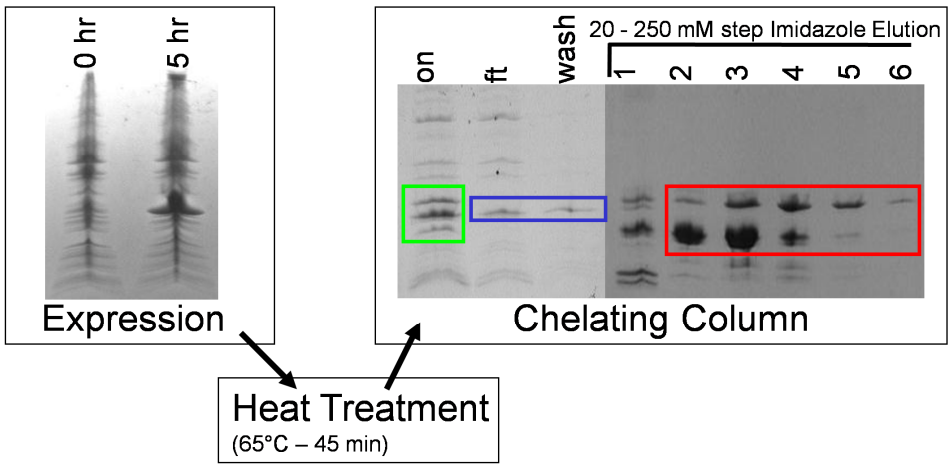


Figure 6.5 – $(His)_6$-Thrombin-*Tth* σ^E Expression and Purification.
Cell pellets after 5 hr of IPTG induction revealed a prominent band around the correct size. Heat treatment results in the enrichment of three bands (green region). The chelating column enriched two prominent bands (red region). The other heat treatment band went into the flow though (blue region).

To further characterize this degradation we performed thrombin treatment to remove the N-term $(His)_6$ tag and SP column ion exchange to separate the full-length and degraded protein products. Using this approach we were able to separate out a small amount of full-length *Tth* σ^E. We performed MALDI-TOF Mass Spec on several SP column fractions (Figure 6.6A) which revealed that the protein sample contained full-length *Tth* σ^E along with several protein species split between $\sigma^E_{4.1}$ and $\sigma^E_{4.2}$ and at the flexible liner between σ^E_2 and σ^E_4 (Figure 6.6B).

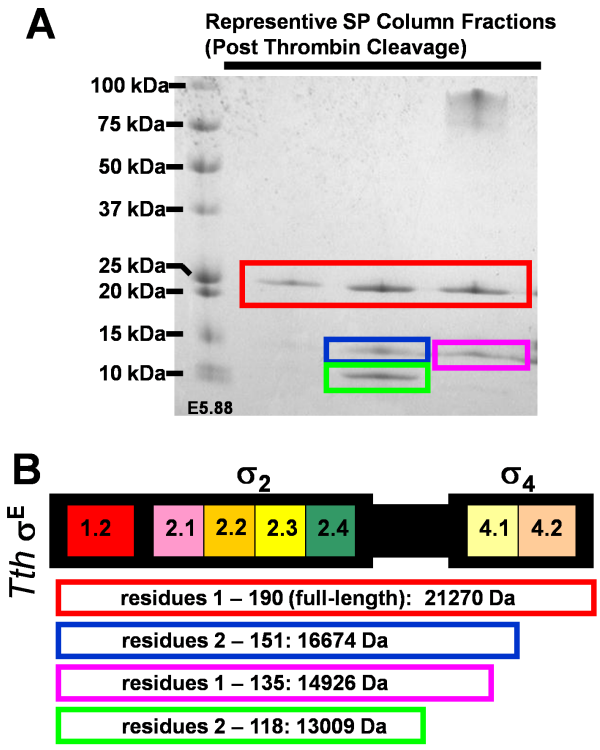


Figure 6.6 – Identification of $(His)_6$-Thrombin-*Tth* σ^E Degradation Products.

(A) Coomassie stained 8-25% phast gel of SP column fractions. The colored regions correspond to the identification of the degraded products on *Tth* σ^E domain architecture.

(B) Schematic showing the results of the MALDI-TOF Mass Spec analysis. We were able to positively identify full-length and three degraded products.

Subsequent, SDS gel analysis revealed that after chelating column enrichment the full-length and degraded $(His)_6$-Thrombin-*Tth* σ^E were completely stable even when left at RT for 30 days. Therefore, separating the full-length $(His)_6$-Thrombin-*Tth* σ^E from the truncated version would likely yield a stable homogenous sample of full-length $(His)_6$-Thrombin-*Tth* σ^E. Unfortunately, the full-length *Tth* σ^E purified using the SP column contained a small amount of degraded products after concentration. In addition, since it resulted in a very small yield we sought to develop another approach.

As seen in Figure 6.5, the final fraction (#6) from the chelating column imidazole step elution contained mostly full-length $(His)_6$-Thrombin-*Tth* σ^E protein. Unfortunately, $(His)_6$-Thrombin-*Tth* σ^E and *Tth* σ^E both precipitate out of solution between 1-2 mg/mL. Though this low solubility would normally be problematic for crystallization, in general it was not a concern for us since we were going to form *Tth* σ^E holoenzyme, in which *Tth* σ^E and RNAP core have a 1:13 molar mass ratio. However, the low solubility did severely limit the amount of purification steps we could perform, since at such low concentrations we would have a decreased percent recovery over each purification step. We therefore tried to optimize the chelating column in order to develop a one column purification. First we determined that by using extra long washes (50 mL lysis buffer at 5 mL/min followed by 20 mL of lysis buffer with 25 mM imidazole at 0.5 mL/min) and a slow imidazole gradient elution (25-250 imidazole over 32 mL at 0.5 mL/min) we could enhance the amount of full-length $(His)_6$-Thrombin-*Tth* σ^E in the final fractions. Interestingly, *Ec* RNAP bound to the chelating column and

eluted early in the imidazole gradient, indicating that there might be some interaction between *Ec* RNAP and $(His)_6$-Thrombin-*Tth* σ^E. We also increased the amount of soluble full-length $(His)_6$-Thrombin-*Tth* σ^E by optimizing the expression to the following steps: transformed expression vector into Rosetta(DE3)pLysS cells and grew overnight at 37 °C, put plates at 4 °C overnight, inoculated expression cultures and grew at 37 °C until OD 0.2 then lowered temperature to 20 °C with induction using 1 mM IPTG at OD 0.6 for 4 hrs, and after cell lysis we recovered some addition protein from the insoluble fraction using a high salt wash.

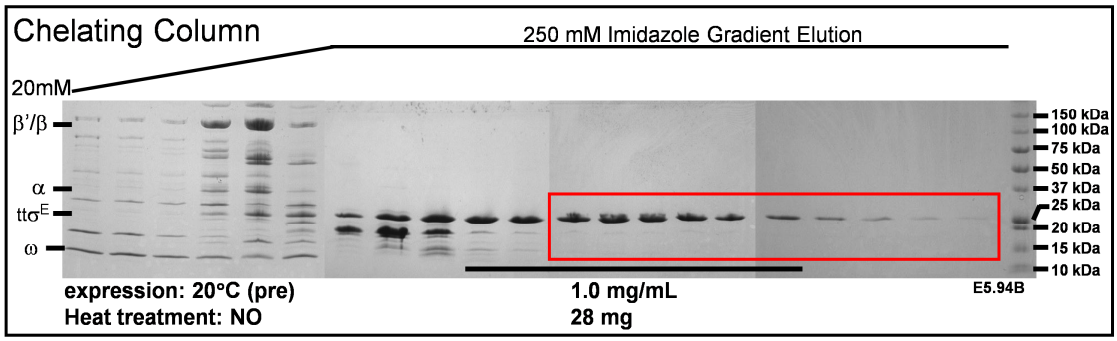


Figure 6.7 – Optimized $(His)_6$-Thrombin-*Tth* σ^E Chelating Column Purification.
By optimizing the expression of soluble protein and the chelating column separation we were able to purify relatively clean full-length $(His)_6$-Thrombin-*Tth* σ^E (red region).

Expressing proteins at a low temperature is a well established method for increasing the solubility of difficult proteins. However, to our knowledge, there is no evidence in the literature about leaving transformed plates at 4 °C overnight prior to growth and expression, something that seemed to significantly increase not the just the level of soluble expressed protein for $(His)_6$-Thrombin-*Tth* σ^E, but also for $(His)_6$-Thrombin-*Ec* $\sigma^E{}_4$. Perhaps the overnight incubation at 4 °C, which

would have induced the *Ec* cold shock response, lead to some change that aided in the expression of soluble protein. Taken together, we were able to use the chelating column to do a one column purification of extremely pure full-length $(His)_6$-Thrombin-*Tth* σ^E (Figure 6.7).

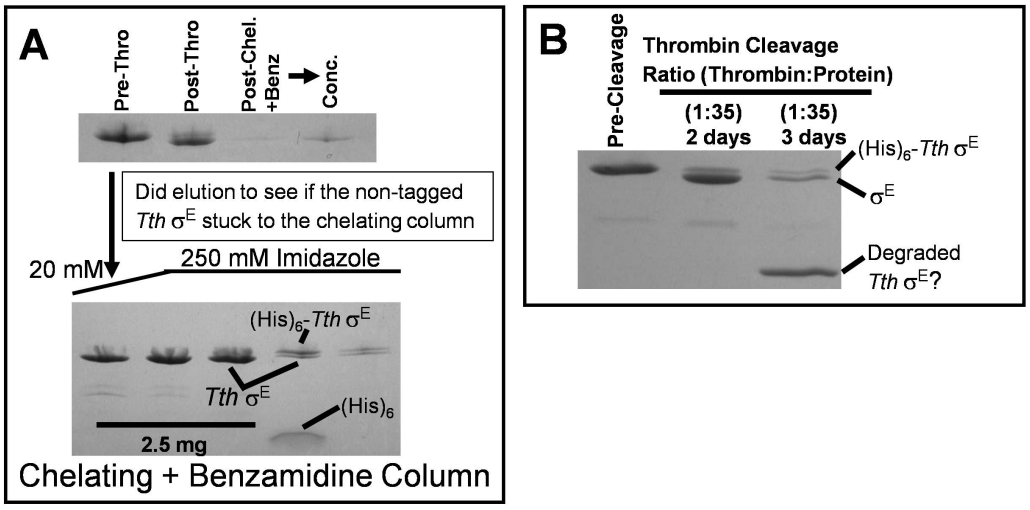


Figure 6.8 – Removal of $(His)_6$ tag from $(His)_6$-Thrombin-*Tth* σ^E using Thrombin.

(A) Thrombin cleavage followed by chelating and benzamidine columns. The bottom gel show the imidazole elution from the chelating column indicating that even the non-tagged *Tth* σ^E has weak affinity.

(B) Attempt to increase the efficiency of thrombin cleavage by increasing the amount of thrombin and the time of incubation. However, neither increased the efficiency to 100% and instead we started to get non-specific thrombin cleavage of the protein.

We encountered additional difficulties when trying to remove the thrombin cleavable $(His)_6$ tag from $(His)_6$-Thrombin-*Tth* σ^E. We initially tried overnight 4 °C incubation with thrombin at a mass ratio of 1:200 (thrombin:$(His)_6$-Thrombin-*Tth* σ^E). However, when the protein was run on a gel there was a small amount of protein that still contained the tag. Normally, this would not be a problem since the next step is to run the protein through a chelating column (removing the protein still tagged along with the cleaved tags) and benazmidine column (to

remove thrombin) with the non-tagged protein in the flow through. Unfortunately, when we did this step there was little to no protein in the flow through. We therefore eluted the chelating column using an imidazole gradient, revealing that the non-tagged *Tth* σ^E possessed weak affinity for the chelating column.

Therefore, since we could not easily use a chelating column to separate out the non-tagged *Tth* σ^E, we tried to increase the amount thrombin and cleavage time. However, at very high mass ratios (1:35) we were not able to get 100% cleavage of the $(His)_6$ tag. Even with three days of thrombin cleavage (some of it at room temperature) we were not able to get 100% cleavage efficiency, rather, the thrombin non-specifically cleaved most of the protein in the sample. This lead us to seek other methods to separate out the non-tagged *Tth* σ^E. However, as mentioned before due to the low solubility of the protein (1-2 mg/mL) this was not very feasible.

Cloning, Expression, and Purification of a PPX Cleavable $(His)_6$ Tagged *Tth* σ^E

As a result of the problems using thrombin to remove the $(His)_6$ tag we decided to try a Precision Protease (PPX) cleavage site between our $(His)_6$-tag and *Tth* σ^E. We sub-cloned *Tth* σ^E from pWJL7 into a pre-existing pET-28a derived vector that contained a PPX cleavage site, creating pWJL8. Since, PPX is more specific than thrombin we reasoned that we could use our optimized chelating column purification as before, but since PPX would be less prone to non-specific cleavage we could safely use high levels of PPX to get 100% tag removal.

Using this approach we were able to increase the levels of PPX for 100% tag removal (1 mg PPX : 5 mg $(His)_6$-PPX-*Tth* σ^E), while at the same time *Tth* σ^E was spared from non-specific cleavage. However, the removal of the PPX proved problematic due to the larger than usual amount of PPX added. In addition, since *Tth* σ^E precipitates at ~1mg/mL the protein sample was in a relatively large volume (usually between 50 – 100 mL). The large sample volume along with the amount of PPX added necessitated three rounds of GST chromatography in order to remove the GST tagged PPX. Nevertheless, we were able to purify 18 mg (0.9 mg/mL) of *Tth* σ^E from an 8 L culture (Figure 6.9).

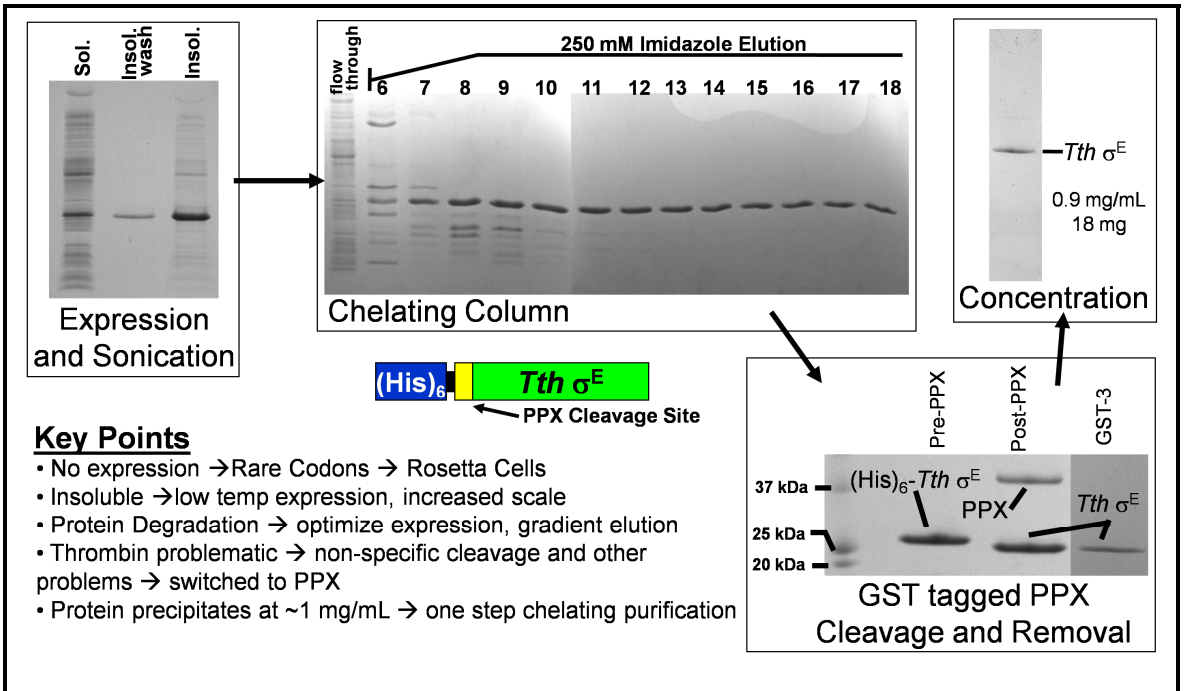


Figure 6.9 – $(His)_6$-PPX-*Tth* σ^E Expression and Purification.

Final expression and purification scheme used for *Tth* σ^E.

Preparation of *Tth* σ^E Holoenzyme

We purified endogenous *Tth* RNAP core using previously established methods as shown in Figure 6.10.

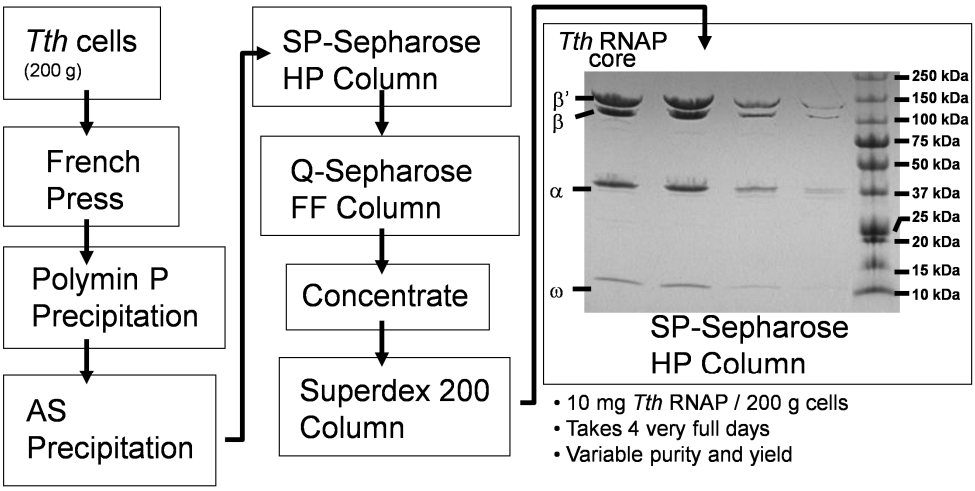


Figure 6.10 – Purification of Endogenous *Tth* RNAP core.
Endogenous *Tth* RNAP core was FPLC purified from *Tth* cells.

We preformed a native gel shift (Figure 6.11) in order to evaluate the binding of *Tth* σ^E to *Tth* RNAP core (*Tth* σ^E Holoenzyme). The gel shift indicated that at even at a 1:1 molar ratio we were able to form *Tth* σ^E Holoenzyme.

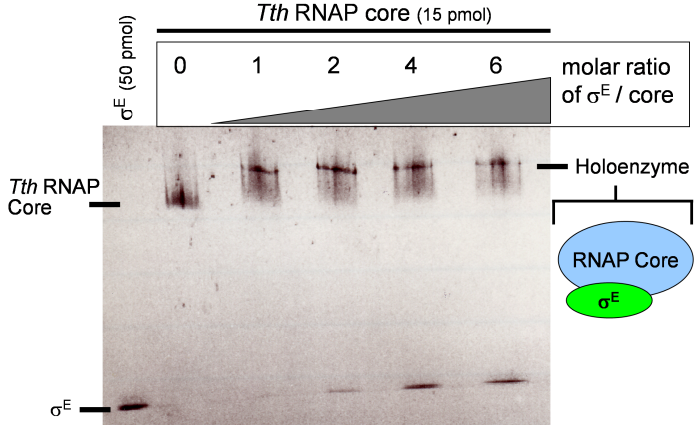


Figure 6.11 – *Tth* σ^E Holoenzyme Gel Shift.
Increasing amounts of *Tth* RNAP core were added to *Tth* σ^E and run on a native gel. As the gel shows *Tth* σ^E efficiently bound to *Tth* RNAP core and was able to shift core at a 1:1 molar ratio.

Crystallization of *Tth* σ^E Holoenzyme

The Index Screen (Hampton Research) revealed a total of three potential crystallization conditions for *Tth* σ^E Holoenzyme.

1st Crystal Form – 0.2 M tri-Sodium Citrate dihydrate, 20% w/v Polyethylene Glycol 3350 (Index Screen Condition #94). The screen produced small needles amongst a background of precipitation. This condition was optimized to yield larger crystals (Figure 6.12A), but in the end this condition was abandoned since the crystals did not diffract at NSLS X25A (Figure 6.12B).

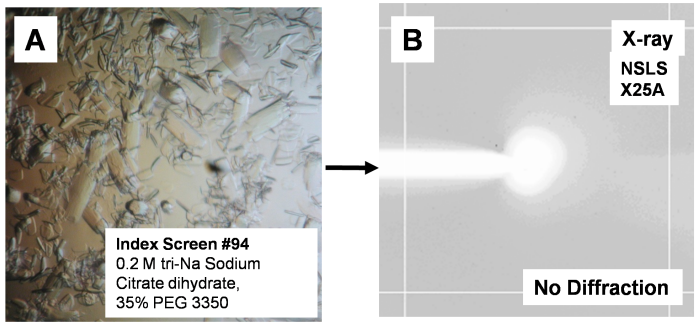


Figure 6.12 – 1st Crystal Form for *Tth* σ^E Holoenzyme.

(A) Crystals grown during screening and later optimized.

(B) The crystals growing during screening (A) did not diffract.

2nd Crystal Form – 0.1 M HEPES pH 7.0, 30% v/v Jeffamine ED-2001 Reagent pH 7.0 (Index Screen Condition #39). This screen produced paper thin 300 μm wide hexagonal plates (Figure 6.13A). These crystals were frozen in ethane and diffracted to ~9 Å at NSLS X25A (Figure 6.13B). However, several initial attempts at reproducing the crystals, by varying the pH, precipitant concentration, drop size, protein concentration, and protein batches all failed. In addition, seeding from the initial crystals did not initiate crystal growth. The crystals were finally reproduced using the Additive Screen 1 (Hampton

Research) along with the initial condition. Hexagonal plates were formed by adding 0.2 uL of 0.1 Magnesium Chloride hexahydrate or 0.1 M Calcium Chloride dihydrate to a 1 uL protein + 1 uL well solution drop. The hexagonal plates from the additive screen were 40 um thick and 150 um wide. However, these crystals did not diffract at NSLS X9A. In order to get diffraction quality crystals we further optimized both the purification of *Tth* RNAP core as well as the crystallization conditions (Well: 1 mL of 0.1 M HEPES pH 7.0, 32% v/v Jeffamine ED-2001 Reagent pH 7.0) (Drop: 1 uL protein range between 12.5 – 20 mg/mL + 1uL well solution + 0.2 uL of 0.1 M $CaCl_2$ additive) to grow 100 um thick and 300 um wide hexagonal plates both at 4 °C and 22 °C (Figure 6.13C). The crystals from the 2^{nd} crystal form were washed, solublized and run on a gel to verify that they contained *Tth* σ^E holoenzyme (Figure 6.13D). The optimized 22 °C crystals diffracted to 9 Å at (NSLS X9A, NSLS X25A). Crystal shrinking, by sequentially incubating the crystals in increasing amount of PEG 400, improved the diffraction to 7.5 Å (NSLS X9A, NSLS X25A) (Figure 6.13E, F). Initial analysis indicated that the space group was hexagonal with unit cell parameters a=b=253 A, c=392.6 A with $\alpha=\beta=90°$, $\gamma=120°$. However, when the crystals were positioned such that the x-ray shot the edge of the hexagonal plate, the diffraction spots became smeared (Figure 6.13F). We believe that this indicates that plates are composed of hexagonal layers which are not properly stacked one upon the next. We tried several batches of crystals, but in the end we were never able to overcome this limitation.

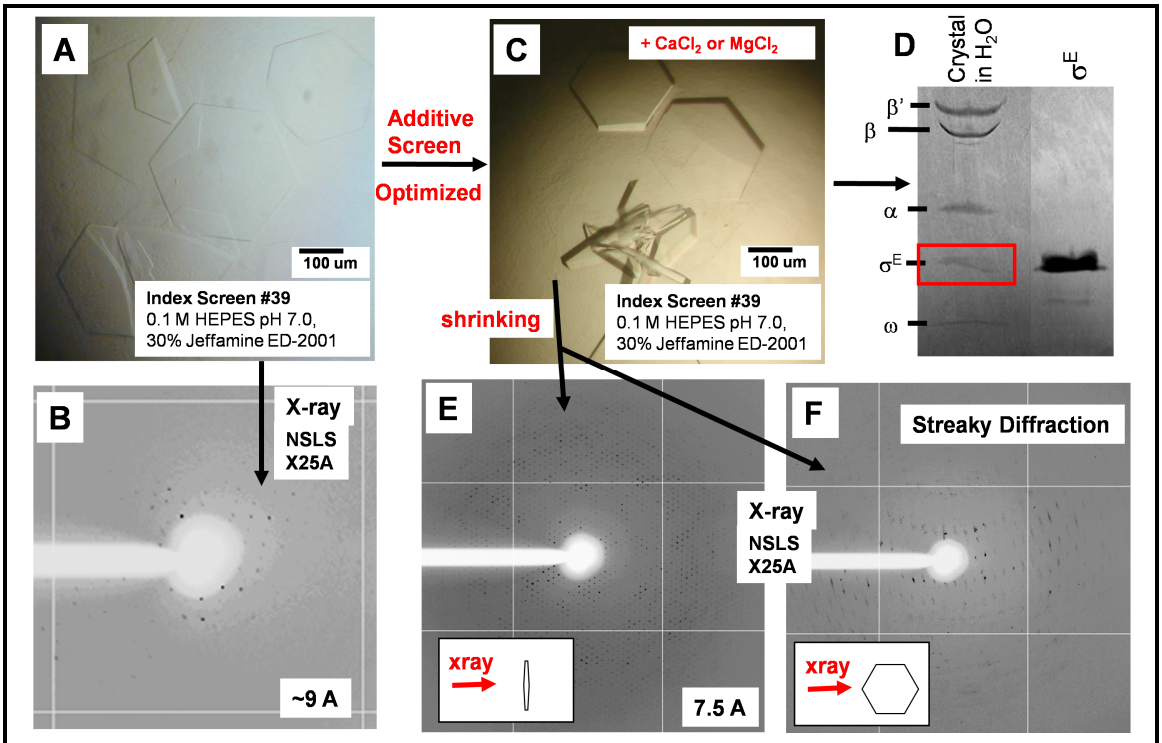


Figure 6.13 – 2nd Crystal Form for *Tth* σ^E Holoenzyme.

(A) Crystals grown during screening.

(B) The crystals growing during screening (A) diffracted weakly.

(C) We were not initially able to re-grow the crystals seen in the screen. However, an additive screen using the crystallization conditions revealed that adding $CaCl_2$ $MgCl_2$ resulted in reproducible crystal growth.

(D) Crystals from (C) were washed twice, solublized in water and run on a SDS gel. This indicated that the crystals contained *Tth* σ^E Holoenzyme.

(E/F) By using crystal shrinking we were able to improve the diffraction to 7.5 Å (E), however when shot edge on the crystals reproducibly gave a smeared diffraction pattern (F).

3rd Crystal Form – 2.8 M Sodium Acetate trihydrate pH 7.0 (Index Screen Condition #24) grew crystals after about 12 months (Figure 6.14A). The well solution and drops evaporated significantly before these crystals formed. One of the almost completely dry drops was washed several times with well solution after which the crystals were dissolved in water and run on a 8-25% phast gel

(Figure 6.14B). Silver staining revealed that the crystals contained *Tth* σ^E holoenzyme, which remarkably had not degraded over time. The crystals were tested at NSLS X9A in both capillary and flash frozen using 6 M Na Formate as a cryoprotectant. Unfortunately the crystals did not diffract. However, this could have been simply due to age and/or crystal damage due to the drop drying. Therefore, we tried to reproduce these crystals without the one year wait. We found that we were able to reproduce crystals by using the following crystallization condition: well solution = 1 mL of 3.5-4.0 M Na Acetate trihydrate pH 7.0, drop = 1 uL of 20 mg/mL *Tth* σ^E holoenzyme + 1 uL well solution, crystals grew after about 2 months at 22 °C. Unfortunately, these crystals also did not diffract. Although this drop had not dried like the first time, the crystals were frozen several months after they appeared. Perhaps they might diffract if frozen immediately after formation.

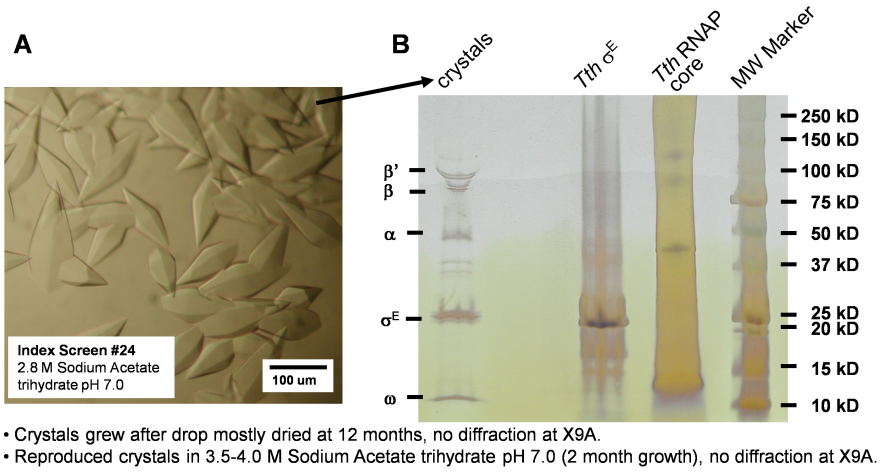


Figure 6.14 – 3rd Crystal Form for *Tth* σ^E Holoenzyme.

(A) Crystals grown after 1 year in a screening condition. This condition was later optimized to give reproducible crystals growth within 2 months.

(B) Crystals from (A) were washed twice, solublized in water and run on a SDS gel, silver staining revealed the presence of *Tth* σ^E Holoenzyme.

Using Crystal Screen (Hampton Research), Crystal Screen 2 (Hampton Research), and PEG/Ion Screen (Hampton Research) we were able to get three additional crystal hits (Figure 6.15) after 6-12 months when the drops had mostly dried up: (1) Crystal Screen Condition #33 – 4.0 M Sodium formate, (2) Crystal Screen 2 Condition #35 – 0.1 M HEPES pH 7.5, 70% v/v (+/-)-2-Methyl-2,4-pentanediol, and (3) Crystal Screen #12 – 0.2 M Magnesium chloride hexahydrate, 0.1 M HEPES sodium pH 7.5, 30% v/v 2-Propanol. However, these were never followed up.

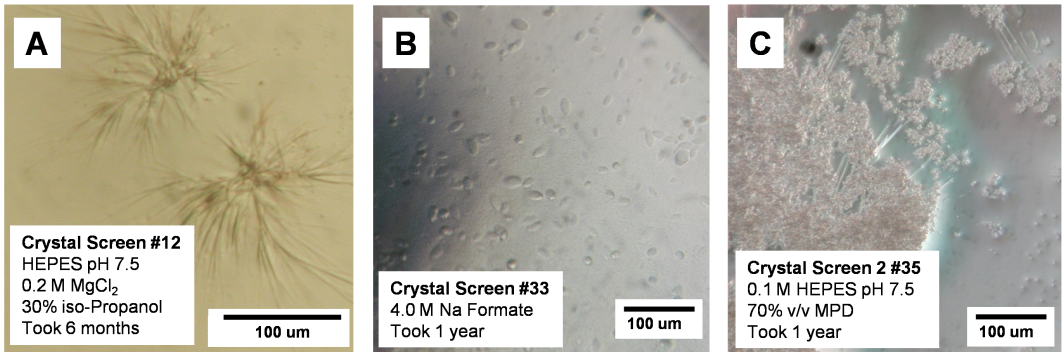


Figure 6.15 – Additional Crystal Hits.
(A/B/C) Crystals grown after 6-12 months in a screening condition. These were never followed up.

An Improved Method of *Tth* RNAP Core Purification

The lack of large quantities of consistent and homogenous protein can be a real bottleneck when trying to crystallize *Tth* RNAP associated structures, since the prep to prep variability made systematic crystal screening difficult. In addition, since the purification does not rely on affinity tags or over expression the protein yield per purification is low. Unfortunately, previous attempts by others to over express *Taq* and *Tth* RNAP core in *Ec* and then crystallize have proven technically challenging. As a result Lars Westblade (a post doc in the lab) $(His)_{10}$

tagged the chromosomal copy of rpoC in *Tth* HB8 cells. Thereby, allowing for tagged affinity purification of endogenously expressed *Tth* RNAP.

Together with Lars Westblade, we did a trial purification from 2x 2L of Tth HB8 cells containing $(His)_{10}$ tagged β′ (Figure 6.16). From approximately 5 g of cells were able to quickly get 2 mg of very pure *Tth* RNAP core using only a single chelating column. The previous purification strategy used 200 g of cells to get at best 10 mg of heterogeneous fractions after several days and may purification steps. If scaled up this new strategy could prove very useful for future attempts at crystallizing *Tth* σ^E holoenzyme as well as other *Tth* RNAP associated structures.

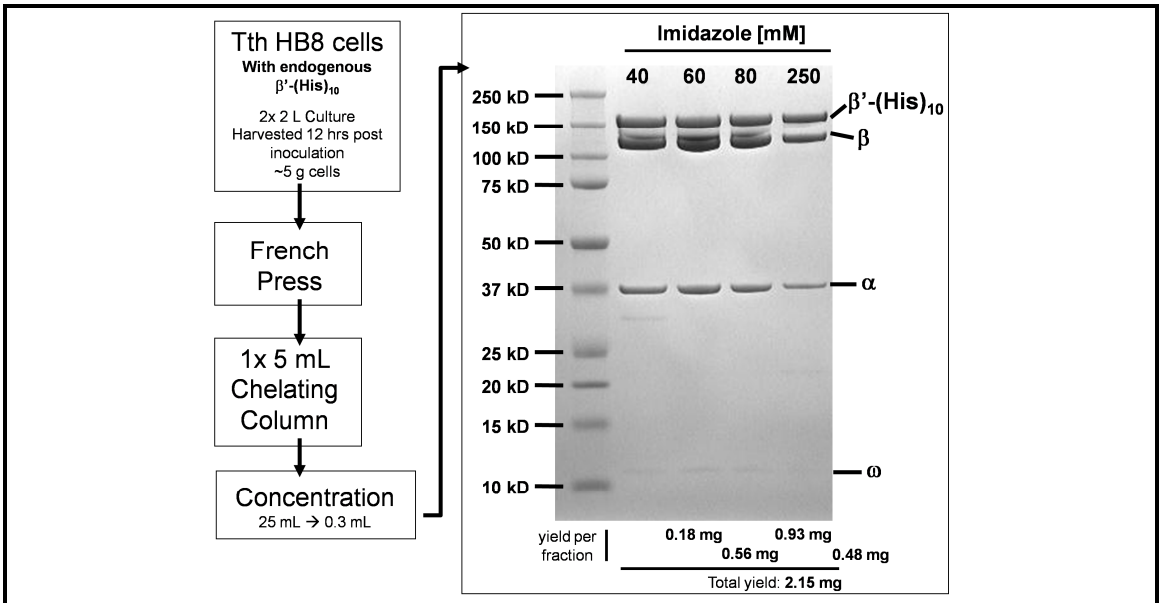


Figure 6.16 – Improved Purification of Endogenous *Tth* RNAP core.
A $(His)_{10}$ tag was added to the chromosomal copy rpoC, the gene that encodes the β′ subunit. This tag can be used to FPLC purify endogenous *Tth* RNAP core.

In particular, if the final fractions are more homogenous or just of larger yield we could consider pooling them together and forming *Tth* σ^E holoenyzme in batch followed by an additional gel purification step. Previously, we formed *Tth*

σ^E holoenyzme prior to each crystallization. We did this since in general when using the old purification method each fraction of *Tth* RNAP core from the final column contained small levels of different molecular weight contaminants and each seemed to posses noticeable crystallization differences. Therefore, since we choose not to pool together the *Tth* RNAP core fractions from the final column, we needed to form *Tth* σ^E holoenyzme in small batches and it was feared that most of the sample would be lost if an additional purification step was attempted.

Conclusions

Despite a large amount of work we were never able to produce diffraction quality crystals of a *Tth* σ^E holoenzyme. However, we were successful in overcoming several technical challenges in determining the conditions necessary to express and purify *Tth* σ^E (Figure 6.9). Importantly, we were able to express soluble full-length protein using expression strains with rare codon coverage and expressing at low temperature. We then optimized the chelating column step and used a PPX cleavable $(His)_6$ tag to successfully purify *Tth* σ^E. We have also started to develop new and improved methods of *Tth* RNAP core purification (Figure 6.16) that might allow for better crystal screening. In addition to obtaining and optimizing 3 crystal forms that did not produce satisfactory diffraction, we found 3 crystal hits that were never optimized to grow crystals suitable for diffraction testing (Figure 6.15).

Materials and Methods

Cloning of $(His)_6$-Thrombin-*Tth* σ^E (pWJL7)

Upstream Primer

5′-GGGTGGAGGCatATGGGCTTGCCCGCCGATTC-3′ ttSigE-s-NdeI

Downstream Primer

5′-AGAAGTTCCCGGggaTCCTCAGGGCTCACGGA-3′ ttSigE-stop-BamHI

Sequence of Full-Length *Tth* rpoE (σ^E)

Nde I

5′GGGTGGAGGCa|tATGGGCTTGCCCGCCGATTCCCTCTCCGACGAGGCCCTCCTCGCCTTGGTGGCCCGCGGGG
ACGAGGAGGCCTTCCGCGCCCCTTTTCCGACGCTACGCGGGAAGCTTCCTCGCCCTGGCGCGCAGGATGGGCCTGGA
TGGGGCGGCGGCCGAGGACGTGGTGCAGGAGGCGATGATAAGGGTCTGGCAGAAGGCCAAGGAGTTTGACCCCAGG
CGCGGGAGCGCCCGGGCTTTTCTCCTCACCCTGGGGCACCACGCGGCGGTGGACGAGGTGCGCAGGCGGGCCGCGA
GGCCCCTTCCCCTGGAGCCCGATCCCGAGAGGGAGGAGGAGGCCTTTGACCTTCCGGGGCCTGGGCTGGACGAGGA
GGGCCACCTGGACCGCACCCGCCTCGGCCGGGCCCTGAAGGCCCTATCCCCGGAGGAGCGGCGGGTGATTGAGGTC
CTCTACTACCAGGGCTACACCCACCGGGAGGCGGCGAGGCTCCTCGGCCTCCCCCTGGGAACCCTCAAGACCTGGG
CCCGAAGGGCCCTTTCCAAGCTCAAGGAGGTGCTCCGTGAGCCCTGAG|GAtccCCGGGAACTTCT-3′

BamH I

pWJL7 Cloning and Expression Region Sequence

- rbs, $(His)_6$, thrombin site, upstream primer, ttSigE, STOP codon, downstream primer

AGATCTCGATCCCGCGAAATTAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTAGAAATAA
TTTTGTTTAACTTTAAGAAGGAGATATACCATGGGCAGCAGCCATCATCATCATCATCACAGCAGCGGCCTGGTGCC
GCGCGGCAGCCATATGGGCTTGCCCGCCGATTCCCTCTCCGACGAGGCCCTCCTCGCCTTGGTGGCCCGCGGGGACG
AGGAGGCCTTCCGCGCCCCTTTTCCGACGCTACGCGGGAAGCTTCCTCGCCCTGGCGCGCAGGATGGGCCTGGATGGG
GCGGCGGCCGAGGACGTGGTGCAGGAGGCGATGATAAGGGTCTGGCAGAAGGCCAAGGAGTTTGACCCCAGGCGCGG
GAGCGCCCGGGCTTTTCTCCTCACCCTGGGGCACCACGCGGCGGTGGACGAGGTGCGCAGGCGGGCCGCGAGGCCCC
TTCCCCTGGAGCCCGATCCCGAGAGGGAGGAGGAGGCCTTTGACCTTCCGGGGCCTGGGCTGGACGAGGAGGGCCAC
CTGGACCGCACCCGCCTCGGCCGGGCCCTGAAGGCCCTATCCCCGGAGGAGCGGCGGGTGATTGAGGTCCTCTACTA
CCAGGGCTACACCCACCGGGAGGCGGCGAGGCTCCTCGGCCTCCCCCTGGGAACCCTCAAGACCTGGGCCCGAAGGG
CCCTTTCCAAGCTCAAGGAGGTGCTCCGTGAGCCCTGAGGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTT
GGCTGCTGCCACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTG

Expression Product

- Molecular weight (PAWS Average Mass): 23151.39 D

MGSSHHHHHHSSGLVPRGSHMGLPADSLSDEALLALVARGDEEAFRALFRRYAGSFLALARRMGLDGAAAEDVVQEA
MIRVWQKAKEFDPRRGSARAFLLTLGHHAAVDEVRRRAARPLPLEPDPEREEEAFDLPGPGLDEEGHLDRTRLGRAL
KALSPEERRVIEVLYYQGYTHREAARLLGLPLGTLKTWARRALSKLKEVLREP

Post Thrombin Cleavage Product

- Molecular weight (PAWS Average Mass): 21269.33 D

GSHMGLPADSLSDEALLALVARGDEEAFRALFRRYAGSFLALARRMGLDGAAAEDVVQEAMIRVWQKAKEFDPRRGS
ARAFLLTLGHHAAVDEVRRRAARPLPLEPDPEREEEAFDLPGPGLDEEGHLDRTRLGRALKALSPEERRVIEVLYYQ
GYTHREAARLLGLPLGTLKTWARRALSKLKEVLREP

Figure 6.17 – Cloning Details for pWJL7.

Figure 6.17 shows the primers, cloning and expression region sequence, and protein products for pWJL7 which was used to express $(His)_6$-Thrombin-*Tth* σ^E. We used PCR on a sequencing clone (ttgi90) used by the *Tth* genome sequencing project (upstream primer ttSigE-s-NdeI and downstream primer ttSigE-stop-BamHI) to generate *Tth* σ^E fragments with engineered restriction sites. We then cloned the *Tth* σ^E fragments into a pET15-b expression vector for expression of *Tth* σ^E with an N-term $(His)_6$ tagged removable by thrombin

cleavage creating $(His)_6$-Thrombin-*Tth* σ^E (pWJL7). The parent vector for pWJL7 was an ampicillin resistant pET-15b expression vector (Novagen).

Cloning of $(His)_6$-PPX-*Tth* σ^E (pWJL8)

Upstream Primer

5′-GGGTGGAGGCatATGGGCTTGCCCGCCGATTC-3′ ttSigE-s-NdeI

Downstream Primer

5′-AGAAGTTCCCGGggaTCCTCAGGGCTCACGGA-3′ ttSigE-stop-BamHI

Sequence of Full-Length *Tth* rpoE (σ^E)

Nde I

5′GGGTGGAGGCa|tATGGGCTTGCCCGCCGATTCCCTCTCCGACGAGGCCCTCCTCGCCTTGGTGGCCCGCGGGG
ACGAGGAGGCCTTCCCCCCCCTTTTCCGACGCTACGCGGGAAGCTTCCTCGCCCTGGCGCGCAGGATGGGCCTGGA
TGGGGCGGCGGCCGAGGACGTGGTGCAGGAGGCGATGATAAGGGTCTGGCAGAAGGCCAAGGAGTTTGACCCCAGG
CGCGGGAGCGCCCGGGCTTTTCTCCTCACCCTGGGGCACCACGCGGCGGTGGACGAGGTGCGCAGGCGGGCCGCGA
GGCCCCTTCCCCTGGAGCCCGATCCCGAGAGGGAGGAGGAGGCCTTTGACCTTCCGGGGCCTGGGCTGGACGAGGA
GGGCCACCTGGACCGCACCCGCCTCGGCCGGGCCCTGAAGGCCCTATCCCCGGAGGAGCGGCGGGTGATTGAGGTC
CTCTACTACCAGGGCTACACCCACCGGGAGGCGGCGAGGCTCCTCGGCCTCCCCCTGGGAACCCTCAAGACCTGGG
CCCGAAGGGCCCTTTCCAAGCTCAAGGAGGTGCTCCGTGAGCCCTGAG|GAtccCCGGGAACTTCT-3′

BamH I

pWJL8 Cloning and Expression Region Sequence

- rbs, $(His)_6$, PreScission Protease site, upstream primer, ttSigE, STOP codon, downstream primer

TAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTAGAAATAATTTTGTTTAACTTTAAGAAG
GAGATATACCATGGGCAGCAGCCATCATCATCATCATCACAGCAGCGGCCTGGAAGTTCTGTTCCAGGGGCCCCATA
TGGGCTTGCCCGCCGATTCCCTCTCCGACGAGGCCCTCCTCGCCTTGGTGGCCCGCGGGGACGAGGAGGCCTTCCGC
GCCCTTTTCCGACGCTACGCGGGAAGCTTCCTCGCCCTGGCGCGCAGGATGGGCCTGGATGGGGCGGCGGCCGAGGA
CGTGGTGCAGGAGGCGATGATAAGGGTCTGGCAGAAGGCCAAGGAGTTTGACCCCAGGCGCGGGAGCGCCCGGGCTT
TTCTCCTCACCCTGGGGCACCACGCGGCGGTGGACGAGGTGCGCAGGCGGGCCGCGAGGCCCCTTCCCCTGGAGCCC
GATCCCGAGAGGGAGGAGGAGGCCTTTGACCTTCCGGGGCCTGGGCTGGACGAGGAGGGCCACCTGGACCGCACCCG
CCTCGGCCGGGCCCTGAAGGCCCTATCCCCGGAGGAGCGGCGGGTGATTGAGGTCCTCTACTACCAGGGCTACACCC
ACCGGGAGGCGGCGAGGCTCCTCGGCCTCCCCCTGGGAACCCTCAAGACCTGGGCCCGAAGGGCCCTTTCCAAGCTC
AAGGAGGTGCTCCGTGAGCCCTGAGGATCCGAATTCGAGCTCCGTCGACAAGCTTGCGGCCGCACTCGAGCACCACC
ACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCACCGCTGAGCAATAA
CTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTG

Expression Product

- Molecular weight (PAWS Average Mass): 23425.71 D

MGSSHHHHHHSSGLEVLFQGPHMGLPADSLSDEALLALVARGDEEAFRALFRRYAGSFLALARRMGLDGAAAEDVVQ
EAMIRVWQKAKEFDPRRGSARAFLLTLGHHAAVDEVRRRAARPLPLEPDPEREEEAFDLPGPGLDEEGHLDRTRLGR
ALKALSPEERRVIEVLYYQGYTHREAARLLGLPLGTLKTWARRALSKLKEVLREP

Post Thrombin Cleavage Product

- Molecular weight (PAWS Average Mass): 21279.37 D

GPHMGLPADSLSDEALLALVARGDEEAFRALFRRYAGSFLALARRMGLDGAAAEDVVQEAMIRVWQKAKEFDPRRGS
ARAFLLTLGHHAAVDEVRRRAARPLPLEPDPEREEEAFDLPGPGLDEEGHLDRTRLGRALKALSPEERRVIEVLYYQ
GYTHREAARLLGLPLGTLKTWARRALSKLKEVLREP

Figure 6.18 – Cloning Details for pWJL8.

We sub-cloned *Tth* σ^E from pWJL7 into a pre-existing pET-28a derived vector that contained a PPX cleavage site, creating pWJL8. Figure 6.18 shows the primers, cloning and expression region sequence, and protein products for pWJL8 which was used to express $(His)_6$-PPX-*Tth* σ^E. The parent vector for pWJL8 was a kanamycin resistant vector derived by replacing the thrombin cleavage site of a pET-28a expression vector (Novagen) with a PPX cleavage site.

Expression and Purification of *Tth* σ^E from $(His)_6$-PPX-*Tth* σ^E (pWJL8)

Figure 6.9 shows the expression and purification scheme. pWJL8 was transformed into *Ec* Rosetta(DE3)pLysS (Novagen) cells, plates left at 4 °C overnight, and transformants were grown at in LB medium with chloramphenicol (30 μg/mL) for selection of the expression vector and kanamycin (30 μg/mL) for selection of the rare codon plasmid contained in the expression host. The cells were grown at 37 °C to an OD_{600} of 0.2, temperature lowered to 20 °C and grown overnight until an OD_{600} of 0.7. Protein expression was induced with 1 mM IPTG for 4 hrs. Cells containing the overexpressed protein were harvested and resuspended in lysis buffer [20 mM Tris-HCl (pH 8.0), 0.5 M NaCl, 5% glycerol, 0.1 mM EDTA, 5 mM imidazole (pH 8.0), 0.5 mM β-mercaptoethanol (β-ME), and 1x protease inhibitor cocktail]. 100x protease inhibitor cocktail was made in cold ethanol with 17.4 mg/mL PMSF, 31.2 mg/mL Benzamidine, 0.5 mg/mL Chymostatin, 0.25 mg/mL Leupeptin, 0.1 mg/mL Pepstatin, and 0.5 mg/mL Aprotinin. Cells were lysed using a sonicator and clarified by centrifugation. Supernatants were applied at 5 mL/min to 2x5 mL Ni^{2+}-charged HiTrap metal-chelating columns (Amersham Biotech), followed by a 50 mL wash with lysis buffer. Lysis buffer with 25 mM imidazole was then used to further wash the column (20 mL at 0.5 mL/min or till baseline), followed by elution of the tagged protein using lysis buffer with a gradient of 25 mM imidazole to 250 mM imidazole (over 32 mL at 0.5 mL/min). To remove the $(His)_6$-tag, samples were treated with high levels of PPX (1 mg PPX : 6 mg protein) at 4 °C overnight with gentle rocking. To remove the PPX (which was GST tagged) we used 2x 1 mL

HiTrap GST columns (Amersham Biotech) and the flowthrough collected (this was repeated 3x). The purified *Tth* σ^E was concentrated to 0.6 mg/mL by centrifugal filtration (ViaScience). Glycerol (10% final concentration) was added the purified *Tth* σ^E which was then aliquoted, flash frozen and stored at -80 °C.

References

1. Gross, C.A., et al., *The functional and regulatory roles of sigma factors in transcription.* Cold Spring Harb Symp Quant Biol, 1998. **63**: p. 141-55.

2. Darst, S.A., *Bacterial RNA polymerase.* Curr Opin Struct Biol, 2001. **11**(2): p. 155-62.

3. Murakami, K.S. and S.A. Darst, *Bacterial RNA polymerases: the wholo story.* Curr Opin Struct Biol, 2003. **13**(1): p. 31-9.

4. Zhang, G., et al., *Crystal structure of Thermus aquaticus core RNA polymerase at 3.3 A resolution.* Cell, 1999. **98**(6): p. 811-24.

5. Murakami, K.S., et al., *Structural basis of transcription initiation: an RNA polymerase holoenzyme-DNA complex.* Science, 2002. **296**(5571): p. 1285-90.

6. Murakami, K.S., S. Masuda, and S.A. Darst, *Structural basis of transcription initiation: RNA polymerase holoenzyme at 4 A resolution.* Science, 2002. **296**(5571): p. 1280-4.

7. Vassylyev, D.G., et al., *Crystal structure of a bacterial RNA polymerase holoenzyme at 2.6 A resolution.* Nature, 2002. **417**(6890): p. 712-9.

8. Vassylyev, D.G., et al., *Structural basis for transcription elongation by bacterial RNA polymerase.* Nature, 2007. **448**(7150): p. 157-62.

9. Vassylyev, D.G., et al., *Structural basis for substrate loading in bacterial RNA polymerase.* Nature, 2007. **448**(7150): p. 163-8.

10. Mittenhuber, G., *A phylogenomic study of the general stress response sigma factor sigmaB of Bacillus subtilis and its regulatory proteins.* J Mol Microbiol Biotechnol, 2002. **4**(4): p. 427-52.

11. Helmann, J.D., *The extracytoplasmic function (ECF) sigma factors.* Adv Microb Physiol, 2002. **46**: p. 47-110.

12. Lonetto, M., M. Gribskov, and C.A. Gross, *The sigma 70 family: sequence conservation and evolutionary relationships.* J Bacteriol, 1992. **174**(12): p. 3843-9.

13. Raivio, T.L. and T.J. Silhavy, *Periplasmic stress and ECF sigma factors.* Annu Rev Microbiol, 2001. **55**: p. 591-624.

14. Gruber, T.M. and C.A. Gross, *Multiple sigma subunits and the partitioning of bacterial transcription space.* Annu Rev Microbiol, 2003. **57**: p. 441-66.

15. Manganelli, R., et al., *Sigma factors and global gene regulation in Mycobacterium tuberculosis.* J Bacteriol, 2004. **186**(4): p. 895-902.

16. De Las Penas, A., L. Connolly, and C.A. Gross, *The sigmaE-mediated response to extracytoplasmic stress in Escherichia coli is transduced by RseA and RseB, two negative regulators of sigmaE.* Mol Microbiol, 1997. **24**(2): p. 373-85.

17. De Las Penas, A., L. Connolly, and C.A. Gross, *SigmaE is an essential sigma factor in Escherichia coli.* J Bacteriol, 1997. **179**(21): p. 6862-4.

18. Missiakas, D., et al., *Modulation of the Escherichia coli sigmaE (RpoE) heat-shock transcription-factor activity by the RseA, RseB and RseC proteins.* Mol Microbiol, 1997. **24**(2): p. 355-71.

19. Dartigalongue, C., D. Missiakas, and S. Raina, *Characterization of the Escherichia coli sigma E regulon.* J Biol Chem, 2001. **276**(24): p. 20866-75.

20. Mani, N. and B. Dupuy, *Regulation of toxin synthesis in Clostridium difficile by an alternative RNA polymerase sigma factor.* Proc Natl Acad Sci U S A, 2001. **98**(10): p. 5844-9.

21. Campbell, E.A., et al., *Structure of the bacterial RNA polymerase promoter specificity sigma subunit.* Mol Cell, 2002. **9**(3): p. 527-39.

22. Voskuil, M.I. and G.H. Chambliss, *The TRTGn motif stabilizes the transcription initiation open complex.* J Mol Biol, 2002. **322**(3): p. 521-32.

23. Spassky, A., et al., *Correlation between the conformation of Escherichia coli -10 hexamer sequences and promoter strength: use of orthophenanthroline cuprous complex as a structural index.* Embo J, 1988. **7**(6): p. 1871-9.

24. Korzheva, N., et al., *A structural model of transcription elongation.* Science, 2000. **289**(5479): p. 619-25.

25. Kuznedelov, K., et al., *Structure-based analysis of RNA polymerase function: the largest subunit's rudder contributes critically to elongation complex stability and is not involved in the maintenance of RNA-DNA hybrid length.* Embo J, 2002. **21**(6): p. 1369-78.

26. Naryshkina, T., K. Kuznedelov, and K. Severinov, *The role of the largest RNA polymerase subunit lid element in preventing the formation of extended RNA-DNA hybrid.* J Mol Biol, 2006. **361**(4): p. 634-43.

27. Toulokhonov, I. and R. Landick, *The role of the lid element in transcription by E. coli RNA polymerase.* J Mol Biol, 2006. **361**(4): p. 644-58.

28. Adelman, J.L., et al., *Mechanochemistry of transcription termination factor Rho.* Mol Cell, 2006. **22**(5): p. 611-21.

29. Alba, B.M. and C.A. Gross, *Regulation of the Escherichia coli sigma-dependent envelope stress response.* Mol Microbiol, 2004. **52**(3): p. 613-9.

30. Rhodius, V.A., et al., *Conserved and variable functions of the sigmaE stress response in related genomes.* PLoS Biol, 2006. **4**(1): p. e2.

31. Humphreys, S., et al., *The alternative sigma factor, sigmaE, is critically important for the virulence of Salmonella typhimurium.* Infection and Immunity, 1999. **67**: p. 1560-1568.

32. Kovacikova, G. and K. Skorupski, *The alternative sigma factor sigma(E) plays an important role in intestinal survival and virulence in Vibrio cholerae.* Infection and Immunity, 2002. **70**: p. 5355-5362.

33. Testerman, T.L., et al., *The alternative sigma factor sigmaE controls antioxidant defences required for Salmonella virulence and stationary-phase survival.* Molecular Microbiology, 2002. **43**: p. 771-782.

34. Craig, J.E., A. Nobbs, and N.J. High, *The extracytoplasmic sigma factor, final sigma(E), is required for intracellular survival of nontypeable Haemophilus influenzae in J774 macrophages.* Infection and Immunity, 2002. **70**: p. 708-715.

35. Campbell, E.A., et al., *Crystal Structure of Escherichia coli sigma(E) with the Cytoplasmic Domain of Its Anti-sigma RseA.* Mol Cell, 2003. **11**(4): p. 1067-78.

36. Gaal, T., et al., *Promoter recognition and discrimination by EsigmaS RNAP.* Molecular Microbiology, 2001. **42**: p. 939-954.

37. Crooks, G.E., et al., *WebLogo: a sequence logo generator.* Genome Res, 2004. **14**(6): p. 1188-90.

38. Wintjens, R., et al., *Contribution of cation-pi interactions to the stability of protein-DNA complexes.* J Mol Biol, 2000. **302**(2): p. 395-410.

39. Rooman, M., et al., *Cation-pi/H-bond stair motifs at protein-DNA interfaces.* J Mol Biol, 2002. **319**(1): p. 67-76.

40. Miticka, H., et al., *Identification of nucleotides critical for activity of the sigmaE-dependent rpoEp3 promoter in Salmonella enterica serovar Typhimurium.* FEMS Microbiol Lett, 2004. **238**(1): p. 227-33.

41. Lu, X.-J. and W.K. Olson, *3DNA: A software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures.* Nucleic Acids Research, 2003. **31**: p. 5108-5121.

42. Prive, G.G., et al., *Helix geometry, hydration, and G.A mismatch in a B-DNA decamer.* Science, 1987. **238**(4826): p. 498-504.

43. Nelson, H.C., et al., *The structure of an oligo(dA).oligo(dT) tract and its biological implications.* Nature, 1987. **330**(6145): p. 221-6.

44. El Hassan, M.A. and C.R. Calladine, *Propeller-twisting of base-pairs and the conformational mobility of dinucleotide steps in DNA.* Journal of Molecular Biology, 1996. **259**: p. 95-103.

45. Mack, D.R., T.K. Chiu, and R.E. Dickerson, *Intrinsic bending and deformability at the T-A step of CCTTTAAAGG: A comparative analysis of T-A and A-T steps within A-tracts.* Journal of Molecular Biology, 2001. **312**: p. 1037-1049.

46. Stefl, R., et al., *DNA A-tract bending in three dimensions: Solving the dA4T4 vs. dT4A4 conundrum.* Proceedings of the National Academy of Sciences USA, 2004. **101**: p. 1177-1182.

47. Davey, C.A., et al., *Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 a resolution.* J Mol Biol, 2002. **319**(5): p. 1097-113.

48. Moyle, H., C. Waldburger, and M.M. Susskind, *Hierarchies of base pair preferences in the P22 ant promoter.* J Bacteriol, 1991. **173**(6): p. 1944-50.

49. Dove, S.L., S.A. Darst, and A. Hochschild, *Region 4 of sigma as a target for transcription regulation.* Mol Microbiol, 2003. **48**(4): p. 863-74.

50. Jain, D., et al., *Structure of a ternary transcription activation complex.* Molecular Cell, 2004. **13**: p. 45-53.

51. Lonetto, M.A., et al., *Analysis of the Streptomyces coelicolor sigE gene reveals the existence of a subfamily of eubacterial RNA polymerase sigma factors involved in the regulation of extracytoplasmic functions.* Proc Natl Acad Sci U S A, 1994. **91**(16): p. 7573-7.

52. Missiakas, D. and S. Raina, *The extracytoplasmic function sigma factors: role and regulation.* Mol Microbiol, 1998. **28**(6): p. 1059-66.

53. Huang, X. and J.D. Helmann, *Identification of target promoters for the Bacillus subtilis sigma X factor using a consensus-directed search.* J Mol Biol, 1998. **279**(1): p. 165-73.

54. Cao, M., et al., *Defining the Bacillus subtilis sigma(W) regulon: a comparative analysis of promoter consensus search, run-off transcription/macroarray analysis (ROMA), and transcriptional profiling approaches.* J Mol Biol, 2002. **316**(3): p. 443-57.

55. Hershberger, C.D., et al., *The algT (algU) gene of Pseudomonas aeruginosa, a key regulator involved in alginate biosynthesis, encodes an alternative sigma factor (sigma E).* Proc Natl Acad Sci U S A, 1995. **92**(17): p. 7941-5.

56. Manganelli, R., et al., *The Mycobacterium tuberculosis ECF sigma factor sigmaE: role in global gene expression and survival in macrophages.* Mol Microbiol, 2001. **41**(2): p. 423-37.

57. Manganelli, R., et al., *Role of the extracytoplasmic-function sigma factor sigma(H) in Mycobacterium tuberculosis global gene expression.* Mol Microbiol, 2002. **45**(2): p. 365-74.

58. Paget, M.S., et al., *Defining the disulphide stress response in Streptomyces coelicolor A3(2): identification of the sigmaR regulon.* Mol Microbiol, 2001. **42**(4): p. 1007-20.

59. Xiao, Y. and S.W. Hutcheson, *A single promoter sequence recognized by a newly identified alternate sigma factor directs expression of pathogenicity and host range determinants in Pseudomonas syringae.* J Bacteriol, 1994. **176**(10): p. 3089-91.

60. Aggarwal, A.K., *Crystallization of DNA binding proteins with oligo deoxynucleotides.* Methods: A Companion to Methods Enzymol, 1990. **1**: p. 83-90.

61. Vagin, A. and A. Teplyakov, *MOLREP: An automated program for molecular replacement.* Journal of Applied Crystallography, 1997. **30**: p. 1022-1025.

62. Adams, P.D., et al., *Cross-validated maximum likelihood enhances crystallographic simulated annealing refinement.* Proceedings of the National Academy of Sciences USA, 1997. **94**: p. 5018-5023.

63. Jones, T.A., et al., *Improved methods for building protein models in electron denstiy maps and the location of errors in these models.* Acta crystallographica, 1991. **A47**: p. 110-119.

64. Baker, N.A., et al., *Electrostatics of nanosystems: Application to microtubules and the ribosome.* Proceedings of the National Academy of Sciences USA, 2001. **98**: p. 10037-10041.

65. Cramer, P., *Multisubunit RNA polymerases.* Curr Opin Struct Biol, 2002. **12**(1): p. 89-97.

66. *Analysis of the genome sequence of the flowering plant Arabidopsis thaliana.* Nature, 2000. **408**(6814): p. 796-815.

67. Hu, J. and L. Bogorad, *Maize chloroplast RNA polymerase: the 180-, 120-, and 38-kilodalton polypeptides are encoded in chloroplast genes.* Proc Natl Acad Sci U S A, 1990. **87**(4): p. 1531-5.

68. Hu, J., R.F. Troxler, and L. Bogorad, *Maize chloroplast RNA polymerase: the 78-kilodalton polypeptide is encoded by the plastid rpoC1 gene.* Nucleic Acids Res, 1991. **19**(12): p. 3431-4.

69. Iyer, L.M., L. Aravind, and E.V. Koonin, *Common origin of four diverse families of large eukaryotic DNA viruses.* J Virol, 2001. **75**(23): p. 11720-34.

70. Iyer, L.M., et al., *Evolutionary genomics of nucleo-cytoplasmic large DNA viruses.* Virus Res, 2006. **117**(1): p. 156-84.

71. Sweetser, D., M. Nonet, and R.A. Young, *Prokaryotic and eukaryotic RNA polymerases have homologous core subunits.* Proc Natl Acad Sci U S A, 1987. **84**(5): p. 1192-6.

72. Jokerst, R.S., et al., *Analysis of the gene encoding the largest subunit of RNA polymerase II in Drosophila.* Mol Gen Genet, 1989. **215**(2): p. 266-75.

73. Iyer, L.M., E.V. Koonin, and L. Aravind, *Evolution of bacterial RNA polymerase: implications for large-scale bacterial phylogeny, domain accretion, and horizontal gene transfer.* Gene, 2004. **335**: p. 73-88.

74. Chlenov, M., et al., *Structure and function of lineage-specific sequence insertions in the bacterial RNA polymerase beta' subunit.* J Mol Biol, 2005. **353**(1): p. 138-54.

75. Johnson, J.M., et al., *Protein family annotation in a multiple alignment viewer.* Bioinformatics, 2003. **19**(4): p. 544-5.

76. Chevenet, F., et al., *TreeDyn: towards dynamic graphics and annotations for analyses of trees.* BMC Bioinformatics, 2006. **7**: p. 439.

77. Zakharova, N., et al., *The largest subunits of RNA polymerase from gastric helicobacters are tethered.* J Biol Chem, 1998. **273**(31): p. 19371-4.

78. Zakharova, N., et al., *Fused and overlapping rpoB and rpoC genes in Helicobacters, Campylobacters, and related bacteria.* J Bacteriol, 1999. **181**(12): p. 3857-9.

79. Dailidiene, D., et al., *Urea sensitization caused by separation of Helicobacter pylori RNA polymerase beta and beta' subunits.* Helicobacter, 2007. **12**(2): p. 103-11.

80. Taillardat-Bisch, A.V., D. Raoult, and M. Drancourt, *RNA polymerase beta-subunit-based phylogeny of Ehrlichia spp., Anaplasma spp., Neorickettsia spp. and Wolbachia pipientis.* Int J Syst Evol Microbiol, 2003. **53**(Pt 2): p. 455-8.

81. Mijts, B.N. and B.K. Patel, *Random sequence analysis of genomic DNA of an anaerobic, thermophilic, halophilic bacterium, Halothermothrix orenii.* Extremophiles, 2001. **5**(1): p. 61-9.

82. Pei, J., R. Sadreyev, and N.V. Grishin, *PCMA: fast and accurate multiple sequence alignment based on profile consistency.* Bioinformatics, 2003. **19**(3): p. 427-8.

83. Edgar, R.C., *MUSCLE: multiple sequence alignment with high accuracy and high throughput.* Nucleic Acids Res, 2004. **32**(5): p. 1792-7.

84. Lockless, S.W. and R. Ranganathan, *Evolutionarily conserved pathways of energetic connectivity in protein families.* Science, 1999. **286**(5438): p. 295-9.

85. Suel, G.M., et al., *Evolutionarily conserved networks of residues mediate allosteric communication in proteins.* Nat Struct Biol, 2003. **10**(1): p. 59-69.

86. Hatley, M.E., et al., *Allosteric determinants in guanine nucleotide-binding proteins.* Proc Natl Acad Sci U S A, 2003. **100**(24): p. 14445-50.

87. Shulman, A.I., et al., *Structural determinants of allosteric ligand activation in RXR heterodimers.* Cell, 2004. **116**(3): p. 417-29.

88. Socolich, M., et al., *Evolutionary information for specifying a protein fold.* Nature, 2005. **437**(7058): p. 512-8.

89. Caffrey, D.R., et al., *Are protein-protein interfaces more conserved in sequence than the rest of the protein surface?* Protein Sci, 2004. **13**(1): p. 190-202.

90. Wang, D., et al., *Structural basis of transcription: role of the trigger loop in substrate specificity and catalysis.* Cell, 2006. **127**(5): p. 941-54.

91. Unniraman, S., R. Prakash, and V. Nagaraja, *Conserved economics of transcription termination in eubacteria.* Nucleic Acids Res, 2002. **30**(3): p. 675-84.

92. Kingsford, C.L., K. Ayanbule, and S.L. Salzberg, *Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake.* Genome Biol, 2007. **8**(2): p. R22.

93. Stefano, J.E. and J.D. Gralla, *Spacer mutations in the lac ps promoter.* Proc Natl Acad Sci U S A, 1982. **79**(4): p. 1069-72.

94. Travers, A.A., *Structure and function of E. coli promoter DNA.* CRC Crit Rev Biochem, 1987. **22**(3): p. 181-219.

95. Ayers, D.G., D.T. Auble, and P.L. deHaseth, *Promoter recognition by Escherichia coli RNA polymerase. Role of the spacer DNA in functional complex formation.* J Mol Biol, 1989. **207**(4): p. 749-56.

96. Warne, S.E. and P.L. deHaseth, *Promoter recognition by Escherichia coli RNA polymerase. Effects of single base pair deletions and insertions in the spacer DNA separating the -10 and -35 regions are dependent on spacer DNA sequence.* Biochemistry, 1993. **32**(24): p. 6134-40.

97. deHaseth, P.L. and J.D. Helmann, *Open complex formation by Escherichia coli RNA polymerase: the mechanism of polymerase-induced strand separation of double helical DNA.* Mol Microbiol, 1995. **16**(5): p. 817-24.

98. McKane, M. and G.N. Gussin, *Changes in the 17 bp spacer in the P(R) promoter of bacteriophage lambda affect steps in open complex formation that precede DNA strand separation.* J Mol Biol, 2000. **299**(2): p. 337-49.

99. Cook, V.M. and P.L. Dehaseth, *Strand opening-deficient Escherichia coli RNA polymerase facilitates investigation of closed complexes with promoter DNA: effects of DNA sequence and temperature.* J Biol Chem, 2007. **282**(29): p. 21319-26.

100. Dombroski, A.J., et al., *The sigma subunit of Escherichia coli RNA polymerase senses promoter spacing.* Proc Natl Acad Sci U S A, 1996. **93**(17): p. 8858-62.

101. Sun, L., et al., *An RNA polymerase mutant deficient in DNA melting facilitates study of activation mechanism: application to an artificial activator of transcription.* J Mol Biol, 2004. **343**(5): p. 1171-82.

102. Selvaraj, S., H. Kono, and A. Sarai, *Specificity of protein-DNA recognition revealed by structure-based potentials: symmetric/asymmetric and cognate/non-cognate binding.* J Mol Biol, 2002. **322**(5): p. 907-15.

103. Kono, H. and A. Sarai, *Structure-based prediction of DNA target sites by regulatory proteins.* Proteins, 1999. **35**(1): p. 114-31.

104. Sarai, A. and H. Kono, *Protein-DNA recognition patterns and predictions.* Annu Rev Biophys Biomol Struct, 2005. **34**: p. 379-98.

105. Lane, W.J. and S.A. Darst, *The structural basis for promoter -35 element recognition by the group IV sigma factors.* PLoS Biol, 2006. **4**(9): p. e269.

106. Malhotra, A., E. Severinova, and S.A. Darst, *Crystal structure of a sigma 70 subunit fragment from E. coli RNA polymerase.* Cell, 1996. **87**(1): p. 127-36.

107. Fenton, M.S., S.J. Lee, and J.D. Gralla, *Escherichia coli promoter opening and -10 recognition: mutational analysis of sigma70.* Embo J, 2000. **19**(5): p. 1130-7.

108. Panaghie, G., et al., *Aromatic amino acids in region 2.3 of Escherichia coli sigma 70 participate collectively in the formation of an RNA polymerase-promoter open complex.* J Mol Biol, 2000. **299**(5): p. 1217-30.

109. Sanderson, A., et al., *Substitutions in the Escherichia coli RNA polymerase sigma70 factor that affect recognition of extended -10 elements at promoters.* FEBS Lett, 2003. **544**(1-3): p. 199-205.

110. Schroeder, L.A., A.J. Choi, and P.L. DeHaseth, *The -11A of promoter DNA and two conserved amino acids in the melting region of sigma70 both directly affect the rate limiting step in formation of the stable RNA polymerase-promoter complex, but they do not necessarily interact.* Nucleic Acids Res, 2007. **35**(12): p. 4141-53.

111. Huerta, A.M. and J. Collado-Vides, *Sigma70 promoters in Escherichia coli: specific transcription in dense regions of overlapping promoter-like signals.* J Mol Biol, 2003. **333**(2): p. 261-78.

112. Henne, A., et al., *The genome sequence of the extreme thermophile Thermus thermophilus.* Nat Biotechnol, 2004. **22**(5): p. 547-53.