Skolkovo Institute of Science and Technology

FUNCTIONAL AND STRUCTURAL ANALYSIS OF A NON-CANONICAL MULTISUBUNIT RNA POLYMERASE ENCODED BY GIANT BACTERIOPHAGE AR9

Doctoral Thesis

by

MARIA SOKOLOVA

DOCTORAL PROGRAM IN LIFE SCIENCES

Supervisor
Professor Konstantin Severinov

Moscow - 2018

Abstract

Transcription, the synthesis of RNA from DNA template, is the first step of gene expression. In all cellular organisms from bacteria to humans, transcription of genomic DNA is catalyzed by evolutionarily related multisubunit RNA polymerases (RNAPs). The core of all cellular RNAPs is conserved. While catalytically active, it requires diverse accessory factors for promoter-specific transcription initiation.

During the past decade, extensive genome sequencing revealed genes coding for distant homologs of cellular RNAP catalytic subunits in phage/viral genomes. Some of these genes were shown to encode functional RNAPs, while the products of others remain uncharacterized. These partially characterized and non-characterized putative RNAPs are referred to as “non-canonical RNAPs” since they are highly diverged from multisubunit RNAPs of cellular organisms.

Some bacteriophages belonging to the giant bacteriophage group have genes coding for distant homologs of the two largest subunits of cellular RNAPs but lack genes coding for other compulsory components of canonical RNAP core enzymes. Also, giant phages genomes do not encode recognizable homologs of any known transcription initiation factors. Thus, it is likely that RNAPs of giant phages have unique properties: they may rely on alternative mechanisms of assembly of the core complex and utilize novel transcription initiation strategies.

This thesis is devoted to an investigation of a non-canonical multisubunit RNAP encoded by a giant phage AR9 infecting *Bacillus subtilis*. A distinguishing feature of AR9 phage is the presence of uracils instead of thymines in its double-stranded DNA

genome. Our work revealed that this property plays a crucial role in transcription of AR9 genes by its non-canonical multisubunit RNAP.

Purification and biochemical characterization of the AR9 RNAP are discussed in Chapter 3 of the thesis. The AR9 RNAP recognizes viral promoters in a way that is distinct from those described for all other known RNAPs. *In vitro* analysis of transcription initiation by the AR9 RNAP showed that promoter recognition depends on the presence of conserved uracils in the template strand of viral promoters. Furthermore, the AR9 RNAP is capable of promoter-specific transcription from single-stranded DNA molecules. This ability is unprecedented for any multisubunit RNAP studied to date. In addition to the catalytic subunits, the AR9 RNAP contains a phage protein which was shown to be responsible for unique transcription initiation properties of the enzyme and thus may represent a new class of transcription initiation factors.

To further elucidate the molecular mechanisms of AR9 RNAP function, X-ray crystallography and cryo-electron microscopy were employed to determine the structure of the enzyme. This work is in progress at the time of this writing but low- to medium-resolution structures of AR9 RNAP obtained thus far are discussed in Chapter 4 of the thesis.

Publications

1. **Sokolova M**, Borukhov S, Lavysh D, Artamonova T, Khodorkovskii M, Severinov K. A non-canonical multisubunit RNA polymerase encoded by the AR9 phage recognizes the template strand of its uracil-containing promoters. Nucleic Acids Res. 2017 Jun 2; 45(10):5958-5967

2. Lavysh D, **Sokolova M**, Slashcheva M, Förstner KU, Severinov K. Transcription profiling of *Bacillus subtilis* cells infected with AR9, a giant phage encoding two multisubunit RNA polymerases. MBio. 2017 Feb 14; 8(1)

3. Lavysh D, **Sokolova M**, Minakhin L, Yakunina M, Artamonova T, Kozyavkin S, Makarova KS, Koonin EV, Severinov K. The genome of AR9, a giant transducing *Bacillus* phage encoding two multisubunit RNA polymerases. Virology. 2016 May 26; 495:185-196

Conferences

1. **Sokolova M**, Borukhov S, Lavysh D, Khodorkovskii M, White M, Leiman P and Severinov K. Transcription Strategy of a Giant Bacteriophage AR9 and Characterization of Its Non-Canonical Multisubunit RNA Polymerase. Mechanism and Regulation of Prokaryotic Transcription, Federation of American Societies for Experimental Biology (FASEB), Saxtons River, VT, USA, 25-30 June 2017

2. **Sokolova M**, Borukhov S, Lavysh D, Artamonova T, Khodorkovskii M and Severinov K. Transcription Strategy of a Giant Bacteriophage AR9 and Characterization of Its Non-Canonical Multisubunit RNA Polymerase. 22nd Structural Biology Symposium, Sealy Center for Structural Biology and Molecular Biophysics Symposium, the University of Texas Medical Branch, TX, USA, 6 May 2017

3. **Sokolova M**, Borukhov S, Lavysh D, Artamonova T, Khodorkovskii M and Severinov K. A non-canonical multisubunit RNA polymerase encoded by the AR9 phage recognizes the template strand of its uracil-containing promoters. 29th RNA Polymerase Workshop, Newcastle University, UK, 6-7 April 2017

4. **Sokolova M**, Lavysh D, Borukhov S, Artamonova T, Khodorkovskii M and Severinov K. Functional analysis of AR9 bacteriophage and characterization of its non-canonical multisubunit RNA polymerase. Bacteriophages: Theoretical and Practical Aspects of Their Application in Medicine, Veterinary and Food, Moscow, Russia, 13-15 October 2016

5. **Sokolova M**, Lavysh D, Borukhov S, Artamonova T, Khodorkovskii M and Severinov K. Functional analysis of AR9 bacteriophage and characterization of its non-canonical multisubunit RNA polymerase. Interdisciplinary School and Conference "Information Technologies and Systems" ITAS, Saint-Petersburg, Russia, 25-30 September 2016

6. **Sokolova M**, Lavysh D, Borukhov S, Yakunina M, Artamonova T, Khodorkovskii M and Severinov K. Unusual RNA polymerases encoded by the PBS1 phage. 76th Harden Conference: Total Transcription, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK, 1-5 September 2014

Acknowledgements

I would like to express my sincere gratitude to my supervisor Konstantin Severinov for being such a great teacher for me, for setting an example of a true scientist, for guiding me during the whole way and for giving me many encouraging words. I would like to thank Sergei Borukhov for sharing his richest experience, for willingness to discuss science endlessly and in great detail, for finding very supporting and inspiring words for me when the most disappointing things were happening on my way. I would like to express my deep gratitude to Mikhail Khodorkovskii, a director of the center in Saint Petersburg where I spent most of the time working on the PhD thesis, for his strong support, for protecting me from any bureaucracy so widely spread in Russia, for keeping the environment in the lab very productive even in the situation of high uncertainty in our state, and for all the conversations on science and life. I would like to thank all my colleagues in Saint Petersburg for the fruitful and kind atmosphere in the lab.

I would like to express my deep gratitude to Petr Leiman, a supervisor of the UTMB lab in the USA, where I had an internship, for his willingness to invest a huge amount of his time in the challenging project on determination of the structure of the AR9 nvRNAP, for teaching me the methods of structural biology, and for his brilliant explanation of the difficult theory. I would like to thank Alec Fraser for his major work on collecting and processing the Cryo-EM data. Not less I would like to thank Michel Plattner for his important contribution in Cryo-EM experiments and for collecting the X-ray data on synchrotron at all times, even at night. I would also like to thank Mark White, a manager of the UTMB X-ray Crystallography Laboratory, for training me in crystallization experiments and also for collecting X-ray data for us.

I would like to thank Skoltech for showing me the high international standards of education and research, for giving me the opportunity to try myself in teamwork, for showing me the atmosphere of respect and collaboration between people from different fields, for giving me opportunities to go in different labs and participate in many international conferences.

I would like to thank my friends Iana Fedorova and Anna Shiriaeva for understanding of my feelings about the work and for their supporting and optimistic words. And most importantly, I would like to thank my family for understanding my enthusiasm for working a lot in order to achieve the desired goal and for their unconditional willingness to help along the way.

Table of Contents

ABSTRACT **2**

PUBLICATIONS **4**

CONFERENCES **4**

ACKNOWLEDGEMENTS **6**

TABLE OF CONTENTS **8**

LIST OF SYMBOLS, ABBREVIATIONS **10**

LIST OF FIGURES **13**

CHAPTER 1. LITERATURE REVIEW **14**

1.1 THE SUPERFAMILY OF TWO-BARREL POLYMERASES 15
1.2 DNA-DEPENDENT RNAPS OF CELLULAR ORGANISMS 17
1.2.1 Overall organization of a multisubunit RNAP core enzyme based on bacterial RNAP structure 19
1.2.2 Active center of multisubunit RNAPs 21
1.2.3 Transcription bubble maintenance and RNA displacement from the RNA-DNA hybrid during elongation 22
1.3 TRANSCRIPTION INITIATION BY BACTERIAL RNAP 26
1.3.1 Classification of bacterial σ factors 26
1.3.2 Overall organization of bacterial promoters recognized by σ factors from the σ70 family 28
1.3.3 Promoter recognition by σ factors from the σ^{70} family 29
1.4 NON-CANONICAL MULTISUBUNIT RNAPS 34
1.4.1 Phage single-subunit RNAPs related to multisubunit RNAPs 35
1.4.2 Viral multisubunit RNAPs 37
1.4.3 Multisubunit RNAPs of giant phages 39

CHAPTER 2. MATERIALS AND METHODS **42**

2.1 BACTERIOPHAGE, BACTERIAL STRAIN AND GROWTH CONDITIONS 42
2.2 PURIFICATION OF AR9 NVRNAP FROM INFECTED CELLS 42
2.3 NATIVE GEL ELECTROPHORESIS 44
2.4 DNA TEMPLATES FOR TRANSCRIPTION ASSAY 44
2.5 PRIMER EXTENSION AND SEQUENCING REACTIONS 45
2.6 *IN VITRO* TRANSCRIPTION 46
2.7 FOOTPRINTING REACTIONS 47
2.8 CLONING OF AR9 NVRNAP 48
2.9 PURIFICATION OF RECOMBINANT AR9 NVRNAP 49
2.10 CRYSTALLIZATION OF AR9 NVRNAP 50
2.11 PREPARATION OF HEAVY-ATOM DERIVATIVE CRYSTALS 51

CHAPTER 3. FUNCTIONAL CHARACTERIZATION OF AR9 NVRNAP **52**

3.1 RESULTS 52
3.1.1 Purification of a multisubunit phage RNAP from AR9 infected cells 52
3.1.2 In vitro transcription by AR9 nvRNAP 53
3.1.3 Functional analysis of AR9 late promoter consensus element 56
3.1.4 Characterization of AR9 nvRNAP-promoter complex 57
3.1.5 The nature of uracil requirement by AR9 nvRNAP 59
3.1.6 Template strand recognition by AR9 nvRNAP 61
3.1.7 Promoter specific transcription by AR9 nvRNAP from single-stranded DNA 62
3.1.8 Characterization of AR9 nvRNAP-promoter complex formed on partially single-stranded DNA 64

3.1.9 AR9 nvRNAP lacking gp226 subunit is catalytically active but unable to initiate transcription from promoters *66*
3.1.10 RNA transcript displacement from RNA-DNA hybrid during transcription of ssDNA *68*
3.2 DISCUSSION 69

CHAPTER 4. DETERMINATION OF THE AR9 NVRNAP STRUCTURE **75**

4.1 RESULTS 75
4.1.1 Recombinant AR9 nvRNAP *76*
4.1.2 Crystallization of the AR9 nvRNAP core enzyme *77*
4.1.3 Crystallization of AR9 nvRNAP core enzyme without the histidine tag *82*
4.1.4 Phase problem solution for crystals of tagless AR9 nvRNAP core *83*
4.1.6 Crystallization of AR9 nvRNAP holoenzyme in complex with promoter DNA *85*
4.1.7 Cryo electron microscopy with AR9 nvRNAP *87*
4.2 DISCUSSION 88

CHAPTER 5. CONCLUSIONS **91**

APPENDIX A **93**

APPENDIX B **102**

APPENDIX C **104**

APPENDIX D **107**

APPENDIX E **110**

APPENDIX F **111**

BIBLIOGRAPHY **113**

List of Symbols, Abbreviations

A – adenine

ATP – adenosine triphosphate

B. subtilis – bacteria *Bacillus subtilis*

BH – bridge helix

bp – base pairs

C – cytosine

Cryo-EM – cryo-electron microscopy

CTP – cytosine triphosphate

DNA – deoxyribonucleic acid

dNTPs – deoxynucleotides

dNTPs – deoxynucleotides

DPBB – double-psi β-barrel

dsDNA – double-stranded DNA

E. coli – bacteria Escherichia coli

ECF – Extra Cytoplasmic Function

G – guanine

GTP – guanosine triphosphate

IPTG – isopropyl β-D-1-thiogalactopyranoside

kDa – kiloDalton

LB – Luria-Bertani broth

LEF – late expression factor

LUCA – last universal common ancestor

MOI – multiplicity of infection

MR – molecular replacement

mRNA – messenger RNA

NCLDV – nucleocytoplasmic DNA viruses

NCS – noncrystallographic symmetry

nt – nucleotides

NTP – nucleoside triphosphate

nvRNAP – non-virion RNAP

OD – optical density

ORF – open reading frame

PAAG – polyacrylamide gel

PAGE - polyacrylamide gel electrophoresis

PCR – polymerase chain reaction

PEG – polyethylene glycol

PEI – polyethyleneimine

PFU – plaque forming unit

RNA – ribonucleic acid

RNAP – RNA polymerase

RNase H – ribonuclease H

RO – run off

RT – reverse transcriptase

S. shibatae – archaea *Sulfolobus shibatae*

SAD (MAD) – single (multiple) anomalous dispersion

SDS – sodium dodecyl sulfate

SIR (MIR) – single (multiple) isomorphous replacement

ssDNA – single-stranded DNA

T – thymine

T. aquaticus – bacteria *Thermus aquaticus*

T. thermophilus – bacteria *Thermus thermophilus*

TEC – transcription elongation complex

TL – trigger loop

TSS – transcription start site

TTP – thymidine triphosphate

U – uracil

UP – upstream

UTP – uridine triphosphate

vRNAP – virion RNAP

w/v – weight/volume

List of Figures

Figure 1. The catalytic center of Thermus thermophilus DNA-dependent RNAP. 16
Figure 2. Multisubunit RNAPs of organisms from three domains of life. 18
Figure 3. Structural overview of bacterial RNAP core. 19
Figure 4. Domain architecture of σ factors from the σ70 family. 28
Figure 5. Promoter motifs recognized by primary σ factors. 29
Figure 6. The model of E. coli σ^{70}-RNAP holoenzyme bound with the promoter DNA. 30
Figure 7. Binding of the conserved nucleotides of the -10 element by primary and ECF σ factors. 33
Figure 8. Transcription strategy and promoters of the AR9 phage. 41
Figure 9. Purification of nvRNAP from AR9 infected Bacillus subtilis cells. 52
Figure 10. Analysis of AR9 nvRNAP transcriptional activity. 54
Figure 11. In vitro transcription by AR9 nvRNAP from late AR9 promoters. 55
Figure 12. Primer extension analysis of in vitro transcripts synthesized from templates containing late AR9 promoters. 56
Figure 13. Late promoter consensus analysis. 57
Figure 14. Promoter binding and promoter opening by AR9 nvRNAP. 58
Figure 15. In vitro run-off transcription by AR9 nvRNAP of double-stranded P007 promoter templates carrying uracils and thymines at different positions. 60
Figure 16. Analysis of the strand requirement for promoter recognition by AR9 nvRNAP. 62
Figure 17. Specific transcription initiation by AR9 nvRNAP using single-stranded promoter DNA templates. 63
Figure 18. Late promoter consensus analysis in transcription from ssDNA. 64
Figure 19. $KMnO_4$ probing of the AR9 nvRNAP-promoter complex formed on a fork DNA template. 65
Figure 20. Functional analysis of the two forms of AR9 nvRNAP. 67
Figure 21. RNA displacement by AR9 nvRNAP. 69
Figure 22. Recombinant AR9 nvRNAP. 77
Figure 23. Crystals of the AR9 nvRNAP core. 78
Figure 24. A fragment of the electron density map of AR9 nvRNAP core. 80
Figure 25. Model of the AR9 nvRNAP core. 81
Figure 26. Crystals of tagless AR9 nvRNAP core. 83
Figure 27. A fragment of the improved electron density map of AR9 nvRNAP core. 84
Figure 28. Improved model of the AR9 nvRNAP core. 85
Figure 29. DNA template used to prepare AR9 nvRNAP holoenzyme/promoter complex prior to crystallization. 86
Figure 30. Crystals of the AR9 nvRNAP holoenzyme/promoter complex. 86
Figure 31. Comparison of the AR9 nvRNAP core model with T. aquaticus RNAP core structure. 89

Chapter 1. Literature Review

Template-dependent nucleic acids polymerases are essential and ancient enzymes required for maintenance, transfer, and expression of genetic information. There are two most common, yet evolutionarily unrelated superfamilies of nucleic acids polymerases: those that share a right-hand-shaped fold [1] and those that contain two double-psi β-barrel domains [2-4]. These superfamilies are further divided into several families and subfamilies depending on enzymatic functions and template and substrate specificities. The enzymes of transcription – DNA-depended RNA polymerases (RNAPs) – are present in both major superfamilies: 'Right-handed' RNAPs are single-subunit enzymes transcribing genes of mitochondria, chloroplasts, and some bacteriophages [5], while 'two-barrel' RNAPs are mostly multisubunit enzymes, transcribing genes of cellular organisms [6].

In recent years, several two-barrel RNAPs were discovered to be encoded in phage/viral genomes [2, 7-9]. They have unique features in terms of structure and function and thus are often referred to as "non-canonical RNAPs" [9]. Investigation of these enzymes may shed light on the evolution of two-barrel polymerases and through comparative analysis will contribute to the understanding of the transcription process in greater detail.

This Chapter starts with description of the two-barrel polymerase superfamily and further zooms in into the structure and function of canonical multisubunit RNAPs. Subsequently, mechanisms of transcription initiation by the simplest multisubunit RNAPs are addressed. The discussion proceeds with an overview of recently characterized non-

canonical transcription enzymes and ends up with information on highly unusual phage-encoded multisubunit RNAPs, one of which is the object of the current investigation.

Multisubunit RNAPs are complex molecular machines whose many elements have multiple modes of action. Not all multisubunit RNAP structural and functional features are described at the same level of detail in this Chapter. Properties of canonical multisubunit RNAPs that are most relevant for discussion of the thesis results are addressed more thoroughly.

1.1 The superfamily of two-barrel polymerases

The double-psi β-barrel domain (DPBB) is formed by two β-sheets which have an arrangement that resembles two Greek letters psi (ψ) with three β-strands constituting each letter [10, 11]. A catalytic center of two-barrel polymerases is wedged between the two DPBB domains, which are located in a head-to-tail manner with respect to each other [3, 11]. In canonical multisubunit DNA-dependent RNAPs, the two barrels are parts of two distinct subunits (β and β' in bacterial RNAP (Fig. 1)). Each DPBB contributes to the active site distinct amino acid residues coordinating two metal ions (usually, magnesium), which catalyze the formation of phosphodiester bond between the α phosphate of the incoming nucleoside triphosphate (NTP) substrate and the 3' hydroxyl of the nucleotide at the growing end of RNA, with elimination of pyrophosphate [12, 13]. The first Mg^{2+} (Mg-I) is stably bound by three aspartate residues of the invariant metal binding motif DbDGD (where b is a bulky amino acid) which forms a catalytic loop of the enzyme [13]. The second Mg^{2+} (Mg-II) is not stably bound by RNAP in the absence of substrate because its co-ordination requires phosphate residues of the entering NTP [13].

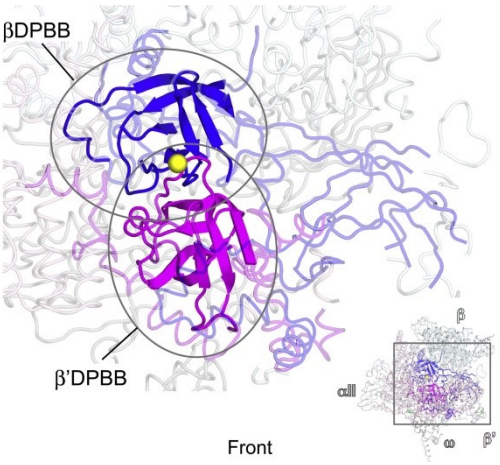


Figure 1. The catalytic center of Thermus thermophilus DNA-dependent RNAP.

The DPBB domains belonging to β and β' subunits are colored in blue and purple, correspondingly. Magnesium ion (Mg-II) coordinated by conserved aspartate residues of the β' DPBB is colored in yellow. The figure is taken from [14], with permission.

Initially, the ubiquity of the DPBB architecture was thought to be restricted to DNA-dependent RNAPs only [2, 3]. However, a similar organization of catalytic center was also found for Qde-1 enzyme – a eukaryotic RNA-dependent RNAP involved in RNA silencing in *Neurospora crassa* [15]. In Qde-1, two DPBBs have the same conserved amino acid residues as DPBBs of DNA-dependent RNAPs, but are located within a single polypeptide chain.

Sauguet and co-workers have recently reported a high-resolution crystal structure of archaeal replicative DNA polymerase D (PolD) [4]. Surprisingly, in contrast to all other known DNA-dependent DNA polymerases (which belong either to the right-handed polymerase superfamily or to a distinct Polβ-like DNA polymerase superfamily [16]), PolD is a member of the two-barrel polymerase superfamily. The large PolD catalytic subunit DP2 contains two DPBB domains which are well-superimposed with the catalytic centers of RNAPs mentioned above; however, only two out of three mandatory aspartic residues are present in DP2 [4]. Although there is no further structural similarity beyond

the two DPBB domains, the crystal structure of PolD bridges together DNA transcription and replication enzymes within a single protein superfamily, possibly underscoring antiquity of two-barrel polymerases [4].

A hypothetical evolutionary scenario was proposed, where an ancestor of nucleic acid polymerases consisted of a homodimer of a single-DPBB domain protein and had no catalytic activity but served as a scaffold for a ribozyme [2, 3, 11]. Duplication of the gene coding for a DPBB domain followed by divergent evolution of the duplicated genes probably gave rise to the two DPBBs forming a heterodimer, which eventually displaced the ribozyme when the critical residues required for protein-based polymerase activity appeared [2, 3, 11]. This development of DPBBs must have happened during the RNA-protein era, where the primordial two-barrel polymerase functioned as an RNA-dependent RNA polymerase. It is envisioned that later in evolution, this enzyme acquired DNA binding ability giving birth to DNA-dependent RNAPs [2].

1.2 DNA-dependent RNAPs of cellular organisms

In all cellular organisms, transcription of cellular genome is driven by two-barrel multisubunit DNA-dependent RNAPs [6]. In bacteria and archaea, a single enzyme is employed for transcription of all genes, while in eukarya there are at least three specialized RNAPs dedicated to different subsets of nuclear genes [17]. Archaeal RNAP is remarkably similar to eukaryal RNAP II which synthesizes messenger RNAs [6].

The first crystal structure of multisubunit RNAP was obtained for *Thermus aquaticus* RNAP core enzyme in 1999 [18]. One year later a crystal structure of eukaryotic RNAP II core from *Saccharomyces cerevisiae* was reported [19] by the group of Roger Kornberg who was awarded the Nobel Prize in 2006 for his “studies of the molecular basis of eukaryotic transcription”. The first crystal structure of archaeal RNAP

was determined in 2008 [20]. At the time of this writing, dozens of multisubunit RNAP structures in complex with ligands, inhibitors, regulatory protein factors, and nucleic acid scaffolds are available, making possible extensive structure-function analyses of multisubunit RNAPs from all domains of life.

All cellular RNAP core enzymes look similar and resemble a 'crab claw' with two 'pincers' formed by the two largest RNAP subunits joined at the base by the assembly platform (Fig. 2). The core of the simplest cellular RNAP encoded in most eubacterial genomes has an $\alpha_2\beta\beta'\omega$ subunit composition, while archaeal and eukaryal RNAPs contain up to twelve additional subunits [17]. Conserved DPBB domains of multisubunit RNAPs are located within the two largest subunits - β and β' in bacteria. A homodimer of bacterial α subunits or a heterodimer of archaeal/eukaryal homologs is necessary for assembly of the two largest subunits [14, 21]. The smallest ω subunit and its homologs promote the assembly of RNAP complex and stabilize it [22].

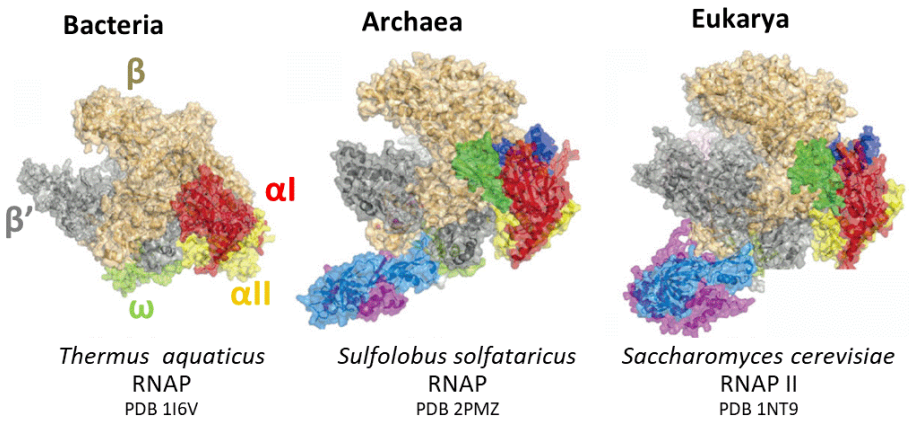


Figure 2. Multisubunit RNAPs of organisms from three domains of life.

The X-ray structures of multisubunit RNAPs from three domains of life are shown (species names and PDB IDs are indicated below the structures). Subunits of bacterial RNAP core are indicated. Homologous subunits are shown in the same color. The figure adapted from [23], with permission.

RNAP synthesizes RNA in a template-dependent manner through a transcription cycle that can be subdivided into three stages: initiation, elongation, and termination of transcription. RNAP core enzyme is catalytically active and operates during the elongation and termination stages but it requires accessory factors for promoter-specific transcription initiation [17].

1.2.1 Overall organization of a multisubunit RNAP core enzyme based on bacterial RNAP structure

The β and β’ subunits of bacterial RNAP form a wide cleft with the catalytic loop positioned at its inner back side (Fig. 3) [18]. Multiple elements required for different aspects of the RNAP function protrude into the cleft [18]. The larger modules and domains of β and β’ subunits form channels which accommodate nucleic acids within the RNAP.

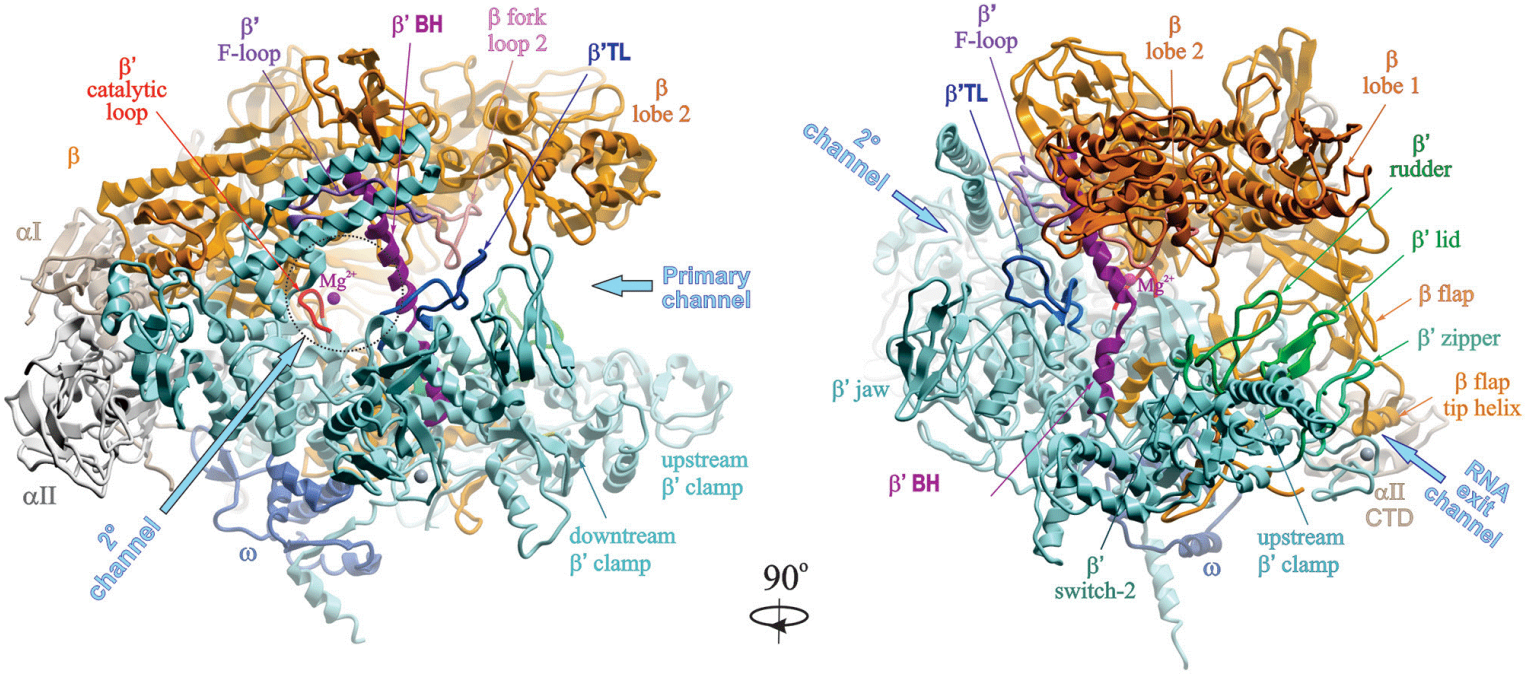


Figure 3. Structural overview of bacterial RNAP core.

Structure of T. aquaticus RNAP core [18]. The β and β’ subunit are colored in orange and blue, correspondingly. The dimer of α subunits is in shadows of gray. The ω subunit is dark blue. Three major RNAP channels are indicated by bold arrows. Conserved

elements (β' catalytic loop, β' trigger loop (TL), β' bridge helix (BH), β' F-loop, β fork loop 2, β' rudder, β' lid, β' zipper, β' switch-2) and major modules (β lobe 1, β lobe 2, β' jaw, β' clamp, β flap) are indicated by thin arrows. The figure is taken from [24], with permission.

In the transcription elongation complex (TEC), the double-stranded DNA molecule is melted around the RNAP active site and the nascent RNA transcript is partially annealed to the template DNA strand [25-27]. The region of untranscribed DNA lying ahead of the RNAP catalytic center (in the direction of the RNAP translocation) is referred to as downstream DNA, while the opposite region is referred to as upstream DNA. The region of melted DNA between the downstream and upstream DNA duplexes is referred to as 'transcription bubble'.

In bacterial TEC, the downstream DNA duplex is located in a 'primary channel' formed by β lobe 2, β' clamp and β' jaw, while the upstream DNA duplex leaves RNAP between β lobe 1 and β' clamp [26, 27]. The nascent RNA transcript is base-paired with the template DNA strand within the transcription bubble. In active TECs, the growing 3' end of the RNA molecule is located in the RNAP active center. The hybrid continues for 8-9 base pairs [28], and the RNA separates from the DNA near the upstream edge of the transcription bubble. The 5' end of the nascent RNA transcript leaves the complex through the 'exit channel' – a narrow cleft formed by the β switch 3, β' clamp and flexible β flap domains [26, 29]. The non-template DNA strand of the transcription bubble is partially accommodated between the two β lobes [29].

The only channel which is not occupied by nucleic acids within the cleft of the elongating RNAP is the 'secondary channel' which is orthogonal to the primary channel, opens on the RNAP surface and directly leads to the active site. This channel serves as a path for NTP substrates [18, 30, 31]. The secondary channel is also used by a variety of

regulatory factors (GreA/B, DksA in bacteria and TFIIS in archaea/eukarya) which directly reach the RNAP catalytic center and modulate its activity [32-34]. During the elongation stage, RNAP may temporarily pause due to different reasons or assume an inactive dead-end state. In the dead-end complex, so-called backtracked configuration is adopted where the 3' end of the nascent RNA is extruded into the secondary channel. Such complexes are rescued by transcript cleavage factors (GreA/B and TFIIS) [32, 34].

Multisubunit RNAPs have a high level of conformational plasticity that is essential for their function. RNAP modules and domains were observed in different conformational states both in crystals and solution [18, 25, 26, 35-37]. The most investigated large conformational change is the opening and closing of the primary channel through a motion of the RNAP β' clamp domain which can swing around a hinge region located at the base of the clamp [36, 38]. The closed conformation is observed in structures of elongation complexes while the open-clamp conformation exists primarily in free RNAP [36]. The closed conformation is stabilized by nonspecific contacts between downstream DNA duplex and the clamp domain and is required for catalysis [36].

1.2.2 Active center of multisubunit RNAPs

In addition to conserved DPBB residues, several other crucial elements constitute the active site of multisubunit RNAPs, among them the most flexible structure of multisubunit RNAPs – the trigger loop. It consists of two base trigger loop α-helices connected by a flexible trigger loop tip. In the first RNAP structure, of *T. aquiticus* core enzyme, only base trigger loop helices were seen while the tip was unstructured [18]. It became apparent later that the trigger loop may adopt a fully folded conformation, where the base α-helices are extended and form a helix-turn-helix fold [13, 39]. Such

conformation is usually observed in TECs [13, 39]. The trigger loop α-helices are in close proximity with another α-helix which is called the bridge helix since it traverses through the main RNAP cleft connecting the pincers of the claw. The bridge helix was also observed in two distinct conformations: kinked and straight [40]. Elements of trigger loop and bridge helix form a metastable three-helix bundle which concertedly changes its conformation during the nucleotide addition cycle [14]. The bridge helix/trigger loop module has many absolutely conserved residues directly interacting with the NTP substrate and the RNA-DNA hybrid and enabling the RNAP to translocate along the DNA template. The module functions in a cyclic manner. Two prominent states may be distinguished, though many intermediate conformations are possible (especially during paused states). Once a correct NTP binds to the DNA template in the active center, the tip of the trigger loop adopts folded conformation blocking the secondary channel [41]. After the release of the pyrophosphate, a motion of the whole unit leads to a transfer from the pre-translocated to post-translocated RNAP state [41]. Then the trigger loop again adopts the unfolded conformation opening the way for a new NTP.

F-loop is another element located in proximity with the bridge helix. It influences the coordinated motion of the bridge helix/trigger loop module during the nucleotide addition cycle [42]. It was suggested that the F-loop interacts directly with the tip of the folded trigger loop likely stabilizing its closed conformation and thus facilitating trigger loop transition between unfolded and folded conformations [26, 42].

1.2.3 Transcription bubble maintenance and RNA displacement from the RNA-DNA hybrid during elongation

During the elongation phase, an 8-9 base-pair RNA-DNA hybrid extends from the RNAP active center to the exit channel [28]. The maintenance of RNA-DNA hybrid of

such length is highly important since it was shown that overextended (>9 bp) or shortened (<7 bp) hybrids dramatically compromise stability of the elongation complex [43, 44]. Analysis of structures and models of elongation complexes led to suggestions that several loop-like elements protruding into the main cleft are involved in the maintenance of the proper RNA-DNA hybrid length and precise borders of the transcription bubble [18, 19, 25, 27, 29, 45-47]. These elements include β' rudder and β' lid arising from the mobile clamp domain, and β fork1/β fork2 emerging from the opposite side of the cleft (β fork1 is absent in bacterial RNAP).

Structures of TECs initially were determined for complexes obtained with RNA/DNA scaffolds composed of downstream DNA duplex and preassembled RNA-DNA hybrid but lacking the non-template DNA strand of the transcription bubble and upstream DNA duplex [25, 26]. In 2015 Barnes and co-workers determined a structure of yeast RNAP II TEC containing the complete nucleic acid scaffold [48]. A structure of the bacterial TEC with complete nucleic acid scaffold is not yet determined. While the main cleft loops generally share common functions among multisubunit RNAPs from all domains of life, some features vary. Effects of deletions of the cleft loops were assessed *in vitro* for bacterial and archaeal RNAPs and are discussed below [45-47, 49].

In the crystal structure of bacterial TEC, the β fork2 element sterically blocks downstream DNA duplex from entering the active site [26]. In RNAP II TECs containing the complete RNA/DNA scaffold, DNA is melted around the β fork2 element and amino acid residues of the β fork2 interact with the non-template DNA strand [48]. Thus, the β fork2 element likely plays a crucial role in downstream DNA strand separation and maintenance of the downstream edge of the transcription bubble [14, 48].

The rudder is another loop-like element of the cleft. It comprises two AT hook-like modules [3, 14]. In the bacterial RNAP TEC structures, the rudder is positioned right between the downstream DNA duplex and the upstream edge of the RNA/DNA hybrid interacting with both nucleic acid elements directly through its conserved arginine residues [14, 26]. In the RNAP II TEC, the rudder together with the β fork1 forms an 'arc', which is located between the template and non-template DNA strands and physically marks the upstream boundary of the transcription bubble [48]. The conserved lysine residue of the rudder in RNAP II interacts with the non-template DNA strand just before the point of its re-annealing with the template strand [48]. Thus the rudder in all multisubunit RNAPs likely stabilizes the transcription bubble architecture. For both bacterial and archaeal RNAPs the rudder was shown to be necessary for open complex formation during transcription initiation stage [45, 49]. It also contributes greatly to transcription elongation complex stability but is not directly involved in RNA transcript displacement [45, 49].

In both bacterial and yeast TECs, the lid element emerges on the way of the RNA transcript at the upstream edge of the RNA-DNA hybrid and forms stacking interactions with the last base pair of the hybrid [26, 48]. It was suggested that the lid serves as a physical barrier preventing extended RNA-DNA hybrid formation and thus plays an active role in RNA displacement [46, 47]. This hypothesis was not confirmed experimentally since deletion of the lid element in bacterial RNAP did not result in extended RNA-DNA hybrid formation during elongation: removal of the lid strongly affected transcription initiation by bacterial RNAP but barely influenced the elongation stage [46, 47].

The most remarkable effect of the lid deletion was observed on transcription by bacterial RNAP from single-stranded DNA (ssDNA) templates [46, 47]. In such experiments, RNA primers are used to initiate transcription, since multisubunit RNAPs are unable to recognize promoters in ssDNA [46, 47]. Two groups independently showed that during transcription from ssDNA bacterial RNAP lacking the lid domain produces an extended RNA-DNA hybrid corresponding to full length of the DNA template [46, 47]. In the same experiment wild-type RNAP also produces an extended RNA-DNA hybrid but stalls after incorporation of about 20 nucleotides [46, 47]. The inability to separate a nascent RNA transcript from ssDNA is therefore attributed to the absence of the non-template DNA strand in both wild-type RNAP and mutant RNAP lacking the lid, whereas the lid element restricts the length of the extended RNA-DNA hybrid during transcription from ssDNA [46, 47].

It was proposed that in the case of transcription from ssDNA by bacterial wild-type RNAP, an extended RNA-DNA hybrid clashes with the lid domain and the RNAP eventually backtracks due to accumulated tension [46, 47]. It was further suggested that in this case the overextended RNA-DNA hybrid may be pulled into the primary channel which is normally occupied by the downstream DNA duplex [46, 47]. Such re-arrangement is not compatible the catalysis and therefore arrests transcription on ssDNA. In the absence of the lid domain, it is likely that an alternative exit pathway opens within the bacterial TEC allowing continuous RNA-DNA hybrid to exit the RNAP complex [46, 47].

Archaeal RNAP also cannot separate RNA from the RNA-DNA hybrid when transcribing ssDNA, supporting the universal importance of the non-template DNA strand in transcription elongation [49]. However, surprisingly, archaeal RNAP is able to

synthesize long RNA-DNA hybrids (>40 bp) regardless of the presence of the lid, pointing out to significant differences between bacterial and archaeal RNAPs in this regard [49].

Overall, it may be assumed that loop-like elements of RNAP cleft described in this section stabilize the architecture of nucleic acids in elongation complexes but are not directly involved in RNA displacement, which depends on the presence of the non-template DNA strand in the TEC.

1.3 Transcription initiation by bacterial RNAP

While the elongation of transcription by multisubunit RNAPs from different domains of life is highly conserved in terms of structure-function relationships, the initiation stage is highly divergent. For all RNAPs, initiation includes finding a promoter sequence in DNA and melting the double-stranded DNA around the transcription start site. For this purpose, bacteria employ one of several σ factors each of which binds the RNAP core forming a holoenzyme able to recognize promoters with different consensus elements [50]. Archaeal and eukaryal RNAPs use a complex set of general and specific transcription factors, which are evolutionarily unrelated to bacterial σ factors [51, 52]. Archaeal and eukaryal transcription initiation factors first bind promoter DNA and then recruit the RNAP core [51, 52].

Transcription initiation by bacterial RNAP is outlined below.

1.3.1 Classification of bacterial σ factors

The first bacterial σ factor was discovered fifty years ago by Richard R. Burgess and, independently, by Ekkehard K. F. Bautz. A short but meaningful abstract of the Burgess paper in *Nature* stated: “A protein component usually associated with RNA polymerase can be separated from the enzyme by chromatography on phosphocellulose.

The polymerase is unable to transcribe T4 DNA unless this factor is added back" [53]. That *Escherichia coli* factor had a molecular weight of about 70 kDa and was later called σ^{70} factor/subunit.

Now, it is known that bacteria have multiple σ factors, up to several dozens in some species [54, 55]. Different σ factors bind the RNAP core and direct it to promoters of different genes orchestrating gene expression in bacteria, which is not a rigid program but rather a complex and changeable network dependent on numerous conditions (various kinds of stress, different stages of growth, and so on).

All σ factors are classified into two evolutionarily unrelated families based on their homology with either σ^{70} or σ^{54} factors from *E. coli* [50]. Most σ factors belong to the former family. The most significant difference between factors from the two families is that RNAP holoenzymes containing σ^{70}-family proteins melt promoter DNA without the input of ATP-hydrolysis while σ^{54}-RNAP holoenzymes require an additional ATP-dependent activator protein to accomplish DNA melting [56]. Since σ^{54}-like factors are a minor family, below we concentrate on σ^{70}-family proteins only.

There are four groups of σ factors within the σ^{70} family [50, 57]. The *E. coli* σ^{70} belongs to Group 1. All σ factors of this group are primary or so-called housekeeping σ factors since they are needed for transcription of a vast majority of bacterial genes, while σ factors of the other three groups are called alternative σ factors: they direct RNAP for transcription of genes required at specific conditions and some of these proteins are non-essential [50, 57]. Group 4 is also known as Extra Cytoplasmic Function (ECF) group since many of its members sense signals generated outside the cell and direct transcription in response to these signals [50, 57].

Members of the σ^{70} family are modular proteins with up to four conserved domains ($\sigma_{1.1}$, σ_2, σ_3, σ_4) connected by linkers (Fig. 4). All four domains are present in σ factors from Group 1, three domains are present in σ factors from Groups 2 and 3 (σ_2, σ_3, σ_4), while σ factors from Group 4 are consist of only two domains (σ_2, σ_4). Each of the σ domains, except the $\sigma_{1.1}$, recognizes a certain promoter element. The $\sigma_{1.1}$ domain was shown to occupy primary channel of the RNAP holoenzyme, inhibiting nonspecific DNA binding [58]. Upon promoter recognition, the $\sigma_{1.1}$ is displaced from the primary channel by the downstream DNA duplex [58].

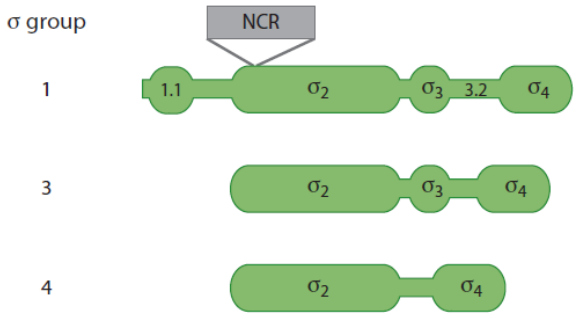


Figure 4. Domain architecture of σ factors from the σ70 family.

Group 1 (primary), 3, and 4 σ factors are shown as green bars with domains and regions labeled with their numerical names. NCR – non-conserved region. The figure is taken from [57], with permission.

1.3.2 Overall organization of bacterial promoters recognized by σ factors from the σ70 family

There are several elements of DNA that may constitute bacterial promoters recognized by RNAP holoenzymes containing σ factors of the σ^{70} family: the -35 and -10 promoter consensus elements (located near the -35 and -10 positions with respect to the transcription start site located at +1, correspondingly), the extended -10 motif, and the

discriminator elements [59] (Fig. 5). In addition to these elements recognized by σ factors in the context of RNAP holoenzyme, some promoters contain the upstream (UP) element, which is recognized by the C-terminal domains of the RNAP α subunits [60, 61]. The -10 element is present in all promoters recognized by σ^{70}-like factors since it plays a crucial role in DNA melting [62]. When appropriately located, either the -35 element or the extended -10 element motif is sufficient for RNAP recognition and binding to DNA sequence containing the -10 element. The UP elements can strongly stimulate recognition of promoters with weak (i.e., poorly matching with consensus) -35 and -10 elements [63-65]. Different combinations of conserved elements with varying degree of matches to their respective consensi create variations of individual promoters' strength necessary for achieving differential levels of basal gene expression [66].

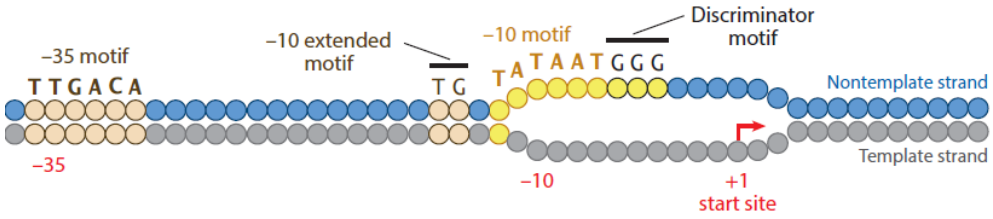


Figure 5. Promoter motifs recognized by primary σ factors.

Non-template and template DNA strands are shown as blue and gray circles, correspondingly. Consensus sequences for each promoter element are shown above. The transcription start site is indicated by "+1". A red arrow indicates direction of the transcription. The figure was taken from [57], with permission.

1.3.3 Promoter recognition by σ factors from the σ^{70} family

Free σ factors of the σ^{70} family either do not bind their respective promoters or bind them very poorly [57, 67, 68]. It is commonly thought that the RNAP core serves as a scaffold that positions the DNA binding domains of a σ such that their simultaneous interactions with all promoter elements is possible (Fig. 6) [57, 69, 70].

The majority of structural studies of the σ^{70} family factors were performed with primary σ factors from *T. aquaticus*, *T. thermophilus*, and *E. coli*, while much less is known about the structures of alternative σ factors. Bellow, all observations and conclusions are given for primary σ factors, unless stated otherwise.

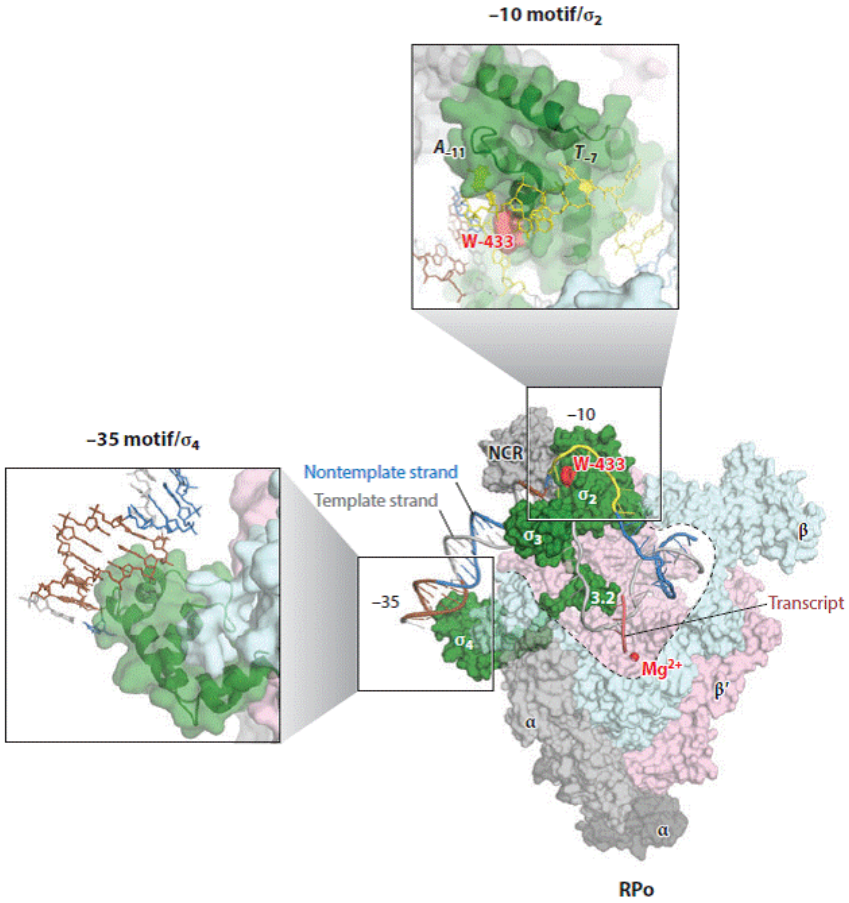


Figure 6. The model of E. coli σ^{70}-RNAP holoenzyme bound with the promoter DNA.

The molecular surface of the RNAP holoenzyme is shown: gray, α-subunits; light cyan, β; light pink, β; green, σ [the non-conserved region of the σ (NCR) is shown in gray]. A part of the β-subunit is omitted to show the active site channel. The non-template and template DNA strands are blue and gray, correspondingly, with the -35 motif shown in brown, and the -10 motif in yellow. A nascent RNA transcript is pink. The catalytic Mg^2+ ion is shown as a red sphere. The close-up views on the left and at the top show the -35 motif/σ_4 and -10 motif/σ_2 interactions, correspondingly. The figure was taken from [57], with permission.

The σ_4 domain interacts with the -35 element through a conserved helix-turn-helix motif [71]. Domain σ_3 contains an α-helix recognizing the major DNA groove of the -10 extended element [72]. Both of these promoter elements are recognized in double-stranded form, while the -10 and discriminator motifs are bound by the σ_2 domain as single-stranded DNA [59, 62].

The common view on promoter recognition is that the RNAP holoenzyme first localizes promoter in DNA through recognition of the UP, -35, or -10 extended elements, forming a closed complex (where DNA strands are not melted yet) [59]. Isomerization of this complex leads to the open complex, where the σ_2 domain binds the non-template DNA strand of the -10 element (it is a speculative and debated question whether σ promotes melting actively or just stabilizes a premelted state) [73].

Long before the first crystal structures directly showed how σ_2 recognizes the -10 element, biochemical experiments revealed the existence of promoter complexes formed by mutant RNAP holoenzymes with transcription bubble shortened in the downstream direction [45, 74]. These observations led to thinking that such complexes could reflect an intermediate state that also occurs during normal promoter melting. Experiments with fluorescent probes showed that opening of promoter DNA indeed occurs first at the upstream edge of the -10 promoter element and then propagates downstream to include the transcription start site [75].

Structural studies revealed extensive interactions between σ_2 and phosphate backbone of every nucleotide of the non-template DNA strand of the -10 promoter element [62, 76]. The selectivity of recognition is ensured through nucleotide-specific binding to most conserved A^{-11} and T^{-7} nucleotides (consensus sequence of the -10 promoter element for primary σ factors is $T^{-12}ATAAT^{-7}$) [62, 76]. These nucleotides are

flipped out of the base stack and accommodated in tight nucleotide binding pockets of the σ_2 domain [62, 76]. Early functional investigations showed that multiple aromatic residues of the σ_2 domain are crucial for melting of promoter DNA by the holoenzyme [77-79]. Later, these residues were shown to constitute the nucleotide binding pockets of the σ_2 domain and participate in the maintenance of the upstream transcription bubble boundary in promoter complexes [62, 76]. For example, as seen in crystal structures of RNAP holoenzymes in complex with promoters, one of the conserved tryptophan residues is positioned in place of the flipped A^{-11} and forms stacking interactions with the -12 base pair, thus stabilizing the junction between the double-stranded upstream DNA and single-stranded DNA within the transcription bubble [76].

It was shown that at least some ECF σ factors, for example σ^E from *E. coli*, to some extent share the paradigm of transcription bubble stabilization through trapping of flipped nucleotides from the non-template DNA strand of their -10 promoter consensus element [80] (Fig. 7). The structure of the σ_2 domain of σ^E bound to a DNA oligo corresponding to the non-template DNA strand of cognate -10 element (consensus sequence is $G^{-12}TC^{-10}$) shows that strictly conserved nucleotide C^{-10} is flipped out of the base stack and is accommodated by the σ_2 specificity loop, which forms a cage-like structure around the flipped base [80]. Substitutions of amino acid residues in this loop alter promoter specificity of the σ^E holoenzyme in *in vitro* [80]. Thus, despite the differences in consensus sequences of the -10 elements, it is possible that the capturing of the non-template DNA strand of promoters through the binding of flipped conserved nucleotides is a universal way through which σ factors of the σ^{70} family ensure promoter melting and stabilize transcription bubble during transcription initiation stage.

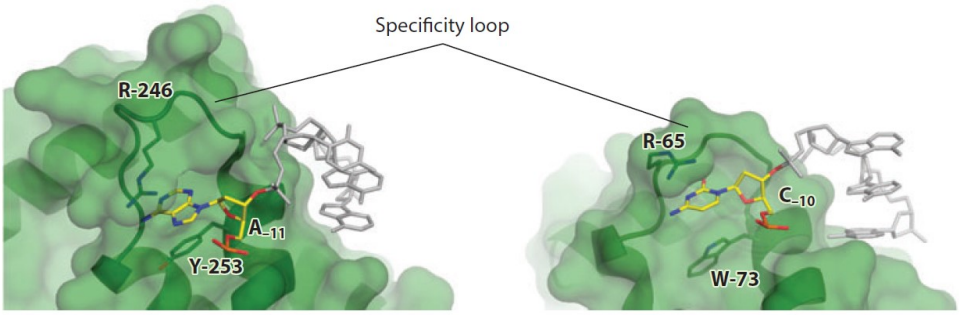


Figure 7. Binding of the conserved nucleotides of the -10 element by primary and ECF σ factors.

Fragments of T. aquaticus primary σA factor [PDB ID 3UGO, [62]] (left) and E. coli ECF σE [PDB ID 4LUP, [80]] (right) bound with non-template strands of their respective -10 promoter elements are shown. The figure is taken from [57], with permission.

The linker between σ_3 and σ_4 domains ($\sigma_{3.2}$), which is approximately 50 amino acids long in primary σ factors, penetrates deep into the RNAP main cleft and was therefore called the 'σ-finger'. In the crystal structures of RNAP holoenzymes bound to promoter DNA, the σ-finger extends along the template DNA strand from the -10 to -4 positions [76].

During early stages of RNA synthesis, the σ-finger blocks the synthesis of RNA transcripts longer than several nucleotides, thus promoting the formation of abortive transcripts [37, 81-84]. The synthesis of longer RNA transcripts can thus only occur when the σ-finger is pushed away. This event leads to the weakening of interactions between σ and the RNAP core, and also between RNAP and promoter [37, 81-85]. Consequently, the enzyme escapes the promoter and proceeds to elongate the transcript. The σ factor remains bound to the RNAP core over some distance of transcribed DNA but eventually dissociates from the complex [37, 81-84].

Since in RNAP holoenzyme/promoter complexes the σ-finger lies close to template DNA strand upstream of the transcription start site, it was proposed that it could

interact sequence specifically with nucleotides of the template DNA strand and contribute to the recognition of the -10 element (at least by some σ factors) [86]. However, there is no structural or biochemical evidence to support this notion.

1.4 Non-canonical multisubunit RNAPs

Multisubunit RNAP core enzymes of all cellular organisms contain homologs of bacterial RNAP α, β, β' and ω subunits, therefore it can be assumed that RNAP of the Last Universal Common Ancestor (LUCA) already contained ancestors of these subunits, which had the same specialized functions as they have in modern multisubunit RNAPs [3, 6].

The exciting area of the transcription field focuses on atypical transcription enzymes, mechanisms of their functioning, and the role these enzymes may have played in the evolution of two-barrel polymerases. Genes coding for distant homologs of cellular RNAP catalytic subunits (β and β' in bacteria) were found in genomes of some viruses, bacteriophages, prophages, as well as in likely mobile selfish elements in the genomes of some firmicutes and cyanobacteria, and in fungal killer plasmids [2, 7-9, 87-89]. Similarity between the products of these genes and catalytic subunits of cellular RNAPs varies significantly and in some cases is limited to a small – as short as 65 amino acids – region containing the metal binding motif [7]. Most of these genes are not accompanied by identifiable genes coding for remaining RNAP subunits. Yet, some of the products of these genes were shown to form functional RNAPs, while others remain uncharacterized. These partially characterized and non-characterized putative RNAPs are referred to as "non-canonical RNAPs" since they are highly diverged from multisubunit RNAPs of cellular organisms and have distinct subunit composition [9].

It is tempting to speculate that the simplified subunit composition of some non-canonical RNAPs and the lack of many conserved regions present in canonical RNAPs could be explained by ancient origin from times predating LUCA [3, 9, 87]. However, another, equally possible scenario is that non-canonical RNAPs are descendants of canonical RNAPs that due to fast evolution of viral/phage genomes and selfish mobile elements encoding these RNAPs and/or simple functional requirements, lost some subunits and conserved regions [89]. Careful phylogenetic analysis of two-barrel polymerases including all non-canonical RNAPs and recently discovered two-barrel DNA polymerases would be required to gain insights on this issue.

Bellow we will review some of the functionally characterized non-canonical RNAPs.

1.4.1 Phage single-subunit RNAPs related to multisubunit RNAPs

Bioinformatic analysis revealed a group of putative proteins encoded in prophages of some firmicutes which are homologous to two-barrel Qde-1 like RNA-dependent RNAPs (see Chapter 1, section 1.1) and thus are also related to canonical multisubunit RNAPs [3]. This group of prophage proteins was named as YonO-like group following the YonO protein of *Bacillus subtilis* phage SPβ [3]. Sequence homology between YonO and bacterial RNAP was found only within the two DPBBs of bacterial RNAP [3]. It was not initially known whether putative YonO-like proteins form RNA-dependent or DNA-dependent RNAPs, or in fact if they are active polymerases at all [3]. Recently, Forrest and co-workers purified and characterized functionally YonO [90]. They showed that this is a DNA-dependent RNAP which is responsible for transcription of late SPβ genes upon induction of a lytic cycle of the phage [90]. The YonO RNAP is more processive but less accurate than bacterial RNAPs [90]. Such features of the enzyme may be beneficial for

the phage [90]. Transcription initiation properties of the YonO RNAP were not investigated yet [90]. The structure of the enzyme is also not determined.

A protein of similar with the YonO size, related to multisubunit RNAPs was also predicted to be encoded in the genome of P23-45 phage infecting *T. thermophilus* [7]. Temporal transcription classes of the P23-45 genes and corresponding promoter consensus sequences were determined [7]. While the phage middle and late genes were shown to be transcribed by the host RNAP, transcription of early phage genes was found to be rifampicin-resistant, and therefore must proceed without involvement of bacterial host RNAP [7]. Bioinformatic analysis revealed the presence of an open reading frame (ORF) 64 in the P23-45 genome that encodes a protein with low similarity to a DPBB domain of bacterial RNAP β' [7]. The product of ORF64 is present in P23-45 virions and thus may function as phage RNAP injected into the host for transcription of early phage genes [7]. Despite a very limited region of homology to a fragment of β', the product of ORF64 was shown to be an active DNA-dependent RNAP *in vitro* (Minakhin, Severinov, personal communications).

Many researchers consider that single-subunit DNA-dependent RNAPs described in this Subchapter together with the Qde-1 like enzymes are direct descendants of a primordial RNAP that existed before LUCA [2, 3, 90]. However, an alternative scenario also exists where the two-barrel single-subunit RNAPs are descendants of canonical RNAPs acquired from bacteria by phage/phages and evolved in a way that genes coding for DPBBs were fused. Subsequent fast evolution of the phage genomes could have led to elimination of some domains and subunits since phages do not require such complex regulation of transcription as is the case of cellular organisms. The fusion of RNAP genes coding for two DPBBs is not a rare event in evolution: the putative RNAPs encoded by

fungal killer plasmids and Cgl1702 protein from *Corynebacterium glutamicum* are also fusions of distant homologs of bacterial β and β' while evolutionary these proteins are much closer to canonical RNAPs rather than to the described two-barrel single-subunit RNAPs and thus evolved independently from the last [3, 9]. The apparently recent independent fusions of the genes of β and β' subunits also occurred in several groups of parasitic bacteria such as *Wolbachia* and *Helicobacter* [91, 92]. Moreover, according to the hypothesis on origins and evolution of eukaryotic RNA interference systems, the two-barrel RNA-dependent RNAPs were acquired by the protomitochondrial (α-proteobacterial) endosymbiont from a bacteriophage [93] contradicting with the hypothesis that the Qde-1 like enzymes are direct descendants of the primordial RNAP existed before the LUCA.

1.4.2 Viral multisubunit RNAPs

Eukaryotic viruses belonging to the superclade of large nucleocytoplasmic DNA viruses (NCLDV) replicate in cytoplasm and thus require special transcription enzymes for expression of their genes [89]. Genomes of viruses from the NCLDV group encode homologs of the two largest subunits of cellular RNAP and up to seven homologs of other RNAP II subunits [89]. All these viruses encode homologs of eukaryotic transcription initiation factors and almost all encode a homolog of eukaryotic elongation factor TFIIS which induces cleavage of nascent RNA in backtracked elongation complexes [89].

The most investigated viral RNAP from this group is a multisubunit RNAP encoded by vaccinia virus [89]. This RNAP was purified from the virus particles [94]. It was shown that the core of vaccinia virus RNAP consists of 8 subunits: two of them are homologs of the largest RNAP II subunits while others are smaller polypeptides among

which only one is non-homologous to RNAP II subunits [89, 94]. The vaccinia virus RNAP does not contain homologs of the assembly platform subunits [89, 94]. Interestingly, the core of vaccinia virus RNAP contains a homolog of TFIIS elongation factor which therefore is tightly associated with viral RNAP in contrast to counterparts from archaea and eukarya, which are dissociable [89, 94]. Such association of the transcription elongation factor rescuing RNAP from a stalled state may be beneficial for a virus which may require faster transcription [89]. The core of the vaccinia virus RNAP was shown to transcribe nonspecifically single-stranded DNA but was inactive on double-stranded DNA [94]. For promoter-specific transcription of early viral genes the core of the vaccinia virus RNAP associates with RAP94, a protein with no homology to any known protein [95]. Promoter specific transcription also requires a heterodimeric virus early transcription factor (VETF), which similarly to TFIID binds to a specific A/T-rich promoter element [95]. The vaccinia virus RNAP core and RAP94 are synthesized at different times during infection and thus expression of viral genes can be partially modulated in this way [95]. Interestingly, it was shown that in multiple-round non promoter-specific transcription of single-stranded DNA the vaccinia virus RNAP is able to displace the RNA molecule from RNA/DNA hybrids [94].

Baculoviruses do not belong to the NCLDV group and replicate in nuclei of eukaryotic cells. Nevertheless, they encode multisubunit RNAP [96]. It was shown for *Autographa californica* baculovirus that transcription of early viral genes is performed by cellular RNAP II, while transcription of late and very late genes is performed by viral enzyme [96]. This RNAP was purified and shown to consist of four proteins: LEF-8, LEF-9, LEF-4 and p47 (LEF here stands for late expression factor) [96]. The purified RNAP was resistant to α-amantin (an inhibitor of RNAP II) and recognized late and very

late viral promoters but not early viral promoters [96]. The LEF-8 and LEF-9 subunits are distant homologs of the largest cellular RNAP subunits [96]. The function of p47 is not known but it has very low similarity to α subunit of bacterial RNAP and thus may serve an assembly platform for baculoviral RNAP [97]. LEF-4 has a guanyltransferase activity which was shown to be essential in vivo for capping of viral mRNAs [98].

Information on transcription initiation by baculoviral RNAP is limited. It was shown that the LEF-5, a baculoviral homolog of TFIIS, is essential for efficient viral transcription *in vivo* [99]. Unexpectedly *in vitro* experiments revealed the LEF-5 does not affect elongation stage but enhances efficiency of transcription initiation from viral late promoters up to 10 times [100]. Thus, in contrast to its cellular homologs, LEF5 is a transcription initiation factor rather than transcription elongation factor [100].

No three-dimensional structure is available for any of non-canonical viral RNAP.

1.4.3 Multisubunit RNAPs of giant phages

Two sets of distant homologs of bacterial RNAP β and β' subunits are encoded in the genomes of some phages from the giant bacteriophage group [88]. Development of several giant phages was shown to be independent of bacterial host RNAP, confirming that these phages rely exclusively on their own transcription machinery for expression of their genes [101, 102].

Giant phages are a highly diverse group of phages that belong to *Myoviridae* family [103]. They have mosaic genomes larger than 200,000 bp coding for up to several hundred proteins [103]. The functions of most of these proteins are not known. Even a subgroup of giant phages whose genomes encode RNAPs is very diverse and includes phages infecting gram-positive and gram-negative bacteria. This could mean that the

ancestor of phages of this subgroup could have acquired RNAP genes from a cell before the divergence of gram-negative and gram-positive bacteria.

No genes coding for homologs of α and ω subunits or promoter-specificity σ factors have been identified in giant phage genomes. Whenever it has been investigated, one set of β/β' homologs is found in giant phage virions [88, 104-107], likely forming a virion RNAP (vRNAP) that is injected into bacterial cell along with phage DNA and transcribes early phage genes. Another set of β/β' homologs forms a non-virion RNAP (nvRNAP) synthesized during subsequent stages of infection and transcribing late phage genes, including the vRNAP genes.

The vRNAP has not been purified to date while the nvRNAP encoded by phiKZ phage infecting *Pseudomonas aeruginosa* was purified and partially characterized [108]. The phiKZ nvRNAP was shown to consist of four phage proteins jointly comprising the full-length β- and β′-like subunits and a fifth subunit gp68 with no sequence similarity to functionally characterized proteins [108]. Homologs of the gp68 are found in all other giant phage genomes encoding β- and β′-like subunits. It was shown that this enzyme specifically recognizes late phage promoters *in vitro* [108]; however further transcription initiation properties were not investigated. The role of the gp68 was not established.

AR9 phage is a giant phage infecting *B. subtilis* [88, 102]. A distinguishing feature of the AR9 phage is the presence of uracils instead of thymines in its double-stranded DNA genome [88]. Sequencing of the phage genome revealed that the AR9 is a close relative of the PBS2 phage judging by the identity of genes coding for uracil DNA glycosylase inhibitor (the only known sequence of the PBS2 phage) [88]. Interestingly, a multisubunit RNAP was purified from *B. subtilis* cells infected with PBS2 long time ago

[109, 110]. The PBS2 RNAP was not characterized in detail but it was found that one of its subunits is dissociable [109, 110].

The global transcript profiling of *B. subtilis* cells infected with AR9 revealed the presence of early and late phage genes and allowed to identify consensus sequences of early and late phage promoters, presumably recognized by AR9 vRNAP and AR9 nvRNAP, correspondingly (Fig. 8) [102].

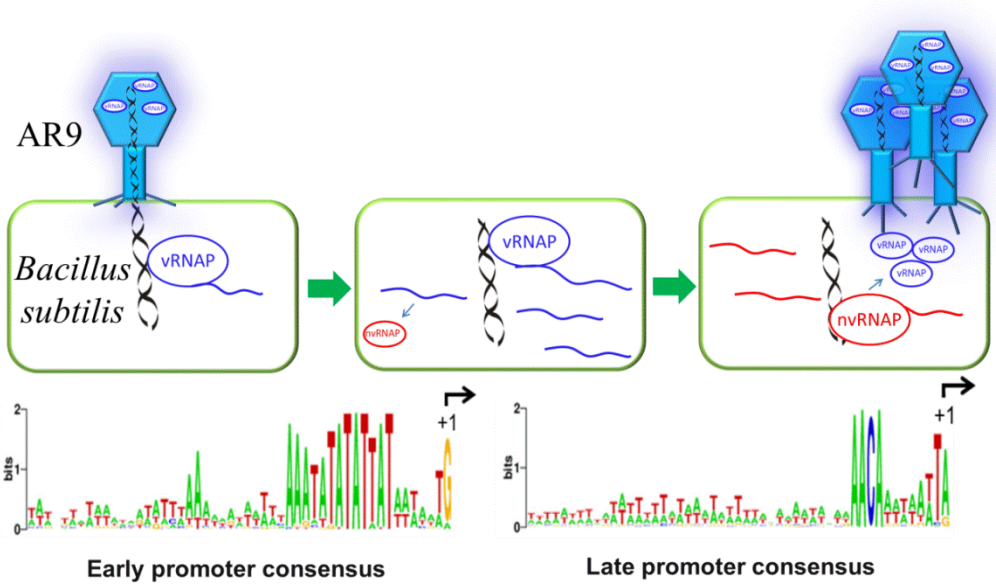


Figure 8. Transcription strategy and promoters of the AR9 phage.

AR9 vRNAP is injected into the B. subtilis cell along with the phage DNA and transcribes early phage genes from early phage promoters characterized by a consensus shown at the left [102]; AR9 nvRNAP is synthesized during the infection and transcribes late phage genes including vRNAP genes from late phage promoters characterized by a consensus shown on the right [102].

The goal of work described in this thesis was to purify and characterize the nvRNAP encoded by the AR9 phage.

Chapter 2. Materials and Methods

2.1 Bacteriophage, bacterial strain and growth conditions

Bacteriophage AR9 was generously provided by David Dubnau from the Public Health Research Institute Center New Jersey Medical School – Rutgers, NJ.

To prepare AR9 lysates, a single plaque was resuspended in 100 µl of LB media and added to the 0.5 L of *B. subtilis* 168 culture at OD_{595}=0.5 and incubated with shaking at 37 °C until complete lysis occurred (3–4 h). Cell debris was removed by centrifugation at 5000 g for 20 min. The resulting phage lysate stock ($5x10^9$–$2x10^{10}$ PFU/ml) was stored at 4 °C.

For purification of AR9 nvRNAP, 20 liters of *B. subtilis* cells were grown up to OD_{595}=1 and infected with AR9 phage at a MOI of 10. The infection was stopped after 22 minutes by chilling the culture on an ice water bath followed by centrifugation at 3500 g for 30 minutes at 4 ^{o}C. The resulting pellets were stored at -20 ^{o}C.

2.2 Purification of AR9 nvRNAP from infected cells

All steps of the following procedure were done on ice or at 4 ^{o}C. Twenty grams of infected *B. subtilis* cells were disrupted by sonication in 100 ml of buffer A (40 mM Tris-HCl pH 8, 5 mM EDTA, 5 mM β-mercaptoethanol, 0.1 mM PMSF) containing 50 mM NaCl followed by centrifugation at 15000 g for 30 min. An 8% Polyethyleneimine (polymin P) solution (pH 8.0) was added with stirring to the cleared lysate to the final concentration of 0.8%. The resulting suspension was incubated on ice for 30 min and centrifuged at 10 000 g for 15 min. The supernatant was removed and the pellet was resuspended in buffer A containing 0.3 M NaCl. After 10 minutes incubation, the PEI pellet was formed by centrifugation as previously. Supernatant containing 0.3 M NaCl

extract from the PEI pellet was saved for further analysis. Then, extraction was repeated twice with buffer A containing 0.5 M NaCl and 1 M NaCl. Eluted proteins were precipitated by addition of ammonium sulfate to 67% saturation and dissolved in buffer A without NaCl. The same procedure also was done for uninfected cells. All samples were loaded onto a 5 ml HiTrap heparin-sepharose HP column (GE Healthcare) equilibrated with buffer A with 0.1 M NaCl. The column was washed with buffer A with 0.1 M NaCl. Then, step elution with buffer A containing 0.3 M NaCl, 0.6 M NaCl and 1 M NaCl was carried out. Heparin-sepharose chromatography was done for three PEI extracts from infected and uninfected cells. All fractions were analyzed by denaturing SDS polyacrylamide gel electrophoresis (SDS-PAGE). The bands missing in samples obtained from uninfected cells were analyzed by mass-spectrometry. Following this way, fractions containing gp089 and gp154 were found. They corresponded to fractions eluted in 0.6 M and 1 M NaCl, respectively, from the Heparin-sepharose column during chromatography of 1 M NaCl PEI-extract. The bacterial RNAP was separated from the nvRNAP during heparin-sepharose chromatography, where it was eluted at 0.6 M NaCl in fractions ahead of the nvRNAP.

These fractions were pooled and concentrated by ultrafiltration (Amicon Ultra-4 Centrifugal Filter Unit with Ultracel-30 membrane, EMD Millipore) and loaded onto a Superdex 200 Increase 10/300 (GE Healthcare) gel filtration column equilibrated with buffer A containing 200 mM NaCl. As a final purification step, the combined nvRNAP fractions eluted from the Superdex 200 column were diluted 4-fold with buffer A and applied to a MonoQ HR 5/5 column (GE Healthcare). Bound proteins were eluted with a linear 0.25–0.45 M NaCl gradient in buffer A. The nvRNAP was eluted from the column at 0.34–0.38 M NaCl. The fractions containing nvRNAP subunits were concentrated to a

final concentration 0.5 mg/ml, then glycerol was added up to 50% to the sample for storage at −20 °C.

2.3 Native gel electrophoresis

One microgram of AR9 nvRNAP was resolved by a native 5%-PAGE. A single band was revealed by Coomassie blue staining. To determine the protein composition of this band, it was excised from the native gel and the gel piece was placed into a well of an SDS 8%-polyacrylamide gel, supplemented with 5–8 µl of Laemmli loading buffer and subjected to electrophoresis. The SDS gel was silver stained.

2.4 DNA templates for transcription assay

Genomic DNA of AR9, phiR1-37 and phiKZ bacteriophages for transcription assay were purified using the QIAGEN Lambda Midi Kit according to the manufacturer's instructions.

DNA templates containing late AR9 promoters and their derivatives were prepared by polymerase chain reaction (PCR). PCRs were done with Encyclo DNA polymerase (Evrogen, Moscow) and the AR9 genomic DNA as a template, with a standard concentration of dNTPs to obtain DNA fragments with thymine or in the presence of dUTP in place of dTTP to obtain DNA fragments with uracil. Oligonucleotide primers used for PCR are listed in Table 1, Appendix A.

To synthesize promoter templates for analysis of the consensus sequence, PCR with oligonucleotide primers bearing single substitution at desired positions of the promoter was performed. Since thymine-containing oligonucleotide primers were used, the final templates were hybrids with respect to their thymine/uracil content (full sequences of the primers and resulting templates are shown in Table 2, Appendix A). This strategy was used to constrain costs (uracil-containing oligonucleotides are

expensive). Since we found that the AR9 nvRNAP efficiently and specifically transcribes from such templates containing the wild-type P007 and P077 promoters with thymines in functionally important positions of the non-template strand (Fig. 13, lanes 1), we concluded that such "hybrid" strategy is appropriate for mutational analysis.

Double-stranded and partially single-stranded DNA templates containing the P007 and P077 promoters with uracils and thymines at certain positions were prepared by annealing of oligonucleotides ordered from Integrated DNA Technologies (IDT) and listed in Table 4, Appendix A. To prepare specific DNA templates, two corresponding oligonucleotides were annealed together by mixing in buffer containing 20 mM Tris–HCl and 40 mM KCl, incubating at 75 °C for 1 minute and cooling down to 4 °C by a decrement of 1°C per minute.

Single-stranded DNA templates containing the P007 promoter were ordered from IDT and listed in Table 5, Appendix A.

To prepare RNA/DNA scaffold, the template DNA oligonucleotide (5'-GGTCCTGTCTGAAATTGTTATCCGCTAC-3'), the non-template DNA oligonucleotide (5'-ACAATTTCAGACAGGACC-3') and the ^{32}P-end-labeled RNA oligonucleotide (5'-GUAGCGGA-3') were mixed in concentrations 1 μM, 1 μM, and 0.5 μM respectively in a buffer containing 20 mM Tris–HCl, 40 mM KCl, 1 mM $MgCl_2$ and 0.5 mM DTT, incubated at 65 °C for 1 minute and cooled down with an increment of 1°C per minute.

2.5 Primer extension and sequencing reactions

For in vitro primer extension reaction RNA was synthesized by AR9 RNAP for 15 min at 37 °C from PCR fragments containing late AR9 promoters in 50 μl of transcription buffer (20 mM Tris–HCl, 40 mM KCl, 1 mM $MgCl_2$, 0.5 mM DTT, and 100

μg/ml bovine serum albumin) in the presence of 100 μM each of ATP, CTP, GTP, UTP. RNA was purified with TRIzol reagent (Invitrogen) according to manufacturer's protocol and used for primer extension reaction. The primers indicated by an asterisk in Table 1, Appendix A were labeled with [γ-^{32}P]-ATP by phage T4 polynucleotide kinase (New England Biolabs), as recommended by the manufacturer. The purified RNA was reverse-transcribed from a ^{32}P-end-labeled primer with Maxima enzyme (Thermo Fisher Scientific) according to the manufacturer's protocol. The reactions were stopped by addition of a loading buffer and heating at 85 °C. Sequencing reactions were carried with USB Thermo Sequenase Cycle Sequencing Kit (Thermo Fisher Scientific) on the PCR products containing corresponding start sites, with the primers used for primer extension reactions. The reaction products of sequencing and reverse transcription reactions were resolved on 6-8% (w/v) denaturing polyacrylamide gels and visualized using a PhosphorImager (Molecular Dynamics).

2.6 *In vitro* transcription

Multiple-round run-off transcription reactions were performed in 10 μl of transcription buffer (20 mM Tris–HCl, 40 mM KCl, 1 mM $MgCl_2$, 0.5 mM DTT, and 100 μg/ml bovine serum albumin) and contained 30-50 nM AR9 nvRNAP and either 0.06 nM phage genomic DNA or 30-50 nM of indicated DNA template. The reactions were incubated for 10 min at 37 °C, followed by the addition of 100 μM each of ATP, CTP, and GTP; 10 μM UTP and 3 μCi [α-^{32}P] UTP (3000 Ci/mmol). Where indicated, rifampicin was added to the final concentration of 10 μg/ml. Reactions proceeded for 30 min at 37 °C and were terminated by the addition of an equal volume of denaturing loading buffer. The reaction products were resolved by electrophoresis on 6-20 % (w/v)

denaturing 7 M urea polyacrylamide gel and visualized by PhosphorImager (Molecular Dynamics).

Abortive transcription initiation reactions were set at the same general conditions as run-off transcription reactions but supplemented with 175 μM of initiating RNA dinucleotides specified by the -1/+1 positions of promoters studied (UpG for P007 and UpA for P077 promoters were used). Reactions were incubated for 10 min at 37 °C, followed by the addition of 3 μCi [α-^{32}P] UTP (3000 Ci/mmol). The reactions were allowed to proceed for 15 min at 37 °C and terminated by the addition of an equal volume of denaturing loading buffer. Abortive initiation reaction products were resolved by electrophoresis on 20% (w/v) denaturing 7 M urea polyacrylamide gels and visualized by PhosphorImager (Molecular Dynamics).

Transcription reactions from RNA/DNA scaffold were set at the same buffer as run-off transcription reactions and contained 15 nM RNA/DNA scaffold and 15 nM AR9 nvRNAP. Reactions were incubated for 10 min at 30 °C, followed by the addition of 1 mM each of ATP, CTP, GTP, and UTP. Reactions proceeded for 15 min at 37 °C and were terminated by the addition of an equal volume of denaturing loading buffer. The reaction products were resolved on 18% (w/v) denaturing 7 M urea polyacrylamide gel and visualized as described above.

2.7 Footprinting reactions

DNA templates for footprinting reactions were prepared by PCR (as templates for transcription reactions) with a ^{32}P-end-labeled reverse primer to obtain template strand labeled or with a ^{32}P-end-labeled forward primer to obtain non-template strand labeled (Table 3, Appendix A). Promoter complexes were formed in 20-μl reactions containing 50 nM AR9 nvRNAP and 30 nM ^{32}P-end-labeled DNA fragment in a buffer with 20 mM

Tris–HCl, 40 mM KCl, 1 mM $MgCl_2$, and 100 μg/ml bovine serum albumin. Reactions were preincubated for 10 min at 37°C. DNase footprinting reaction was initiated by addition of 1 Unit of DNase I (Ambion). The reaction proceeded for 30 s at 37 °C and was terminated by addition of EDTA to 15 mM followed by phenol extraction and ethanol precipitation. For $KMnO_4$ probing, promoter complexes were treated with $KMnO_4$ (2 mM) for 20 s at 37 °C. Reactions were terminated by addition of β-mercaptoethanol to 450 mM, followed by ethanol precipitation, and 15 min treatment with 10% piperidine at 95 °C followed by addition of chloroform up to 10% and vortexing. Then, samples were centrifuged followed by ethanol precipitation of DNA from the aqueous phase. Products of footprinting reactions were resolved by electrophoresis on 8% (w/v) denaturing 7M urea sequencing polyacrylamide gels and visualized by PhosphorImager (Molecular Dynamics).

2.8 Cloning of AR9 nvRNAP

Plasmids were constructed using synthetic gene-Blocks (gBlocks) ordered from IDT, containing AR9 nvRNAP gene fragments with codons optimized for expression in *E. coli*. First, two gBlocks coding for gp270 N-terminally fused to a hexahistidine tag, gp154, and pETDuet-1 digested by NcoI and BamHI were assembled using the NEBuilder HiFi DNA Assembly Master Mix (New England Biolabs). Then, the obtained plasmid digested by BglII and XhoI was assembled with two gBlocks coding for gp105 and gp089. The resulting plasmid encoded the AR9 nvRNAP core. To get the plasmid encoding AR9 nvRNAP holoenzyme, the plasmid coding for AR9 nvRNAP core was linearized by XhoI and assembled with the gBlock coding for gp226. In both plasmids each RNAP gene was preceded with T7 RNAP promoter and *lac* operator.

To get the plasmid coding for AR9 nvRNAP core lacking the tag, a fragment containing four AR9 nvRNAP genes (without sequence coding the tag) was amplified by PCR from one of the obtained plasmids and was assembled with pETDuet-1 digested by NcoI and XhoI.

2.9 Purification of recombinant AR9 nvRNAP

For protein purification 3 liters of *E. coli* BL21 Star (DE3) transformed with a corresponding plasmid were grown to OD_{595}=7 and induced with 1 mM IPTG for 3.5 hours. Cells were pelleted by centrifugation at 4000 g for 30 minutes at 4 ^{o}C and stored at -20 ^{o}C.

All steps of the following procedure were done on ice or at 4 ^{o}C. Seven grams of cells were resuspended in 50 ml of buffer C (50 mM NaH_2PO_4 pH 8, 300 mM NaCl, 3 mM β-mercaptoethanol, 0.1 mM PMSF). Lysozyme was added to the final concentration of 1 mg/ml; after 30 min incubation cells were disrupted by sonication followed by centrifugation at 15000 g for 30 min. Then the lysate was loaded on 5 ml Ni-NTA column (Qiagen) equilibrated with buffer C, washed with 5 column volumes of buffer C and with 5 column volumes of buffer C containing 20 mM Imidazole. Then, elution with buffer C containing 200 mM Imidazole was carried out.

All fractions were analyzed by denaturing SDS polyacrylamide gel electrophoresis (SDS-PAGE). Further, fractions containing AR9 nvRNAP were pooled and diluted ten times by buffer A for anion exchange chromatography and applied to a MonoQ 10/100 column (GE Healthcare) (for AR9 nvRNAP core the buffer A is 20 mM Tris pH 8, 0.5 mM EDTA, 1 mM DTT and for AR9 nvRNAP holo the buffer A is 20 mM Bis-tris propane pH 6.8, 0.5 mM EDTA, 1 mM DTT). Bound proteins were eluted with a linear 0.25–0.45 M NaCl gradient in buffer A.

The AR9 nvRNAP core was also subjected to gel-filtration on a Superdex 200 10/300 (GE Healthcare) column equilibrated with buffer A containing 100 mM NaCl. The AR9 nvRNAP holoenzyme was not subjected to gel-filtration, salt concentration in the sample was lowered during the concentration procedure.

The fractions containing AR9 nvRNAP were concentrated to a final concentration 20 mg/ml and stored at 4 °C for 5 days.

2.10 Crystallization of AR9 nvRNAP

Crystals of AR9 nvRNAP were grown by vapor diffusion method.

AR9 nvRNAP core carrying the tag:

1.5 µl of protein solution (4.5 mg/ml) were mixed with the same volume of a solution containing 100 mM Tricine pH 8.8, 270 mM KNO_3, 15 % PEG 6000, 5 mM $MgCl_2$, and incubated as a hanging drop over the same solution. Crystals grew in 1 week at 19 C° temperature. For flash freezing crystals were shortly soaked in the crystallization solution containing 25% ethylene glycol.

AR9 nvRNAP core lacking the tag:

1.5 µl of protein solution (7.5 mg/ml) were mixed with the same volume of a solution containing 150 mM Malic acid pH 7, 150 mM NaCl, 14 % PEG 3350 and incubated as a hanging drop over the same solution. Crystals grew in 1 week at 19 C° temperature. For flash freezing crystals were shortly soaked in the crystallization solution containing 25% ethylene glycol.

AR9 nvRNAP holoenzyme in complex with promoter containing DNA:

To prepare DNA template for crystallization, two corresponding oligonucleotides (shown in Fig. 29) at final concentrations 100 µM each were annealed together by mixing in buffer containing 20 mM Bis-tris propane pH 6.8, 100 mM NaCl, 4 mM $MgCl_2$, 0.5

mM EDTA incubating at 65 °C for 1 minute and cooling down to 4 °C by a decrement of 1°C per minute.

A 1.5-fold molar excess of the DNA template was added to the holoenzyme and incubated for 30 min at room temperature (the final concentrations: 10 mg/ml of the protein (34 μM) and 50 μM of the DNA).

1.5 μl of protein/DNA complex solution were mixed with the same volume of a solution containing 150 mM MIB pH 5, 150 mM LiCl, 14 % PEG 1500 and incubated as a hanging drop over the same solution. Crystals grew in 1-2 weeks at 19 C° temperature. For flash freezing crystals were shortly soaked in the crystallization solution containing 25% ethylene glycol.

2.11 Preparation of heavy-atom derivative crystals

The following compounds were tried to derivatize crystals (by co-crystallization or soaking): $SrCl_2$, $GdCl_3$, Na_2WO_4, $HgCl_2$, $Pb(NO_3)_2$, Thimerosal (2-$(C_2H_5HgS)C_6H_4CO_2Na$), 10 compounds containing Eu and Yb (JBS Lanthanide Phasing Kit), 3 compounds containing W (JBS Tungstate Cluster Kit), 1 compound containing Ta (JBS Tantalum Cluster Derivatization Kit).

For soaking, the crystallization solution was supplemented with different concentrations of a corresponding compound (between 0.1 mM and 100 mM). The time of soaking varied between 2 hours and 2 days. Among all cheeked conditions soaking in 10 mM Thimerosal and in 1 mM Tantalum Cluster for a night gave derivatized crystals (judging by the presence of anomalous signal in X-ray diffraction data).

Chapter 3. Functional characterization of AR9 nvRNAP

The work described in this Chapter was performed by the author in Skoltech Research Center in the center of nano- and biotechnologies of Peter the Great St. Petersburg Polytechnic University in Saint Petersburg (http://www.nanobio.spbstu.ru) and in Konstantin Severinov's laboratory in the Institute of Molecular Genetics of Russian Academy of Sciences in Moscow (https://www.img.ras.ru/en).

3.1 Results

3.1.1 Purification of a multisubunit phage RNAP from AR9 infected cells

To purify phage-encoded RNAP(s), cell lysates of *B. subtilis* cultures infected with AR9 at high multiplicity of infection (MOI) and collected midway through the infection cycle were subjected to fractionation following the standard bacterial RNAP purification scheme involving polyethyleneimine (Polymin P) fractionation, heparin-sepharose affinity chromatography, gel-filtration, and anion exchange chromatography (Fig. 9 A, left panel). Extraction of Polymin P pellet with a buffer containing 1.0 M NaCl yielded, after heparin-sepharose chromatography, fractions that contained two prominent protein bands with apparent molecular weights of ~80 and ~75 kDa (indicated by asterisks in Fig. 9 A, right panel, lane 3).

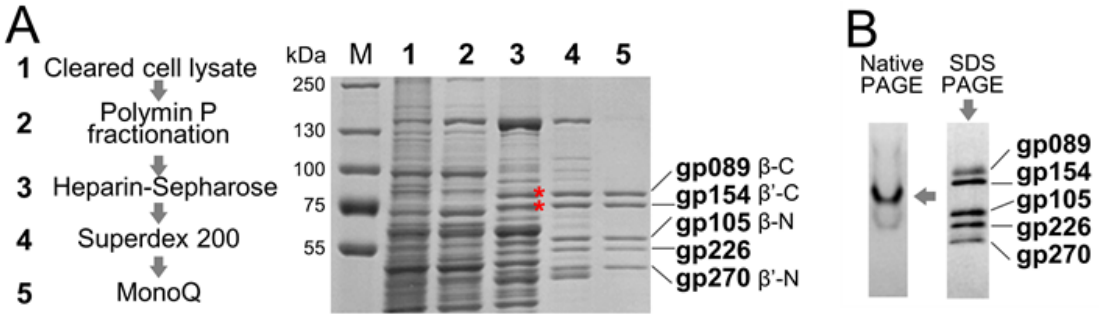


Figure 9. Purification of nvRNAP from AR9 infected Bacillus subtilis cells.

(A) Left: main steps of nvRNAP purification. Right: SDS-PAGE analysis of fractions containing gp089 and gp154 (marked by red asterisks) during the course of AR9 nvRNAP purification. A Coomassie-stained gel is shown; lane numbers correspond to the steps of purification shown on the left. (B) Left: a Coomassie-stained gel after native PAGE analysis of the five-subunit form of AR9 nvRNAP after the final MonoQ purification (step V in panel A). Right: a silver-stained gel after SDS-PAGE analysis showing polypeptides present in the native gel band marked by an arrow.

Mass-spectrometric analysis of these bands identified them as AR9 gp089 and gp154, the presumed subunits of nvRNAP homologous to C-terminal parts of bacterial RNAP β and β' subunits, respectively [88]. By following the gp089 and gp154 bands during subsequent chromatographic steps, a fraction from a MonoQ column that contained five protein bands as judged by SDS-PAGE (Fig. 9 A, right panel, lane 5) was obtained. In addition to gp089 and gp154, this fraction also contained two AR9 polypeptides homologous to the N-terminal parts of bacterial RNAP β and β' subunits, gp105 and gp270, respectively [88]. The fifth polypeptide was gp226, a distant homolog of phiKZ gp68, a subunit of the recently purified phiKZ nvRNAP with unknown function [108]. All five polypeptides migrated in a single band during non-denaturing gel electrophoresis (Fig. 9 B), indicating that they form a complex, which we will refer to as AR9 nvRNAP. The subunit composition of AR9 nvRNAP corresponds to that reported long ago for an RNAP isolated from *B. subtilis* culture infected with a closely related PBS2 phage, with gp089, gp154, gp105, gp226, and gp270 of AR9 likely matching PBS2 P80, P76, P58, P53, and P48, respectively [109, 110].

3.1.2 In vitro transcription by AR9 nvRNAP

The PBS2 RNAP was reported to transcribe genomic DNA of the phage *in vitro* [109]. Transcription of several other phage genomes was much less efficient [109]. We tested the AR9 nvRNAP for transcription from genomic DNA of the AR9, phiR1-37, and

phiKZ phages. For each template, transcription reactions were conducted in the presence or in the absence of rifampicin, a host RNAP inhibitor. The result, shown in Fig. 10, revealed that the AR9 nvRNAP was highly active on the AR9 template, was partially active on the phiR1-37 template, and was inactive on the phiKZ template. Whenever transcription was observed, it was rifampicin-resistant. Control transcription by host RNAP was sensitive to rifampicin.

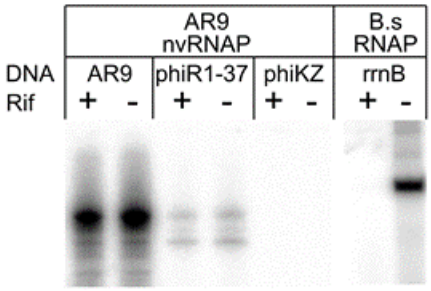


Figure 10. Analysis of AR9 nvRNAP transcriptional activity.

In vitro transcription by AR9 nvRNAP of genomic DNA of AR9, phiR1-37, and phiKZ phages in the presence and in the absence of rifampicin. Transcription by B. subtilis RNAP of a PCR-fragment containing the rrnB promoter was used as a control.

As mentioned above, the nvRNAP likely transcribes late viral genes. Late AR9 promoters were previously identified in the course of global transcript profiling of AR9-infected cells [102]. When PCR fragments containing several predicted late promoters were tested as templates in *in vitro* transcription reactions with the nvRNAP, no transcription products were detected (Fig. 11, top panel). Since AR9 nvRNAP transcribed the AR9 and phiR1-37 genomic DNA both of which contain uracil instead of thymine [88, 111], we considered whether the presence of uracil is required for transcription. Accordingly, DNA templates with late promoters containing uracil instead of thymine were tested for *in vitro* transcription. Robust transcription by AR9 nvRNAP was observed from every template tested (Fig. 11, bottom panel).

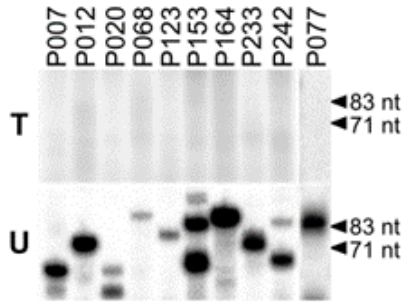


Figure 11. In vitro transcription by AR9 nvRNAP from late AR9 promoters.

Multiple-round run-off transcription by AR9 nvRNAP was performed using templates containing indicated late AR9 promoters. The templates for transcription were prepared by PCR either with dTTP (top panel) or dUTP (bottom panel). The primers used to prepare the DNA templates are listed in Table 1, Appendix A.

The 5' ends of transcripts generated by the AR9 nvRNAP *in vitro* were mapped by primer extension analysis and matched late promoter transcription start sites (TSSs) revealed *in vivo* (Fig. 12). We therefore conclude that the five-subunit AR9 nvRNAP recognizes late AR9 promoters. We further conclude that AR9 nvRNAP specifically transcribes late promoter-containing templates with uracil in place of thymine.

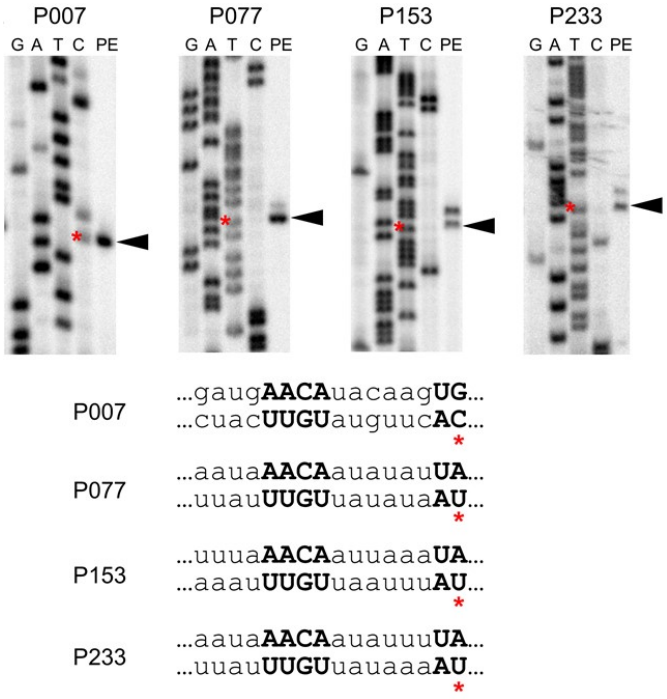


Figure 12. Primer extension analysis of in vitro transcripts synthesized from templates containing late AR9 promoters.

Late AR9 promoters are indicated. For each template DNA sequencing reactions obtained with a primer used in a primer extension reaction are shown as markers. Arrows indicate primer extension products while red asterisks indicate TSSs predicted from global transcription profiling of infected cells. Below, the sequences at and immediately upstream of the determined TSSs for analyzed promoters are shown. Conserved positions are shown in capital bold letters.

3.1.3 Functional analysis of AR9 late promoter consensus element

To determine the role of the late promoter 5'-A^{-11}ACA-(6N)-UA/G^{+1}-3' consensus motif [102] in transcription by the AR9 nvRNAP, DNA templates bearing single-substitutions at conserved and non-conserved positions of the motif were tested in an *in vitro* multiple-round run-off transcription assay (Fig. 13). Mutations were introduced into the P007 and P077 late phage promoters. For both promoters, substitutions at the positions -11, -10, -9, and -8 with respect to the TSS fully abolished transcription, indicating that the conserved 5'-$A^{-11}$$ACA^{-8}$-3' motif plays a crucial role in

promoter specific transcription. Substitutions at the +1 position also strongly decreased transcription. Substitutions at non-conserved promoter positions and at conserved position -1 had a smaller effect. We therefore conclude that the late promoter 5'-A^{-11}ACA-(6N)-UA/G^{+1}-3' consensus motif is necessary for in vitro transcription by the AR9 nvRNAP.

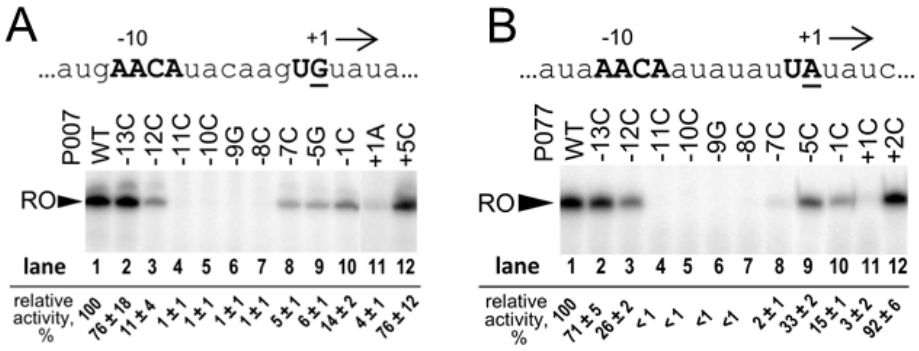


Figure 13. Late promoter consensus analysis.

Mutational analysis of the AR9 late promoters: (A) – P007, (B) – P077. Atop: nucleotide sequence of the non-template strand at and around the TSS of the promoter DNA. The position of the +1 start site is underlined. The conserved nucleotides of the late promoters are shown in capital bold letters. Below: In vitro run-off transcription by AR9 nvRNAP of the DNA templates containing the late promoters and its derivatives are shown. RO, run-off transcripts ((A) – 62nt; (B) – 59 nt). The numbers under the gel indicate the transcription activities relative to the positive controls (wild-type promoters). Average values and standard deviations from three independent experiments are presented. The full sequences of the templates can be found in Table 2, Appendix A.

3.1.4 Characterization of AR9 nvRNAP-promoter complex

To characterize AR9 nvRNAP promoter complexes we performed DNase I footprinting and $KMnO_4$ probing. DNase I footprinting of the P077 promoter complex on uracil-containing DNA revealed that AR9 nvRNAP protected the template strand positions between ca. -20 to +20 and the non-template strand positions from ca. -20 to

+13 (Fig. 14 A, lanes 2 on the left and right panels, respectively). Some upstream positions (-25, -36, -44) became hypersensitive to DNase I attack in the presence of AR9 nvRNAP. When AR9 nvRNAP was added to thymine-containing promoter template no significant protection from DNase I digestion was observed (Fig. 14 A, lanes 4 on the left and right panels). Thus, the absence of transcription from thymine-containing late promoters is caused by the inability of AR9 nvRNAP to bind to such templates.

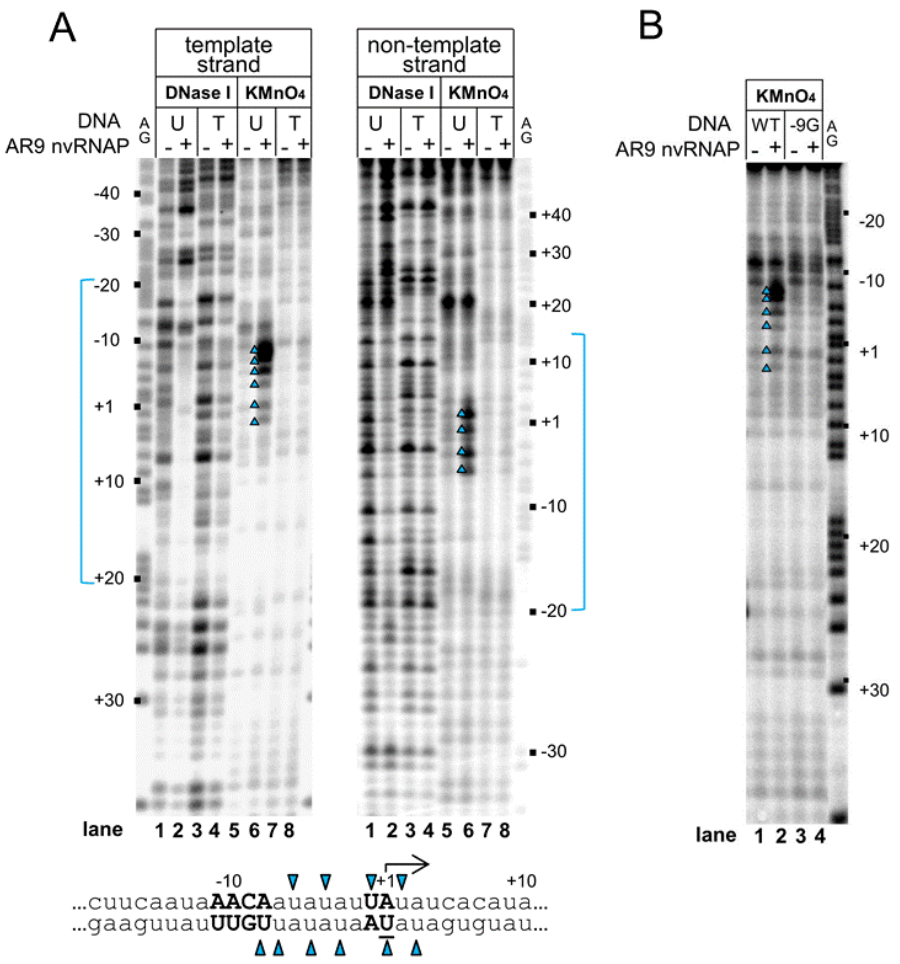


Figure 14. Promoter binding and promoter opening by AR9 nvRNAP.

(A*) DNase I footprinting and $KMnO_4$ probing of nvRNAP complexes with the P077 promoter DNA was performed with DNA templates containing uracil (U) or thymine (T). Positions relative to the TSS (+1) are indicated. Lanes indicated as "AG" show markers. Areas protected from DNase I attack are indicated in blue. A fragment of the P077 promoter sequence is shown below, with uracils that undergo oxidation by $KMnO_4$ in the*

presence of AR9 nvRNAP indicated by blue triangles. (B) $KMnO_4$ probing of nvRNAP–promoter P077 complexes formed with the wild-type promoter (WT) and the promoter bearing single substitution at the -9 position (-9G). The experiment was performed using uracil-containing templates (with template strand radiolabeled). The full DNA sequences of the templates can be found in Table 3, Appendix A.

$KMnO_4$-sensitive bands between positions -8 to +3 of the uracil-containing template were observed (Fig. 14 A, lanes 6 on the left and right panels), delineating a transcription bubble. No $KMnO_4$ sensitivity was observed in reactions with the thymine-containing template (Fig. 14 A, lanes 8 on the left and right panel). Introduction of non-consensus G at the position -9 abrogates transcription (Fig. 13) and also abolished promoter melting on uracil-containing template (Fig. 14 B, lane 4 in comparison to the lane 2).

3.1.5 The nature of uracil requirement by AR9 nvRNAP

To further investigate the uracil requirement for AR9 nvRNAP transcription we designed a set of double-stranded DNA templates based on the P007 late promoter with uracils and thymines at different positions, and tested them in a multiple-round run-off transcription assay. As expected, the nvRNAP did not transcribe the thymine-only template but efficiently transcribed from uracil-only template (Fig. 15, lanes 1 and 2, respectively).

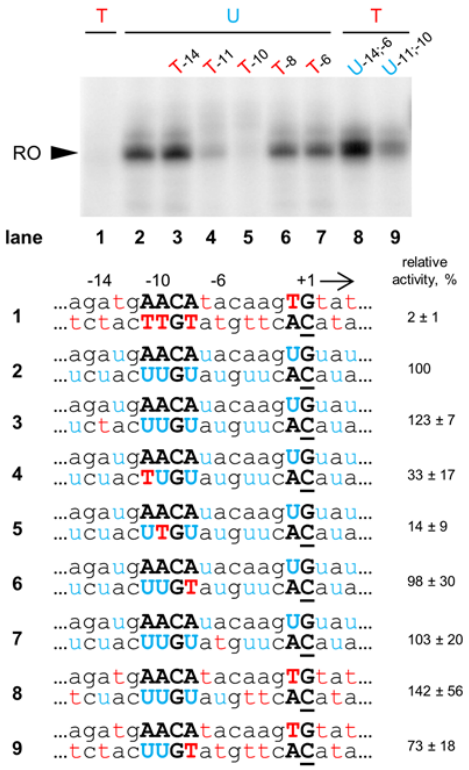


Figure 15. In vitro run-off transcription by AR9 nvRNAP of double-stranded P007 promoter templates carrying uracils and thymines at different positions.

"RO" - run-off transcripts (18 nt). The numbers under the gel indicate transcription activities relative to uracil-only control template. Below, DNA sequences around the TSS of the DNA templates used in the experiment are shown. Uracils and thymines are highlighted in blue and red, respectively. The position of the +1 start site is underlined. Conserved nucleotides of the late promoter are shown in capital bold letters. Average values and standard deviations from three independent experiments are presented. The full DNA sequences of the templates can be found in Table 4, Appendix A.

Introduction of single thymines at the -11 and -10 positions in the template strand of the consensus element 5'-$A^{-11}ACA^{-8}$-3' led to dramatic decrease in transcription (Fig. 15, lanes 4 and 5, respectively) while thymines at the -14, -8, and -6 positions had little or no effect (Fig. 15, lanes 3, 6, and 7, respectively). Transcription of the thymine-containing template with uracils at positions -14, -11, -10, -8, and -6 of the template strand was even more efficient than transcription of uracil-only template (Fig. 15, lane 8). The nvRNAP also transcribed from a thymine-containing template with uracils at the -11

and -10 positions (Fig. 15, lane 9). Therefore, we conclude that the presence of uracils instead of thymines at the -11 and -10 positions of the template strand is both necessary and sufficient for promoter specific transcription by AR9 nvRNAP; the presence of neighboring uracils increases transcriptional activity.

3.1.6 Template strand recognition by AR9 nvRNAP

To investigate which strand of the promoter DNA is recognized by the AR9 nvRNAP we designed fork-junction templates based on the P007 and P077 promoters where parts of either template strand or non-template strand were absent, while the transcribed part was double-stranded (Fig. 16). The AR9 nvRNAP transcribed from templates without the non-template strand with same efficiency as from the fully double-stranded templates (Fig. 16, lanes 3 and 1, respectively). No transcription from templates with missing template strand of promoters was detected (Fig 16, lanes 2). Transcription from the partially double-stranded templates was abolished when thymines were introduced instead of uracils in the consensus positions (Fig. 16, lanes 4). Thus, the AR9 nvRNAP specifically recognizes single-stranded late promoter consensus motif in the template strand (3'-$U^{-11}UGU^{-8}$-5').

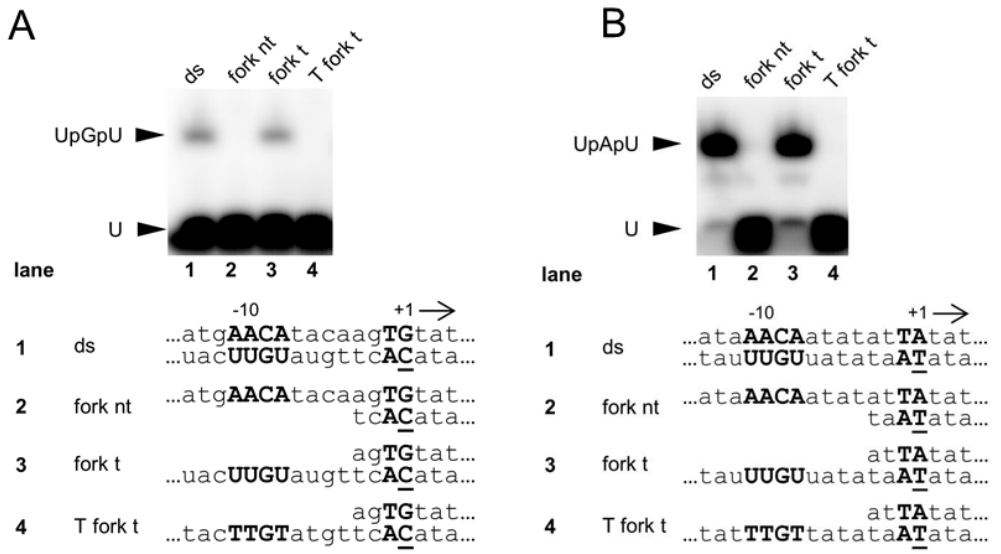


Figure 16. Analysis of the strand requirement for promoter recognition by AR9 nvRNAP.

The results of in vitro abortive initiation reactions by AR9 nvRNAP from the double-stranded and fork-junction P007-based (A) and P077-based (B) DNA templates shown below. RNA dinucleotides UpG (A) and UpA (B) were used to initiate transcription. The full DNA sequences of the templates can be found in Table 4, Appendix A.

3.1.7 Promoter specific transcription by AR9 nvRNAP from single-stranded DNA

The fact that nvRNAP recognizes the promoter consensus element in single-stranded form and in the template strand suggested that the enzyme may be capable of specific transcription of single-stranded DNA. Indeed, we observed robust multiple-round transcription by AR9 nvRNAP from single-stranded P007 promoter template containing uracil and no transcription from thymine-only template (Fig. 17, lanes 2 and 1, respectively). Introduction of thymines at the -11 and -10 positions strongly inhibited transcription from single-stranded templates containing uracils in other positions (Fig. 17, lanes 4 and 5, respectively). Introduction of thymines in several randomly chosen non-consensus positions or consensus position -8 had small or no inhibitory effect. As was the case with the double-stranded templates, introduction of uracils at the -11 and -10

positions was sufficient to allow transcription from a single-stranded template containing thymines in all other positions (Fig. 17, lane 9). Thus, the AR9 nvRNAP requirement for uracils in single-stranded and double-stranded promoters is the same.

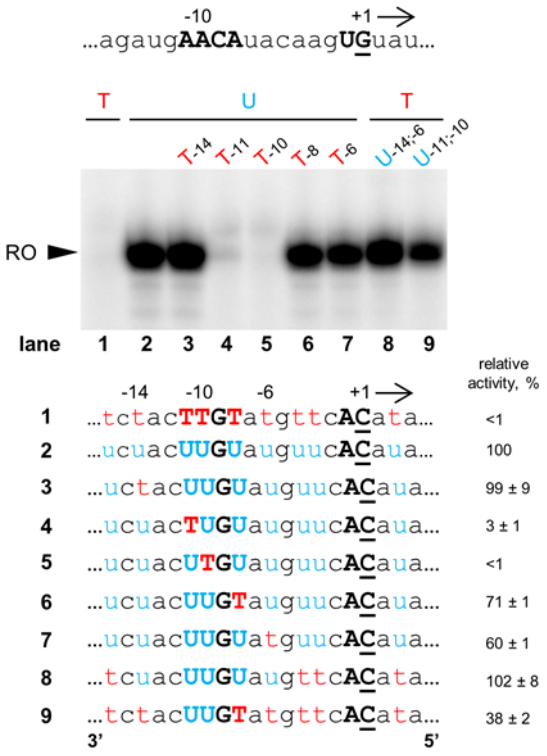


Figure 17. Specific transcription initiation by AR9 nvRNAP using single-stranded promoter DNA templates.

In vitro run-off transcription assay of AR9 nvRNAP using the single-stranded DNA templates matching the template strand of the P007 promoter carrying uracils and thymines at different positions. "RO" - run-off transcripts (18 nt). The nucleotide sequence of the non-template strand at and around the TSS of the P007 promoter DNA is shown at the top. Below the gel, DNA sequences around the TSS of the DNA templates used in the experiment are shown. The full DNA sequences of the templates can be found in Table 4, Appendix A.

Mutational analysis of the promoter consensus element in the context of single-stranded DNA was also performed (Fig. 18). While the consensus requirement appeared less strong for single-stranded DNA transcription than for double-stranded DNA

transcription, nevertheless, a common pattern of important positions was observed in both cases.

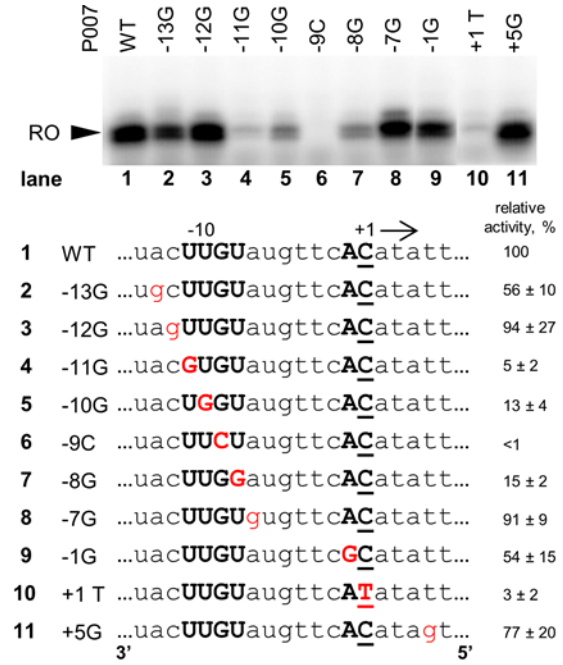


Figure 18. Late promoter consensus analysis in transcription from ssDNA.

In vitro run-off transcription assay of AR9 nvRNAP using single-stranded templates based on the P007 promoter DNA and its derivatives. Average values and standard deviations from three independent experiments are presented. The full sequences of the templates can be found in Table 5, Appendix A.

3.1.8 Characterization of AR9 nvRNAP-promoter complex formed on partially single-stranded DNA

Footprinting analysis of the AR9 nvRNAP-promoter complexes formed on dsDNA showed $KMnO_4$-sensitive bands between positions -8 and +3 of uracil-containing DNA template establishing that these positions are located within the transcription bubble. The U^{-11} and U^{-10} bases crucial for promoter recognition were resistant to $KMnO_4$ oxidation and thus appear to be outside the melted region. Alternatively, U^{-11} and U^{-10} could be part of the transcription bubble but remain resistant to chemical modification

due to their tight interactions with the AR9 nvRNAP. To differentiate between these possibilities we performed $KMnO_4$ probing of the AR9 nvRNAP-promoter complex formed on a fork DNA template with a partially absent non-template DNA strand. In free DNA, uracils of the template strand located in the single-stranded region are expected to be reactive in such experiment, while the addition of the AR9 nvRNAP could decrease the reactivity of some of the bases. Indeed, U^{-11} and U^{-10} appeared to be less reactive upon the AR9 nvRNAP addition (Fig. 19). We conclude that U^{-11} and U^{-10} interact with and/or are shielded by the AR9 nvRNAP. Conversely, U^{-8} became more reactive in the presence of AR9 nvRNAP. This may indicate increased exposure, due, for example, to a disruption of stacking interactions between the U^{-8} base and its neighboring bases caused by DNA bending near this position in the AR9 nvRNAP-promoter complex. Signals from all other uracils of the DNA template were not changed in the presence of the enzyme.

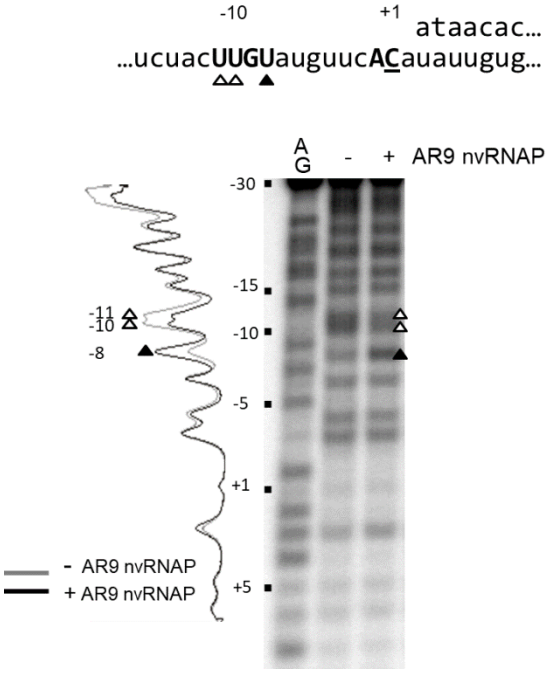


Figure 19. $KMnO_4$ probing of the AR9 nvRNAP-promoter complex formed on a fork DNA template.

A fragment of DNA template containing the P007 promoter used in the assay is shown at the top. Positions relative to the TSS (+1) are indicated. Lane indicated as "AG" is a marker lane (cleavage at purines). A densitogram of the signal intensities from the gel is shown on the left. The U^{-11} and U^{-10} bases that undergo protection upon the AR9 nvRNAP addition are marked by white triangles. U^{-8}, a base that becomes more reactive in promoter complex is marked by a black triangle. The experiment was performed using the promoter substrate radiolabeled at the template strand.

3.1.9 AR9 nvRNAP lacking gp226 subunit is catalytically active but unable to initiate transcription from promoters

At the last step of AR9 nvRNAP purification we obtained a minor fraction that contained trace amounts of gp226, the AR9 nvRNAP subunit with unknown function (Fig. 20 A, bottom). This finding is in agreement with the earlier observation that P53, the likely counterpart of gp226, is dissociable from the PBS2 RNAP [109, 110]. We compared AR9 nvRNAP lacking gp226 with the five-subunit form in *in vitro* transcription reactions from an RNA/DNA scaffold and promoter-containing templates (Fig. 20 B). AR9 nvRNAP lacking gp226 extended the RNA primer from the RNA/DNA scaffold with the same efficiency as the five-subunit enzyme but was unable to transcribe promoter-containing templates. No binding or melting of the promoter-containing template by AR9 nvRNAP lacking gp226 was observed when it was tested in footprinting experiments (Fig. 20 C). Thus, we conclude that the four-subunit AR9 nvRNAP form composed of the β/β' bacterial homologs but lacking gp226 is catalytically active but unable to bind to promoters.

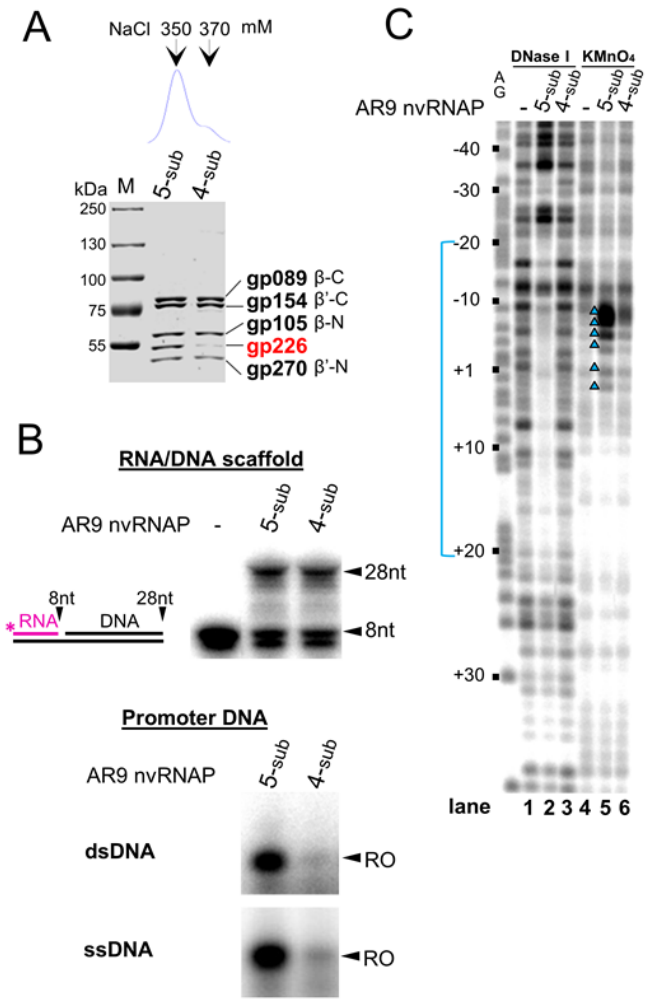


Figure 20. Functional analysis of the two forms of AR9 nvRNAP.

(A) Atop: chromatographic profile of AR9 nvRNAP eluted from a MonoQ column with a NaCl concentration gradient. Below: a Coomassie-stained SDS gel of the MonoQ fractions containing five-subunit (5-sub) and four-subunit (4-sub) forms of AR9 nvRNAP. (B) Comparison of transcriptional activities of the 5s and 4s nvRNAP. Top panel: RNA extension assay using the RNA/DNA scaffold schematically shown on the left. Below: in vitro run-off transcription from the uracil-containing promoter P007 in double- and single-stranded DNA (dsDNA and ssDNA). "RO" - a run-off transcript (62 nt for dsDNA and 18 nt for ssDNA). (C) DNase I footprinting and $KMnO_4$ probing of nvRNAP–promoter P077 complexes formed by 5s and 4s nvRNAP. The experiment was performed using the uracil-containing template (with template strand radiolabeled).

3.1.10 RNA transcript displacement from RNA-DNA hybrid during transcription of ssDNA

All multisubunit RNAPs studied to date do not displace RNA from extended RNA-DNA hybrids formed during transcription of ssDNA (see Chapter 2, section 2.2.3). To test the ability of the AR9 nvRNAP to displace RNA, we performed transcription from dsDNA and matching ssDNA template at conditions that support multiple rounds of transcription, and treated completed reactions with RNase H, the enzyme which degrades RNA in RNA-DNA hybrids but not free RNA. As expected, RNA transcript was fully resistant to RNase H treatment in transcription reactions from dsDNA, indicating its proper displacement from the RNA-DNA hybrid (Fig. 21 A, lane 2). In contrast, and also as expected, the most RNA products synthesized from the ssDNA template was sensitive to RNase H, indicating that it was hybridized to DNA (Fig. 21 A, lane 4). However, a small amount of RNA product of transcription from ssDNA was RNase H-resistant (Fig. 21 A, lane 4). We hypothesized that the RNase H-sensitivity of the most RNA products synthesized from ssDNA template could have resulted due to post-transcriptional re-annealing of released transcripts to ssDNA template. To address this possibility, we performed transcription reactions in the presence of increased or decreased concentrations of ssDNA template, while keeping the AR9 nvRNAP amounts constant (50 nM). Decreasing the DNA concentration was expected to decrease the RNA re-annealing efficiency and thus increase the fraction of RNase H-resistant material. Indeed, we found that in reactions with decreased ssDNA concentrations, a percentage of RNA that was resistant to RNAse H treatment was significantly increased (up to 40 %, Fig. 21 B, C). Therefore, we conclude that at least at certain conditions RNA transcripts generated by AR9 nvRNAP are separated from ssDNA template.

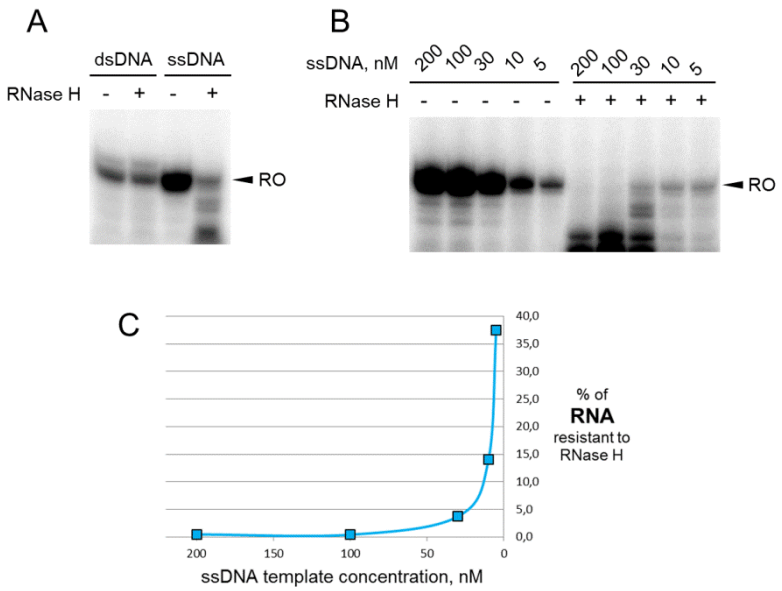


Figure 21. RNA displacement by AR9 nvRNAP.

(A) Effect of RNAase H treatment of RNA synthesized by AR9 nvRNAP during transcription from dsDNA and ssDNA containing the P007 promoter. "RO" - run-off transcripts (18 nt) (B) Effect of varying the concentration of DNA template on the RNase H sensitivity of RNA synthesized by AR9 nvRNAP. The concentrations of ssDNA in reactions are indicated. (C) Percentage of RNA transcript resistant to RNase H treatment as a function of concentration of ssDNA added in the reactions.

3.2 Discussion

In this Chapter, we described purification and functional characterization of nvRNAP – one of the two RNAPs encoded by giant bacteriophage AR9. Since AR9 nvRNAP subunits are the products of early phage genes, the enzyme was expected to transcribe late phage genes from promoters characterized by the presence of a 5'-A^{-11}ACA-(6N)-UA/G^{+1}-3' consensus sequence revealed by the dRNA-seq analysis of infected cells [102]. This expectation was fulfilled, however, the enzyme was also found to strictly require the presence of uracils instead of thymines in the template strand positions -11 and -10 of promoter DNA (Fig. 15). The presence of thymines in other

positions of the promoter or transcribed DNA has little or no effect on AR9 nvRNAP transcription.

The AR9 phage possesses a double-stranded DNA genome with uracil in place of thymine [88]. While unusual nucleotides are generally thought of as a strategy to overcome host defenses by restriction-modification systems [112], they can also help specific transcription of viral genes. *Bacillus* SP01 phage genome contains hydroxymethyl uracil instead of thymine [113]. SP01 utilizes host RNAP core bound to the phage-encoded σ factor, gp28, for transcription of its middle genes that was shown to be dependent on the presence of the modified nucleotides in the phage middle promoters [113]. In the case of T4 bacteriophage, whose genome contains hydroxymethyl cytosine instead of cytosine, the phage-encoded transcription terminator factor Alc terminates transcription elongation from cytosine-containing host DNA, while transcription of viral DNA is unaffected [114]. The requirement for uracils in promoter consensus element is an elegant strategy that should allow AR9 to avoid unnecessary transcription from host DNA, which contains multiple matches to the simple consensus of phage late promoter.

The requirement for uracils in the template strand of phage late promoter consensus element suggested that the nvRNAP recognizes the template strand. This hypothesis was confirmed by the analysis of transcription from fork-junction DNA templates (Fig. 16). Moreover, it was found that the AR9 nvRNAP specifically transcribes from the single-stranded template containing a reverse complement of the 5'-A^{-11}ACA-(6N)-UA/G^{+1}-3' late promoter consensus, provided that uracils are present in positions -11 and -10 (Fig. 17). The ability of the AR9 nvRNAP for promoter-specific transcription of single-stranded DNA is, to our knowledge, unprecedented for a multisubunit RNAP. The AR9 DNA is very AU-rich (72.25 %) and may be present in

partially single-stranded form in infected cells, especially during phage DNA replication. The unique properties of phage nvRNAP may allow utilizing such partially single-stranded DNA for specific transcription of late genes.

Although we do not know yet the biological role of promoter-specific transcription of ssDNA by AR9 nvRNAP, it is clear that *in vivo,* transcription of ssDNA makes sense only if the RNA transcript is separated from the DNA template somehow; otherwise it will remain unavailable for subsequent translation by ribosomes. The non-template DNA strand was shown to play a crucial role in RNA displacement from RNA-DNA hybrid in transcription by multisubunit RNAPs raising a question if the AR9 nvRNAP is able to separate RNA in transcription from ssDNA. We showed that in *in vitro* transcription reactions by AR9 nvRNAP, a significant amount of the RNA transcript is in a free form when the concentration of ssDNA template is low (Fig. 21) indicating the ability of AR9 nvRNAP to separate RNA from ssDNA, i.e., in the absence of non-template strand. An interesting feature of transcription by the AR9 nvRNAP is that potentially the RNA transcript may be separated from the ssDNA template in two ways. The first way is direct displacement of the nascent RNA from the RNA-DNA hybrid achieved through an unknown structural feature that should distinguish the AR9 nvRNAP from all other multisubunit RNAPs. The second scenario is less unusual from the structural side, yet would be a unique transcriptional strategy. Transcription of ssDNA may lead to formation of RNA-DNA fork-junction template, which is the same as the DNA-DNA fork-junction templates efficiently transcribed by AR9 nvRNAP but for the presence of an RNA rather than DNA strand as a non-template strand downstream the TSS. We speculate that in this case the RNA transcript may be displaced from the RNA-DNA hybrid not during its synthesis but in a subsequent round of the transcription cycle.

We envision that during transcription from the RNA-DNA fork junction template the new nascent RNA displaces the annealed RNA transcript generated during the previous transcription cycle. This hypothesis is under investigation at the time of this writing.

The fact that AR9 nvRNAP can recognize its promoter consensus element in single-stranded form suggests that during transcription initiation from dsDNA, the consensus element is also recognized in a single-stranded form and then stabilized to ensure transcription bubble maintenance. This resembles the promoter melting strategy utilized by σ^{70}-RNAP holoenzymes with remarkable difference in that the σ^{70}-RNAP holoenzymes bind sequence specifically to the non-template strand of the -10 element while the AR9 nvRNAP requires only the template DNA strand. Two uracil residues in the -10 and -11 positions of AR9 late promoters may be specifically recognized by AR9 nvRNAP similarly to A^{-11} and T^{-7} bases of the -10 element which are flipped and bound by the σ_2 domain of σ factors in RNAP holoenzymes (C^{-10} for σ^E-RNAP holoenzyme). The fact that AR9 nvRNAP does not bind and does not melt thymine-only late promoter templates suggests that 5-methyl groups of thymines at positions -11 and -10 of the template strand must interfere with the recognition by AR9 nvRNAP. Additionally, we found that U^{-11} and U^{-10} bases become protected from modification by $KMnO_4$ in AR9 nvRNAP-promoter complexes indicating their interaction with the enzyme (Fig. 19). The low sensitivity to $KMnO_4$ treatment was previously observed for single-stranded T^{-7} of the -10 promoter element due to its interaction with the σ^{70}-RNAP holoenzyme in promoter complex [115].

AR9 nvRNAP lacking gp226 is unable to bind promoter DNA but is catalytically active. Therefore, gp226 may be directly responsible for promoter recognition. We have found a limited similarity of gp226 (~ aa170-aa255) with the σ_2 domain of bacterial σ-

factors belonging to the σ^{70} family using HHpred program (Appendix B) [116, 117]. The σ_2 domain is the most conserved part of the σ^{70} family proteins that is involved in core binding and -10 promoter element recognition [14, 62, 76]. Binding to the RNAP core proceeds through a conserved coiled-coil structure in the largest (β') subunit [37]. HHpred and sequence analysis indicates that the corresponding structure should be present in the AR9 gp270, a homolog of the N-terminal part of bacterial RNAP β' subunit (Appendix B). Unlike RNAP holoenzymes containing σ^{70} family proteins, the AR9 nvRNAP recognizes the template strand of promoter DNA. Thus, the function of sequence-specific recognition of promoters must reside in a region of gp226 with no homology to proteins of known function. The promoter recognition region of gp226 may occupy the space of the σ-finger, lying along the template DNA strand in structures of bacterial RNAP holoenzymes.

Homologs of gp226 are encoded in genomes of all giant phages that have genes coding for β/β'-like proteins, suggesting that all nvRNAPs to some extent may share a common mechanism of promoter recognition. Thus, gp226 and its homologs may constitute a new class of transcription initiation factors. However, these proteins may be functionally diverse since many giant phages genomes have thymine in their DNA.

Footprinting experiments show that AR9 nvRNAP complexes on uracil-containing double-stranded templates appear to be similar to bacterial RNAP open complexes, at least as judged from the extent of promoter DNA protection from DNase I digestion and the size and position of transcription bubble. The presence of DNase I hypersensitive sites located with ~10 bp periodicity suggests that upstream DNA is wound around the AR9 nvRNAP, similarly to the situation in open promoter complexes formed by bacterial RNAP [76, 118]. However, in bacterial RNAP, the upstream DNA

contacts are accomplished by the dimer of α subunits, which are absent from the AR9 nvRNAP.

Among further research directions, an investigation of elongation and termination stages by the AR9 nvRNAP and studying the *in vivo* role of the unusual properties of the AR9 nvRNAP appears to be a promising avenue.

Chapter 4. Determination of the AR9 nvRNAP structure

The work described in this Chapter has been performed in Petr Leiman's laboratory at the University of Texas Medical Branch, USA (https://scsb.utmb.edu/labgroups/leiman/) as part of Skoltech academic mobility program. The author performed all biochemical experiments and prepared all protein crystals for X-ray data collection. The X-ray data were collected by Petr Leiman, Mark White and Michel Plattner. All processing of the X-ray data and all calculations were performed by Petr Leiman. Collecting and processing of the Cryo-EM data were performed by Alec Fraser supervised by Petr Leiman.

4.1 Results

Structural studies of proteins revolutionized molecular biology. To understand the molecular mechanisms underlying unique properties of the AR9 nvRNAP we set the ambitious goal to determine the 3D molecular structure of the enzyme. Currently, there are two major methods suitable for structure determination of protein complexes of such size: a) X-ray crystallography, a classical method, which has been applied for protein structure determination for more than sixty years and b) cryo-electron microscopy (Cryo-EM), which was developed much later and, due to recent technical advances, is taking the field by the storm. This Subchapter starts with description of our crystallographic endeavors directed towards the AR9 nvRNAP structure determination and ends with results obtained using Cryo-EM. This Subchapter also contains short notes on the theory of X-ray crystallography method to engage readers who are not closely acquainted with the method.

4.1.1 Recombinant AR9 nvRNAP

The yield of pure AR9 nvRNAP from one liter of *B. subtilis* culture infected with the AR9 phage was about five micrograms. Such a tiny yield precludes crystallization trials, which require several milligrams of highly pure protein just to find conditions suitable for further optimization. It was shown previously that active bacterial RNAP can be purified from *E. coli* co-overexpressing genes coding for the RNAP subunits [119]. Therefore, we decided to clone AR9 nvRNAP genes for co-overexpression and purify recombinant enzyme from the *E. coli* surrogate host. Below, we will refer to the four- and five-subunit versions of AR9 nvRNAP as nvRNAP core and holoenzyme, correspondingly. We created two plasmids: one encoding AR9 nvRNAP core enzyme (β/β'-like subunits genes only) and another one encoding AR9 nvRNAP holoenzyme (β/β'-like subunits and the gp226 genes). Plasmids were constructed using synthetic gene-Blocks containing AR9 nvRNAP gene fragments with codons optimized for expression in *E. coli* to enable high levels of target proteins synthesis (see Materials and Methods for details). Each RNAP gene was preceded with T7 RNAP promoter and *lac* operator. In both plasmids, the AR9 nvRNAP subunit gp270 (a homolog of the N-terminal part of bacterial RNAP β' subunit) was N-terminally fused to a hexahistidine tag.

The AR9 nvRNAP core and holoenzyme were purified from *E. coli* cultures transformed with created plasmids using Ni-NTA affinity chromatography and subsequent anion exchange chromatography and gel-filtration (Fig. 22 A). The recombinant AR9 nvRNAP core and holoenzyme were shown to be active on RNA/DNA scaffold and promoter-containing DNA templates, correspondingly (Fig. 22 B, C). The yield of the recombinant AR9 nvRNAPs (both core and holoenzyme) was about 10

milligrams from one liter of induced *E. coli* culture. The resulting samples were pure enough to initiate crystallization trials.

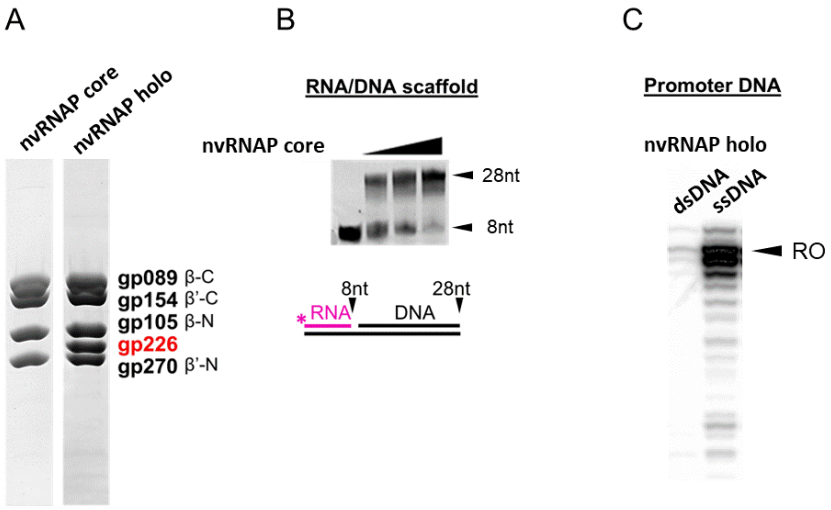


Figure 22. Recombinant AR9 nvRNAP.

(A) A coomassie-stained SDS gel of the MonoQ fractions containing AR9 nvRNAP core and holoenzyme correspondingly. (B) Analysis of transcriptional activity of the AR9 nvRNAP core in RNA extension assay using the RNA/DNA scaffold schematically shown below. (C) Analysis of transcriptional activity of the AR9 nvRNAP holoenzyme on promoter containing dsDNA and ssDNA templates. "RO" - run-off transcripts (18 nt).

4.1.2 Crystallization of the AR9 nvRNAP core enzyme

We performed large-scale screening of crystallization conditions for both AR9 nvRNAP core and AR9 nvRNAP holoenzyme. Crystal growth was only observed with the AR9 nvRNAP core enzyme. Next, manual optimization of promising crystal growth conditions was performed. In optimized conditions small needle-like crystals were seen 30 minutes after setting up crystallization drops, while overnight crystals had a size suitable for diffraction data collection. Even larger crystals appeared in a few days (Fig. 23). Diffraction data were collected for several crystals at LS-CAT synchrotron. The diffraction pattern had different quality in different directions (anisotropic diffraction). The best dataset extended to 3.5 Å resolution (statistics for the dataset are shown in

Appendix C-1). Crystals had a large orthorhombic unit cell (parameters of the asymmetric unit: a = 112Å b = 163Å c = 306Å α=β=γ=90°) with two RNAP molecules per asymmetric unit according to Matthews coefficient probabilities [120] (the table with Matthews coefficients is shown in Appendix D-1).

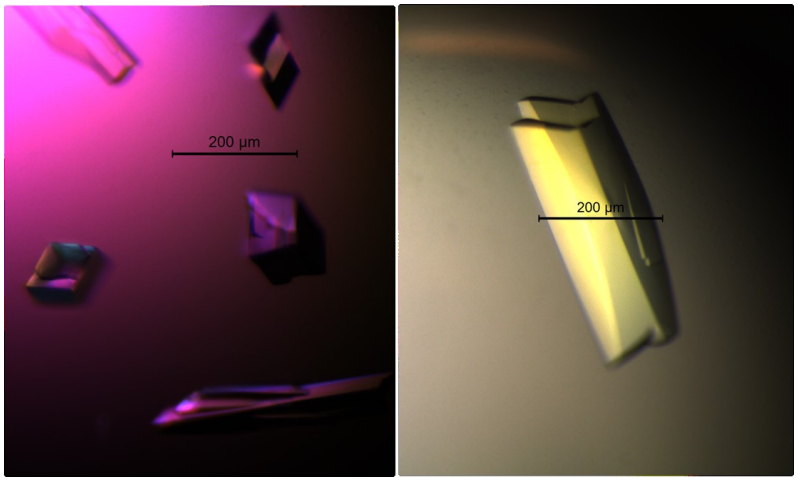


Figure 23. Crystals of the AR9 nvRNAP core.

Since only amplitudes of scattered X-rays are recorded by the detector during diffraction data collection, one has to find phases to determine the structure. The easiest method to find crystallographic phases is the molecular replacement (MR) method, where a known structure with sufficient level of amino acid identity is used to predict phases of X-rays scattered in the experiment [121]. The AR9 nvRNAP is too different from all RNAPs with known structures; therefore, we could not find the solution of the phase problem by MR.

Another technique used for finding crystallographic phases is based on the introduction of heavy atoms (preparation of heavy-atom derivatized crystals) [122]. If there are ordered heavy atoms within a crystal, then phases can be found either by comparison of datasets collected from 'native' (non-modified) crystals and derivative (modified) crystals, or by using the anomalous scattering behavior of certain heavy atoms

at or near their X-ray absorption edges [123]. The first approach includes single and multiple isomorphous replacement (SIR and MIR) methods; the second includes single and multiple anomalous dispersion (SAD and MAD) methods [123].

One of the most commonly used ways for preparation of heavy-atom derivative crystals is crystallization of a Selenomethionine (SeMet) version of the protein, which is produced by expressing the protein in methionine-auxotrophic cells in the presence of Se-methionine instead of the normal S-methionine [124]. A SeMet derivative of AR9 nvRNAP core enzyme was produced and crystallized in conditions similar to the native protein. However, the diffraction of the SeMet crystals extended to only 7 Å resolution or worse.

More than twenty compounds containing different heavy atoms were tried to obtain heavy-atom derivatives of AR9 nvRNAP core crystals, either by soaking of native crystals or co-crystallizing the protein in the presence of these compounds. The work was challenging since even diffraction from native crystals was not reproducible. Crystals were also very fragile and often dissolved spontaneously even without disturbing them with heavy-atom compounds. Overall, about 300 crystals of the AR9 nvRNAP core were analyzed at a synchrotron (in addition to those that were checked using an in-house X-ray source).

Some heavy-atom compounds destroyed diffraction without affecting crystal appearance, while others did not affect diffraction but also did not bind the protein in crystals in an ordered way (the anomalous signal was absent after processing the data).

Eventually, two heavy-atom derivatives suitable for phasing procedure were found. One derivative contained Tantalum Bromide Cluster (Ta_6Br_{12}) and diffracted to 6.5 Å only, while another one contained Thimerosal (2-$(C_2H_5HgS)C_6H_4CO_2Na$) and

diffracted to 3.8 Å in the best direction and to 4.5 Å in the worst. The phases were found separately for both datasets by the SAD method. The electron density map calculated from the Thimerosal derivative crystal allowed to see an overall shape of the AR9 nvRNAP core molecule (Fig. 24). Some α-helices are visible in the map, though the density is mostly discontinuous. Combining the data from two derivative crystals with data from the native crystal allowed us to solve the phase problem by the MIR method. However, the resulting electron density map was still discontinuous and was not enough to determine the structure of the enzyme.

The presence of several protein molecules per asymmetric unit enables one to use noncrystallographic symmetry (NCS) to average the electron density map in order to improve it. However, in the case of AR9 nvRNAP core crystals, one of the two RNAP molecules in asymmetric unit appeared to be disordered, so the averaging could not help to improve the map.

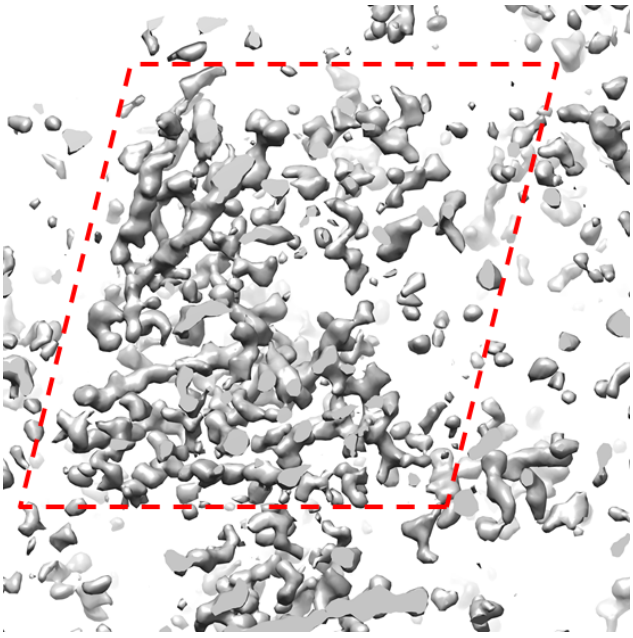


Figure 24. A fragment of the electron density map of AR9 nvRNAP core.

The phases were first calculated by the program SHARP (SAD of a Thimerosal derivative crystal) [125], and then improved by solvent flattening with the program PARROT [126]

and DM [127]. The map is contoured at 2.5 standard deviations above the mean. Shape of the RNAP molecule is highlighted by a dashed red line. The figure was prepared using the program UCSF Chimera [128].

We attempted to build a model based on the obtained electron density map. For this, we chose the structure of the *Sulfolobus shibatae* RNAP (PDB ID 4ayb). According to HHpred program, *S. shibatae* RNAP subunits have the highest coverage of the AR9 nvRNAP sequence among RNAPs with known structures. All fragments of archaeal RNAP that were absent from the AR9 nvRNAP according to HHpred were removed from the *S. shibatae* RNAP structure. Then the structure was converted to polyalanine model and fitted with the electron density map of the AR9 nvRNAP core. Several additional regions were removed from the model and other regions were adjusted to fit the density. A few short α-helices were found in the electron density map but were absent from the archaeal RNAP; they were added to the model manually. The resulting model corresponded to 30% of the total length of all AR9 nvRNAP core polypeptide chains. The model contained DPBB domains and surrounding α-helices (Fig. 25).

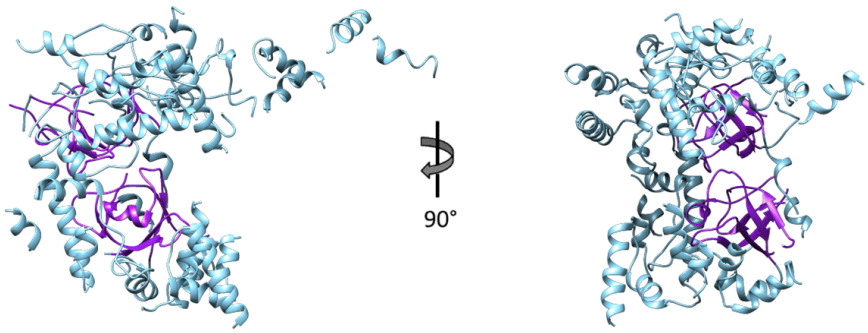


Figure 25. Model of the AR9 nvRNAP core.

The model was built based on the structure of S. shibatae RNAP (PDB ID 4ayb) with using an electron density map obtained from AR9 nvRNAP core crystals. DPBBs are colored in purple. The figure was prepared using the program UCSF Chimera [128].

4.1.3 Crystallization of AR9 nvRNAP core enzyme without the histidine tag

To improve diffraction of the AR9 nvRNAP core crystals we decided to crystallize the protein without the histidine tag since it is known that tags can affect crystallization condition and diffraction quality. To get the AR9 nvRNAP core without the tag ('tagless' protein) we created a new plasmid without a sequence coding the tag and purified the enzyme using the procedure developed for purification of AR9 nvRNAP from infected cells - Polymin P fractionation, heparin-sepharose affinity chromatography, and anion exchange chromatography.

Tagless AR9 nvRNAP core did not crystalize in conditions where AR9 nvRNAP with tag crystalized. We therefore performed new screening of crystallization conditions and found several hits, though not all of conditions could be reproduced during crystal growth optimization. Nevertheless, several crystals of tagless AR9 nvRNAP core (Fig. 26) were analyzed at synchrotron. Diffraction was observed from a few crystals and extended to 4Å only (statistics for the best dataset are shown in Appendix C-2). However, diffraction was much brighter and more isotropic in comparison to crystals of the RNAP carrying tag. We collected datasets from native and Thimerosal derivative crystals. Crystals of tagless AR9 nvRNAP core had a giant orthorhombic unit cell (parameters of the asymmetric unit: a = 173Å b = 234Å c = 594Å $\alpha=\beta=\gamma=90°$) with several RNAP molecules per asymmetric unit according to Matthews coefficient probabilities [120] (the table with Matthews coefficients is shown in Appendix D-2). The size of the asymmetric unit was comparable with that from crystals obtained from small virus particles [129, 130].

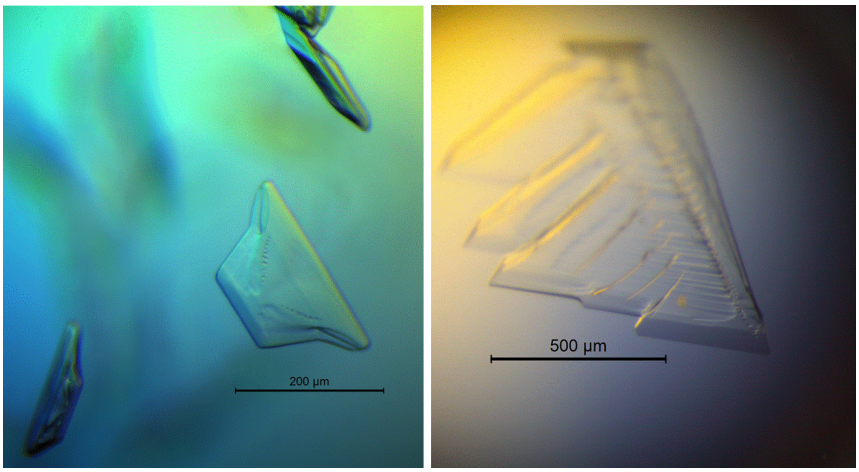


Figure 26. Crystals of tagless AR9 nvRNAP core.

4.1.4 Phase problem solution for crystals of tagless AR9 nvRNAP core

Though we had a dataset from a Thimerosal derivative crystal of the tagless AR9 nvRNAP core, we could not find phases by the SAD method. We tried MR method using the polyalanine model which was built previously based on the electron density map of the protein with histidine tag. This model allowed the Phaser program to find a solution (to find phases of the dataset). It appeared that the asymmetric unit of AR9 nvRNAP core crystals contained eight RNAP molecules (~ 2058 kDa). The polyalanine model that we used in MR had a molecular weight of 45 kDa thus constituting less than 3% of the entire atomic mass of the asymmetric unit. The fact that solution was found using such tiny model supports the high accuracy of the model.

All eight RNAP molecules found in the asymmetric unit of tagless AR9 nvRNAP core crystals were ordered allowing us to average the electron density map using NCS. As a result, we got a much better electron density map in comparison with that obtained from crystals of AR9 nvRNAP core with tag (Fig. 27).

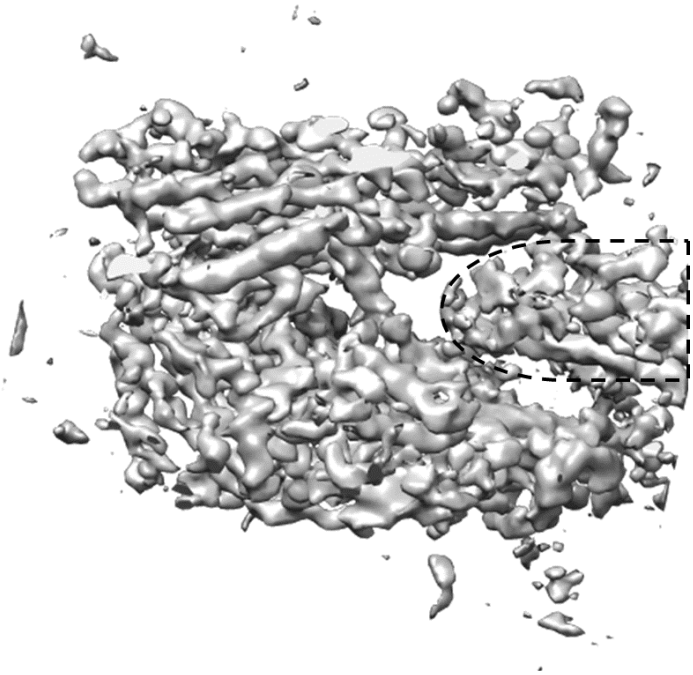


Figure 27. A fragment of the improved electron density map of AR9 nvRNAP core.

The map was obtained using diffraction data collected for a crystal of tagless AR9 nvRNAP. The phases were first calculated by the program PHASER [131] *using a significantly incomplete model derived from the data obtained previously (for AR9 nvRNAP core with the histidine tag); then the phases were improved by NCS averaging (eight molecules in the asymmetric unit) and solvent flattening with PARROT* [126] *and DM* [127]. *A part of a neighboring RNAP molecule is highlighted by a dashed black line. The map is contoured at 1.4 standard deviations above the mean. The figure was prepared using the program UCSF Chimera [128].*

There are still only a few side chains seen in the improved electron density map, thus precluding automatic building of the whole model required for the structure determination. New polyalanine model corresponding to 57% of the total length of all AR9 nvRNAP core polypeptide chains was therefore built manually (Fig. 28 A). Fragments of the electron density map of AR9 nvRNAP core corresponding to some elements common for all multisubunit RNAPs are shown in Figure 28 B and Appendix E. At the time of this writing we continue to improve the model of AR9 nvRNAP core.

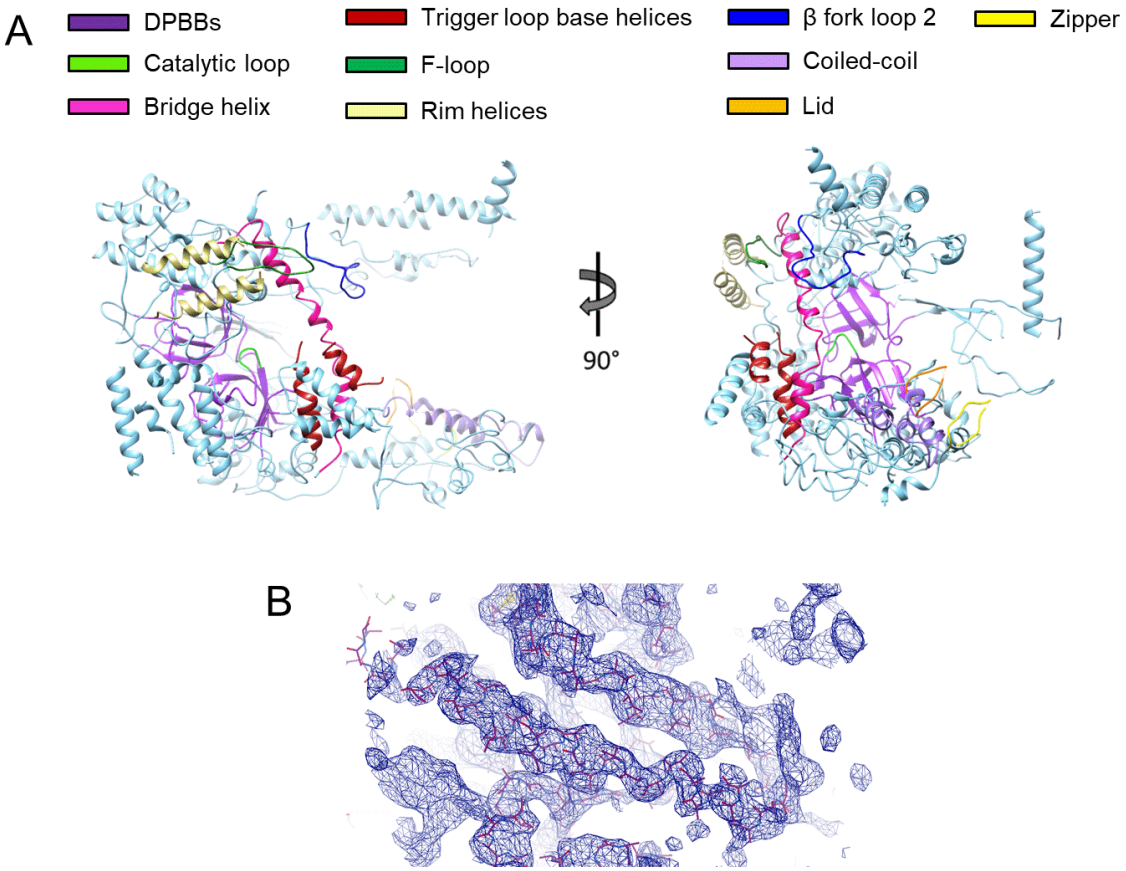


Figure 28. Improved model of the AR9 nvRNAP core.

(A) Model was built using the electron density map obtained for a crystal of tagless AR9 nvRNAP (shown in Fig. 27). Elements, common for all multisubunit RNAPs, that were identified in the model are colored according to the legend. The figure is prepared using the program UCSF Chimera [128] (B) A fragment of the electron density map used to build the model is shown (contoured at 1.3 standard deviations above the mean, for other details about the map see the legend for Fig. 27). The coiled-coil motif is at the first plan. For more fragments of the map see Appendix E. The figure was prepared using the program Coot [132].

4.1.6 Crystallization of AR9 nvRNAP holoenzyme in complex with promoter DNA

The fact that we were unable to get crystals of the AR9 nvRNAP holoenzyme could be caused by its high conformational mobility. In order to stabilize a particular conformation and improve homogeneity of the sample we decided to run crystallization

trial of the AR9 nvRNAP holoenzyme/promoter complex. We chose a fork junction template on which the AR9 nvRNAP was highly active in abortive synthesis (Fig. 29).

Figure 29. DNA template used to prepare AR9 nvRNAP holoenzyme/promoter complex prior to crystallization.

The position of the +1 start site is underlined. The conserved nucleotides of the late promoters are shown in capital bold letters. Uracils are highlighted in blue.

Crystallization conditions screening of preformed AR9 nvRNAP holoenzyme/promoter complex was performed. Four hits were found, however, among them only one was reproduced and gave crystals that were not suitable for data collection (Fig. 30, top). Manual optimization of crystal growth enormously improved crystals' appearance (Fig. 30, bottom). The crystals likely contained the AR9 nvRNAP holoenzyme in complex with the promoter DNA since the protein was not crystalized in the absence of the DNA.

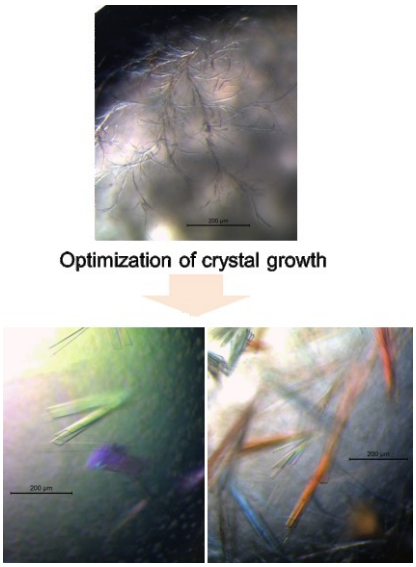


Figure 30. Crystals of the AR9 nvRNAP holoenzyme/promoter complex.
Crystals before optimization are shown at the top. Crystals after optimization are shown at the bottom.

Diffraction data from several native crystals of the AR9 nvRNAP holoenzyme/promoter complex were collected at synchrotron. The crystals had monoclinic unit cell (a = 175Å b = 110Å c = 190Å α=β=90° γ=109.35°) with only one RNAP molecule per asymmetric unit according to Matthews coefficient probabilities [120] (the table with Matthews coefficients is shown in Appendix D-3). The diffraction extended to 3.3Å (statistics for the best dataset are shown in Appendix C-3). The new crystals were less fragile and diffraction was observed more consistently than with previous crystals, most of which did not diffract for no apparent reason.

The collected datasets were solved by MR using previously built model of the AR9 nvRNAP core. The quality of the electron density map is not sufficient for automatic building of the model. At the time of this writing we plan to collect data from heavy-atom derivative crystals of the AR9 nvRNAP holoenzyme/promoter complex to get better crystallographic phases to improve the map.

4.1.7 Cryo electron microscopy with AR9 nvRNAP

While at the time of this writing we continue to work on structure determination by the X-ray diffraction method, we also initiated a Cryo-EM study. The first and major challenge with using of Cryo-EM for AR9 nvRNAP was that both AR9 nvRNAP core and holoenzyme got stuck to the carbon surface during grid preparation instead of falling into holes making data collection impossible. This problem was solved by using carbon grids with a continuous thin layer of carbon deposited upon grid holes. First images of the AR9 nvRNAP core and holoenzyme were collected at the JEOL 2200 FS cryo-electron microscope at the UTMB. The data contained about 84,000 images of single particles for RNAP holoenzyme and 45,000 images for RNAP core. A part of 2D classification results of images is shown in Appendix F. Luckily, the AR9 nvRNAP does not exhibit the often

faced preferred orientation problem during grid deposition. However, for structure determination, images of even more single particles are required. For this reason, new data were collected at a Krios Titan microscope, SLAC-Stanford. At the moment of this writing data processing is in progress. The estimated quantity of single particles in collected images is between 200,000 and 400,000. Such datasets should allow to determine the structure with a resolution in a range of 3.5-4.5Å.

4.2 Discussion

Many elements which are conserved among all multisubunit RNAPs are easily recognized in the model of the AR9 nvRNAP core currently at hand (Fig. 28 A, Appendix E). To compare relative positions of these elements between the AR9 nvRNAP and other multisubunit RNAPs we superimposed the model of the AR9 nvRNAP core with the structure of *T. aquaticus* RNAP core (PDB ID 1HQM [18]) (Fig. 31). The AR9 nvRNAP core in crystals adopts an open clamp conformation similar to the *T. aquaticus* RNAP core structure. Elements forming the secondary channel are clearly seen in the electron density map and thus are present in the model of AR9 nvRNAP core. These elements include rim helices, bridge helix, a part of the trigger loop and F-loop which is adjacent to the bridge helix. The bridge helix is in the kinked conformation and the trigger loop is partially unfolded like is also observed in many RNAP core structures without DNA. The β switch 3, β' clamp and β flap domains forming the RNA exit channel are also present in the AR9 nvRNAP core model (the flap domain is without the flap tip helix in the model).

The β lobes are present in the AR9 nvRNAP core model only partially. There is a difference in the position of β lobe 2 between the AR9 nvRNAP core model and *T. aquaticus* RNAP structure judging by the difference in location of its two edge α helices.

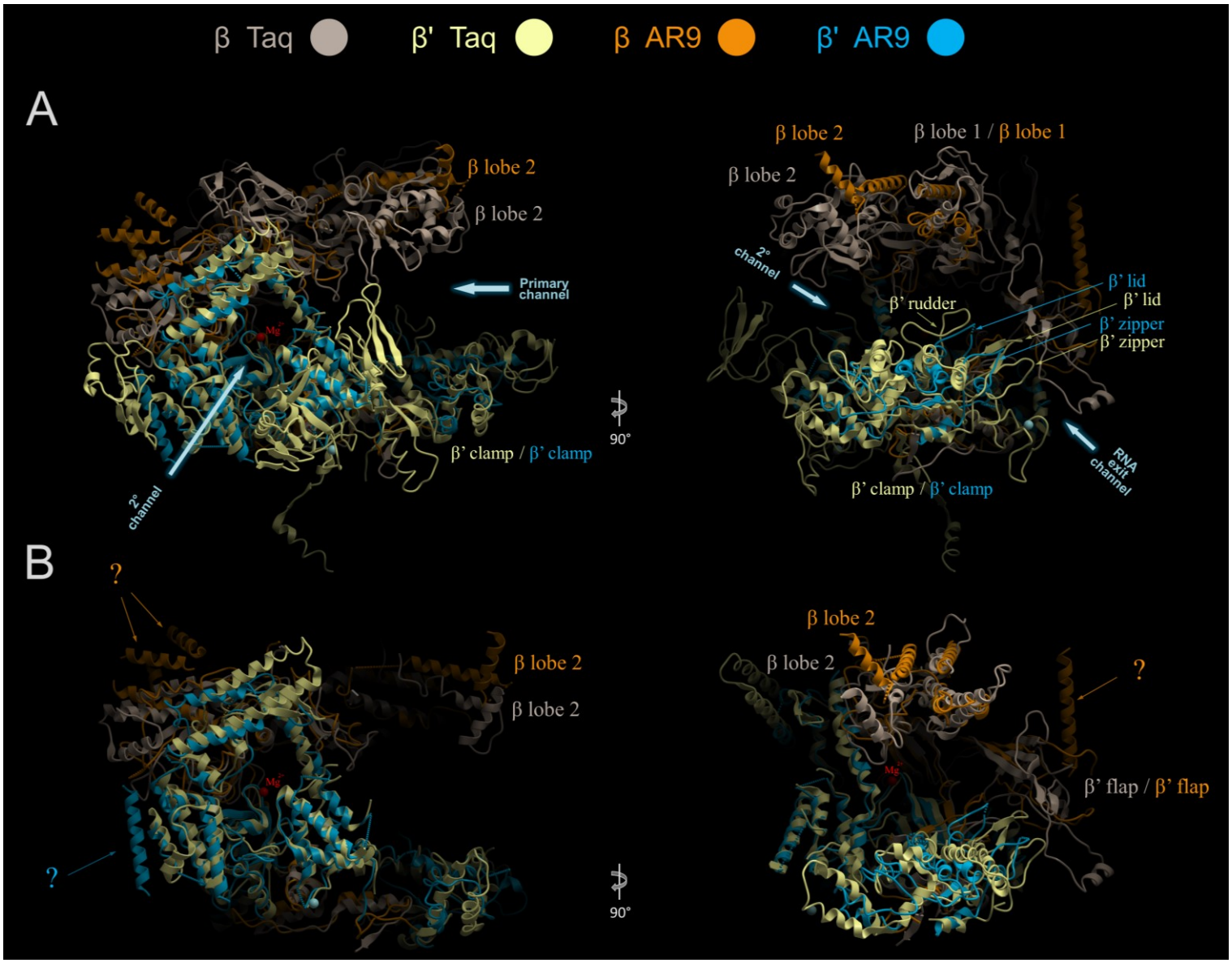


Figure 31. Comparison of the AR9 nvRNAP core model with T. aquaticus RNAP core structure.

(A) A superposition of the AR9 nvRNAP core model with the structure of T. aquaticus RNAP core (Taq) (PDB ID 1HQM [18]). The αI, αII and ω subunits of the Taq RNAP are omitted, the Taq nonconserved domain (NCD1) is also not shown. The β and β' subunits of Taq RNAP and β- and β'-like subunits of the AR9 nvRNAP are colored according to the legend. Those elements of Taq RNAP β and β', which are absent in the AR9 nvRNAP core model, are omitted in views shown in (B). Several α-helices present in the AR9 nvRNAP model but absent in Taq RNAP are indicated by arrows with question marks. The superposition and different views of superimposed structures used for the figure were kindly prepared by Sergey Borukhov. The figure was prepared using ICM-Browser.

Among the loop-like elements protruding into the main cleft of bacterial RNAP, β fork 2, β' lid and β' zipper are present in the AR9 nvRNAP core model, while the rudder is absent. The absence of the rudder in the AR9 nvRNAP was predicted bioinformatically (Appendix B, B) and is also undoubtedly seen in the electron density map of the AR9 nvRNAP core. The rudder is ubiquitously present in all canonical RNAPs and stabilizes the transcription bubble architecture in multisubunit RNAP/DNA complexes (see Chapter 1, section 1.2.3). Thus, its absence in the AR9 nvRNAP predicts altered transcription bubble architecture of the AR9 nvRNAP/DNA complexes and distinct mechanisms of its stabilization. The lid and zipper elements in the AR9 nvRNAP core model are significantly shifted, towards the position where the *T. aquaticus* rudder is located. As a result, the lid and flap domain of the AR9 nvRNAP are located much farther from each other in comparison to any other multisubunit RNAP structure where they are almost connected. This space may be occupied by one of the domains of gp226, which should allow the recognition of the template strand of promoter DNA.

There are several α-helices present in the model of AR9 nvRNAP core which are absent from the structure of *T. aquaticus* RNAP core (Fig. 31 B, indicated by question marks). The functions of these elements remain to be established.

At the time of this writing we continue to work on improvement of the AR9 nvRNAP core model and on the determination of the structure of AR9 nvRNAP holoenzyme/promoter complex.

Chapter 5. Conclusions

In this work we purified and functionally characterized a non-canonical multisubunit RNAP encoded by giant bacteriophage AR9 infecting *B. subtilis.* We also launched a large project on determination of the structure of this enzyme.

We showed that the catalytically active core of the AR9 nvRNAP consists of distant homologs of the β and β' subunits of bacterial RNAP but lacks any other proteins. Such arrangement, where catalytic subunits of a multisubunit RNAP constitute an active RNAP core in the absence of additional subunits is observed for the first time. Subunits forming the assembly platform likely were present in the multisubunit RNAP of LUCA since all cellular multisubunit RNAPs contain them. The absence of the assembly platform from the AR9 nvRNAP core likely means that derived from cellular enzymes nvRNAPs of giant phages must have evolved in a way that strengthened the interactions between the catalytic subunits which allowed the loss of the assembly platform. On the other hand, our findings support a hypothesis that ancient RNAP which existed before the LUCA could function without the assembly platform.

It is an interesting question how β- and β'-like subunits of the AR9 nvRNAP are kept together in the absence of additional subunits (especially considering that both AR9 nvRNAP counterparts of the β and β' are split in two separate polypeptides). In the model of the AR9 nvRNAP core which we built using X-ray crystallography data, several α helices with no structural analogues in other RNAPs are seen. When the complete structure of the AR9 nvRNAP core will be determined, we expect to see more structural features of the AR9 nvRNAP distinguishing it from canonical multisubunit RNAPs.

Some of these structural features should be responsible for AR9 nvRNAP core assembly in the absence of a specialized assembly platform.

Our work revealed that the AR9 nvRNAP holoenzyme containing the phage protein gp226 relies on a novel transcription initiation strategy. Promoter recognition by the AR9 nvRNAP strictly depends on the presence of uracils at certain positions of the template DNA strand of late phage promoters. The AR9 nvRNAP was found to be able for promoter specific transcription from single-stranded DNA. These properties are unprecedented for any multisubunit RNAP studied to date. We showed that the gp226 is responsible for promoter recognition by the AR9 nvRNAP. Gp226 has a very weak limited sequence similarity with bacterial σ factors of σ^{70} family. At the time of this writing we continue to work on the structure of AR9 nvRNAP holoenzyme/promoter complex which should reveal details of the new promoter recognition mechanism and evolutionarily relationships of gp226 to bacterial σ factors.

It is also worth mentioning that a system for obtaining recombinant multisubunit AR9 nvRNAP (both core and holoenzyme) not only allowed us to crystallize both protein complexes for X-ray analysis but is also an important step towards structure-functional analysis of the AR9 nvRNAP which is among our future plans.

In conclusion, work presented in this thesis provides the most accurate to date analysis of transcription initiation strategy utilized by a non-canonical multisubunit RNAP and for the first time gives information on the structure of a non-canonical transcription enzyme.

Appendix A

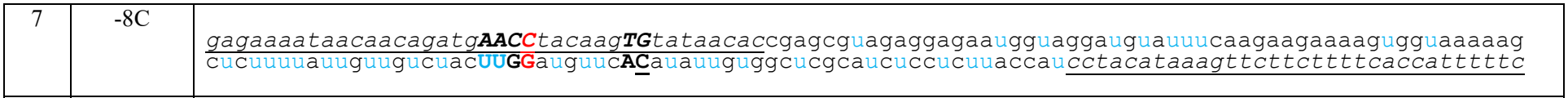
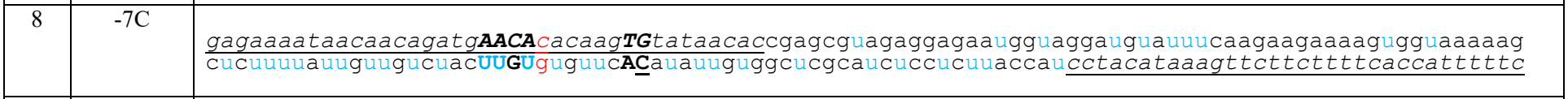
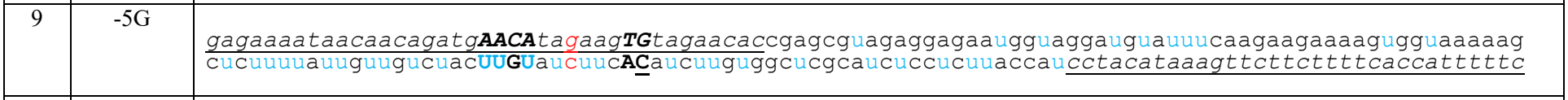
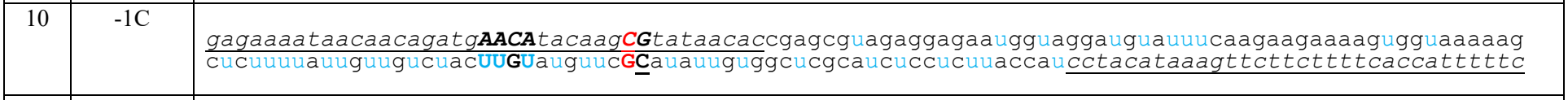
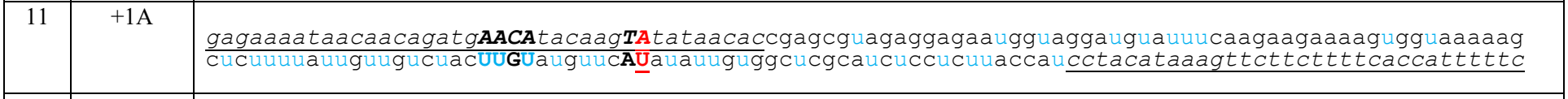
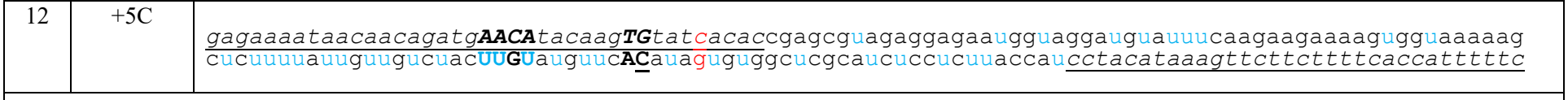
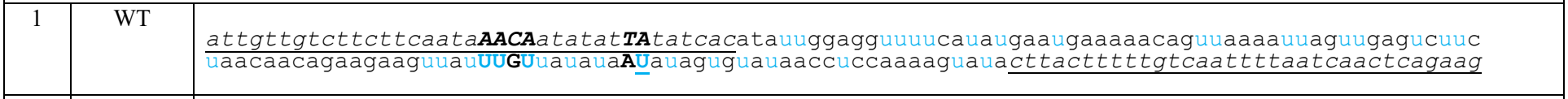
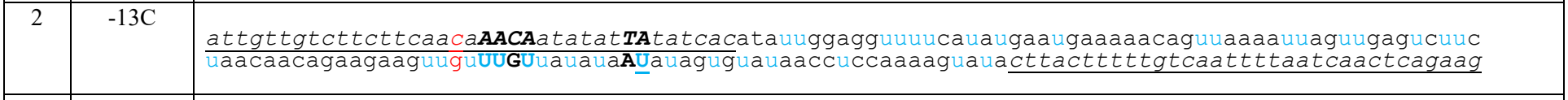
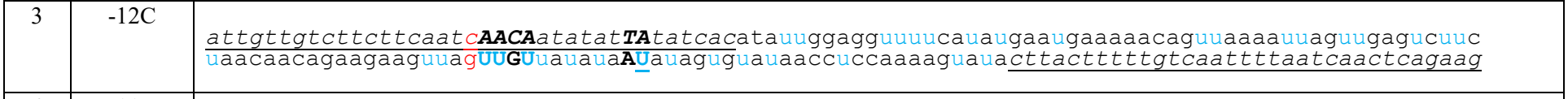
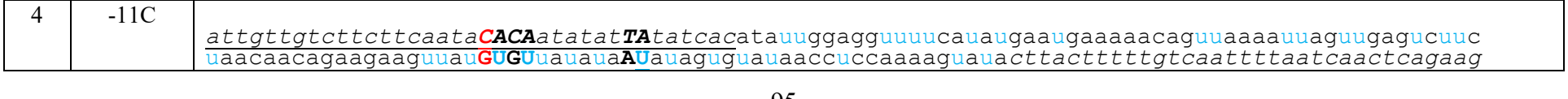
Table 1. Primers used to prepare DNA templates containing late AR9 promoters.

AR9 Promoter	**Primer 1** Sequence (5'-3' orientation)	**Primer 2** Sequence (5'-3' orientation) (primers used in primer extension and sequencing reactions marked by asterisk)
P007	aaataatttaggttaatattcc	tttcttcttgaaatacatcc *
P012	tcactatcaaccttgtcaattac	cttcaaaatttgcatttgttaataag
P020	agtaattcaatttcttcatatccc	agagccttcttcatgtatatc
P068	cactctccataaaaattc	agagtaattgatagcaatgac
P123	aagcctaaattttattactag	ttgcttcttgttcatccatc
P153	tggatatgaatttacataatttttag	gtgagtaagttgaagaatgatgacat *
P164	gattattccatgtattctcttc	caatattaagcaactcagtatc
P233	ggagttcactcttaaatatatc	cttattagaacaaaatggac *
P242	gaaatgaagcggcatctgag	gattattcatctttggtactaacat
P077	tctttctatgttccgaatacc	acttacacgaattgcttcatt *

Table 2. DNA templates with the P007 and P077 promoters and their mutant variants

Templates containing either the P007 or the P077 promoters and their variants were synthesized by PCR with dUTP. The sequences of the primers are italic and underlined. Conserved nucleotides of the promoters are shown in capital bold letters. The position of the +1 start site is underlined. Uracils are highlighted in blue. Single substitutions are highlighted in red.

№	Name	Sequence
		P007
1	WT	*gagaaaataacaacacagatg**AACA**tacaag**TG**tataacac*cgagcguagaggagaaugguaggauguauuucaagaagaaaagugguaaaaag cucuuuuauuguugucuac**UUGU**auguuc**AC**auauuguggcucgcaucuccucuuaccau*cctacataaagttcttcttttcaccatttttc*
2	-13C	*gagaaaataacaacagacg**AACA**tacaag**TG**tataacac*cgagcguagaggagaaugguaggauguauuucaagaagaaaagugguaaaaag cucuuuuauuguugucugc**UUGU**auguuc**AC**auauuguggcucgcaucuccucuuaccau*cctacataaagttcttcttttcaccatttttc*
3	-12C	*gagaaaataacaacagatc**AACA**tacaag**TG**tataacac*cgagcguagaggagaaugguaggauguauuucaagaagaaaagugguaaaaag cucuuuuauuguugucuag**UUGU**auguuc**AC**auauuguggcucgcaucuccucuuaccau*cctacataaagttcttcttttcaccatttttc*
4	-11C	*gagaaaataacaacagatg**CACA**tacaag**TG**tataacac*cgagcguagaggagaaugguaggauguauuucaagaagaaaagugguaaaaag cucuuuuauuguugucuacG**UGU**auguuc**AC**auauuguggcucgcaucuccucuuaccau*cctacataaagttcttcttttcaccatttttc*
5	-10C	*gagaaaataacaacagatg**ACCA**tacaag**TG**tataacac*cgagcguagaggagaaugguaggauguauuucaagaagaaaagugguaaaaag cucuuuuauuguugucuac**UGGU**auguuc**AC**auauuguggcucgcaucuccucuuaccau*cctacataaagttcttcttttcaccatttttc*
6	-9G	*gagaaaataacaacagatg**AAGA**tacaag**TG**tataacac*cgagcguagaggagaaugguaggauguauuucaagaagaaaagugguaaaaag cucuuuuauuguugucuac**UUCU**auguuc**AC**auauuguggcucgcaucuccucuuaccau*cctacataaagttcttcttttcaccatttttc*

7	-8C	
8	-7C	
9	-5G	
10	-1C	
11	+1A	
12	+5C	
P077		
1	WT	
2	-13C	
3	-12C	
4	-11C	

5	-10C	attgttgtcttcttcaataACCAatatatTATatcacatauuggaggguuuucauaugaaugaaaaacaguuaaaauuaguugagucuuc uaacaacagaagaaguuauUGGUuauauaAUauaguguauaaccuccaaaaguauacttacttttgtcaattttaatcaactcagaag
6	-9G	attgttgtcttcttcaataAAGAatatatTATatcacatauuggaggguuuucauaugaaugaaaaacaguuaaaauuaguugagucuuc uaacaacagaagaaguuauUUCUuauauaAUauaguguauaaccuccaaaaguauacttacttttgtcaattttaatcaactcagaag
7	-8C	attgttgtcttcttcaataAACCatatatTATatcacatauuggaggguuuucauaugaaugaaaaacaguuaaaauuaguugagucuuc uaacaacagaagaaguuauUUGGuauauaAUauaguguauaaccuccaaaaguauacttacttttgtcaattttaatcaactcagaag
8	-7C	attgttgtcttcttcaataAACActatatTATatcacatauuggaggguuuucauaugaaugaaaaacaguuaaaauuaguugagucuuc uaacaacagaagaaguuauUUGUgauauaAUauaguguauaaccuccaaaaguauacttacttttgtcaattttaatcaactcagaag
9	-5C	attgttgtcttcttcaataAACAatctatTATatcacatauuggaggguuuucauaugaaugaaaaacaguuaaaauuaguugagucuuc uaacaacagaagaaguuauUUGUuagauaAUauaguguauaaccuccaaaaguauacttacttttgtcaattttaatcaactcagaag
10	-1C	attgttgtcttcttcaataAACAatatatCATatcacatauuggaggguuuucauaugaaugaaaaacaguuaaaauuaguugagucuuc uaacaacagaagaaguuauUUGUuauauaGUauaguguauaaccuccaaaaguauacttacttttgtcaattttaatcaactcagaag
11	+1C	attgttgtcttcttcaataAACAatatatTCtatcacatauuggaggguuuucauaugaaugaaaaacaguuaaaauuaguugagucuuc uaacaacagaagaaguuauUUGUuauauaAGauaguguauaaccuccaaaaguauacttacttttgtcaattttaatcaactcagaag
12	+2C	attgttgtcttcttcaataAACAatatatTAcatcacatauuggaggguuuucauaugaaugaaaaacaguuaaaauuaguugagucuuc uaacaacagaagaaguuauUUGUuauauaAUguaguguauaaccuccaaaaguauacttacttttgtcaattttaatcaactcagaag

Table 3. DNA templates used in footprinting reactions.

Templates containing the P077 were synthesized by PCR either with dUTP (U, WT and -9G templates) or with dTTP (T template). Sequences of primers are italic and underlined. Conserved nucleotides of the promoter are shown in capital bold letters. The position of the +1 start site is underlined. Uracils are highlighted in blue. Single substitutions are highlighted in red.

Name	Sequence
U	*cctacttatctagtctagaattaattcttgtcttc*uucaauaAACAauauauUAuaucacauauuggaggguuuucauaugaaugaaaaacaguuaaaauuaguugagucuuc ggaugaauagaucagaucuuaauuaagaacagaagaaguuauUUGUuauauaAUauaguguauaaccuccaaaaguauaacttacttttttgtcaattttaatcaactcagaag
T	*cctacttatctagtctagaattaattcttgtcttc*ttcaataAACAatatatTAtatcacatattggaggttttcatatgaatgaaaaacagttaaaattagttgagtcttc ggatgaatagatcagatcttaattaagaacagaagaagttatTTGTtatataATatagtgtataacctccaaaagtataacttacttttttgtcaattttaatcaactcagaag
WT	*attgttgtcttcttcaataAACAatatatTAtatcac*atauuggaggguuuucauaugaaugaaaaacaguuaaaauuaguugagucuuc uaacaacagaagaaguuauUUGUuauauaAUauaguguauaaccuccaaaaguauaacttacttttttgtcaattttaatcaactcagaag
-9G	*attgttgtcttcttcaataAAGAatatatTAtatcac*atauuggaggguuuucauaugaaugaaaaacaguuaaaauuaguugagucuuc uaacaacagaagaaguuauUUCUuauauaAUauaguguauaaccuccaaaaguauaacttacttttttgtcaattttaatcaactcagaag

Table 4. Double-stranded and partially single-stranded DNA templates

Templates containing the P007 or P077 were prepared by annealing of oligonucleotides ordered from Integrated DNA Technologies. Conserved nucleotides of the promoters are shown in capital bold letters. The position of the +1 start site is underlined. Uracils are highlighted in blue.

№	Name	Sequence
Templates used for analysis of requirement for uracils		
P007		
1	T	gagaaaataacaacagatg**AACA**tacaag**TG**tataacaccgagcgtag ctcttttattgttgtctac**TTGT**atgttc**AC**atattgtggctcgcatc
2	U	gagaaaauaacaacagaug**AACA**uacaag**UG**uauaacaccgagcguag cucuuuuauuguugucuac**UUGU**auguuc**AC**auauuguggcucgcauc
3	U T-14	gagaaaauaacaacagaug**AACA**uacaag**UG**uauaacaccgagcguag cucuuuuauuguugucuac**UUGU**auguuc**AC**auauuguggcucgcauc
4	U T-11	gagaaaauaacaacagaug**AACA**uacaag**UG**uauaacaccgagcguag cucuuuuauuguugucuac**TUGU**auguuc**AC**auauuguggcucgcauc
5	U T-10	gagaaaauaacaacagaug**AACA**uacaag**UG**uauaacaccgagcguag cucuuuuauuguugucuac**UTGU**auguuc**AC**auauuguggcucgcauc
6	U T-8	gagaaaauaacaacagaug**AACA**uacaag**UG**uauaacaccgagcguag cucuuuuauuguugucuac**UUGT**auguuc**AC**auauuguggcucgcauc
7	U T-6	gagaaaauaacaacagaug**AACA**uacaag**UG**uauaacaccgagcguag cucuuuuauuguugucuac**UUGU**atguuc**AC**auauuguggcucgcauc
8	T U-14;-6	gagaaaataacaacagatg**AACA**tacaag**TG**tataacaccgagcgtag ctcttttattgttgtcuac**UUGU**augttc**AC**atattgtggctcgcatc
9	T T-11;-10	gagaaaataacaacagatg**AACA**tacaag**TG**tataacaccgagcgtag ctcttttattgttgtctac**UUGT**atgttc**AC**atattgtggctcgcatc
Templates used for analysis of transcription of partially single-stranded templates		
P007		
1	ds	gagaaaataacaacagatg**AACA**tacaag**TG**tataacaccgagcgtag

		ctcttttattgttgtcuac**UUGU**augttc**AC**atattgtggctcgcatc
2	fork nt	gagaaaataacaacagatg**AACA**tacaag**TG**tataacaccgagcgtag tc**AC**atattgtggctcgcatc
3	fork t	ag**TG**tataacaccgagcgtag ctcttttattgttgtcuac**UUGU**augttc**AC**atattgtggctcgcatc
4	T fork t	ag**TG**tataacaccgagcgtag ctcttttattgttgtcta**cTTGT**atgttc**AC**atattgtggctcgcatc
P077		
1	ds	attgttgtcttcttcaata**AACA**atatat**TA**tatcacatattggaggt taacaacagaagaagtta**uUUGU**uatata**AT**atagtgtataacctcca
2	fork nt	attgttgtcttcttcaata**AACA**atatat**TA**tatcacatattggaggt ta**AT**atagtgtataacctcca
3	fork t	at**TA**tatcacatattggaggt taacaacagaagaagtta**uUUGU**uatata**AT**atagtgtataacctcca
4	T fork t	at**TA**tatcacatattggaggt taacaacagaagaagttat**TTGT**tatata**AT**atagtgtataacctcca

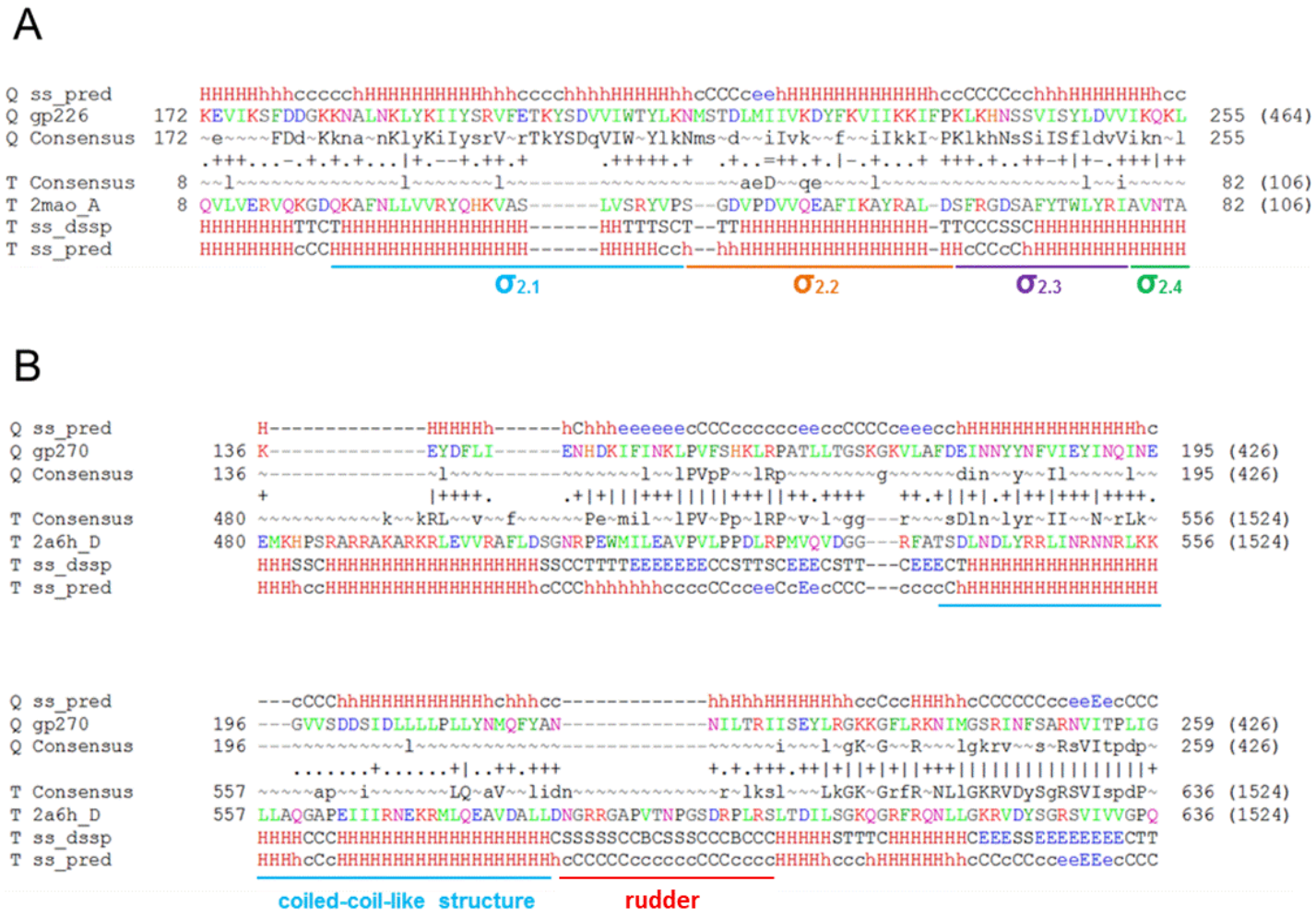
Table 5. Single-stranded DNA templates.

Conserved nucleotides of the P007 promoter are shown in capital bold letters. The position of the +1 start site is underlined. Uracils are highlighted in blue. Single substitutions are highlighted in red.

№	Name	**Sequence** (3'-5' orientation)
Templates used for analysis of requirement for uracils in a context of single-stranded DNA		
1	T	ctcttttattgttgtctac**TTGT**atgttc**AC**atattgtggctcgcatc
2	U	cucuuuuauuguugucuac**UUGU**auguuc**AC**auauuguggcucgcauc
3	U T-14	cucuuuuauuguugucuac**UUGU**auguuc**AC**auauuguggcucgcauc
4	U T-11	cucuuuuauuguugucuac**TUGU**auguuc**AC**auauuguggcucgcauc
5	U T-10	cucuuuuauuguugucuac**UTGU**auguuc**AC**auauuguggcucgcauc
6	U T-8	cucuuuuauuguugucuac**UUGT**auguuc**AC**auauuguggcucgcauc
7	U T-6	cucuuuuauuguugucuac**UUGU**atguuc**AC**auauuguggcucgcauc
8	T U-14;-6	ctcttttattgttgtcuac**UUGU**augttc**AC**atattgtggctcgcatc
9	T T-11;-10	ctcttttattgttgtctac**UUGT**atgttc**AC**atattgtggctcgcatc
Templates used for mutational analysis in the context of single-stranded DNA		
1	WT	ctcttttattgttgtcuac**UUGU**augttc**AC**atattgtggctcgcatc
2	-13C	ctcttttattgttgtcugc**UUGU**augttc**AC**atattgtggctcgcatc
3	-12C	ctcttttattgttgtcuag**UUGU**augttc**AC**atattgtggctcgcatc
4	-11C	ctcttttattgttgtcuac**GUGU**augttc**AC**atattgtggctcgcatc
5	-10C	ctcttttattgttgtcuac**UGGU**augttc**AC**atattgtggctcgcatc
6	-9G	ctcttttattgttgtcuac**UUCU**augttc**AC**atattgtggctcgcatc

7	-8C	ctcttttattgttgtcuacUUGGaugttcACatattgtggctcgcatc
8	-7C	ctcttttattgttgtcuacUUGUgugttcACatattgtggctcgcatc
9	-1C	ctcttttattgttgtcuacUUGUaugttcGCatattgtggctcgcatc
10	+1A	ctcttttattgttgtcuacUUGUaugttcATatattgtggctcgcatc
11	+5C	ctcttttattgttgtcuacUUGUaugttcACatagtgtggctcgcatc

Appendix B



(A) Similarity of gp226 to the region 2 of bacterial σ factors. Alignment of the gp226 fragment with the best hit found in search of homologs by HHpred program (PDB: 2mao_A) is shown. 2mao_A corresponds to the region 2 of σ^E of *E. coli.* Functional regions are indicated under the alignment according to homology of σ^E with primary σ factors [80].

(B) Prediction of β' coiled-coil-like structure in AR9 nvRNAP. The fragment of alignment of AR9 nvRNAP subunit gp270 with β' subunit of *T. thermophilus* RNAP obtained by HHpred program (PDB: 2a6h) is shown. The region forming coiled-coil like structure is indicated under the alignment [14].

The pairwise query-templates alignments with annotation for secondary structure, consensus sequence and column-column match quality are shown [116, 117]. Lines starting with 'Query' are either for AR9 gp226 or gp270.

The alignments consist of blocks with the following lines:

Q ss_pred: query secondary structure as predicted by PSIPRED. Upper case letters: High probability, lower case letters: Low probability (when available)

Q query_name: query sequence

Q Consensus: query alignment consensus sequence

Middle line without a name: quality of column-column match

=	:	very bad match	column score below -1.5
-	:	bad match	column score between -1.5 and -0.5
.	:	neutral match	column score between -0.5 and +0.5
+	:	good match	column score between +0.5 and +1.5
|	:	very good match	column score above +1.5

T Consensus: template alignment consensus sequence

T templ_name: template sequence

T ss_dssp: template secondary structure as determined by DSSP (when available)

T ss_pred: template secondary structure as predicted by PSIPRED. Upper case letters: High probability, lower case letters: Low probability (when available)

The consensus sequence uses capital letters for amino acids that occur with >=60% probability and lower case letters for amino acids that have >=40% probability. For unconserved columns a tilde is used. The line in the middle shows the column score between the query and template amino acid distributions. It gives a valuable indication for the alignment quality. (A unit of column score corresponds approximately to 0.6 bits.)

Appendix C

(1) Statistics of the best dataset collected for the crystal of AR9 nvRNAP core carrying hexahistidine tag

```
UNIT_CELL_CONSTANTS=   112.636   165.706   306.561  90.000  90.000  90.000
SPACE_GROUP = P212121
```

SUBSET OF INTENSITY DATA WITH SIGNAL/NOISE >= -3.0 AS FUNCTION OF RESOLUTION

RESOLUTION LIMIT	NUMBER OF REFLECTIONS OBSERVED	UNIQUE	POSSIBLE	COMPLETENESS OF DATA	R-FACTOR observed	R-FACTOR expected	COMPARED	I/SIGMA	R-meas	CC(1/2)	Anomal Corr	SigAno	Nano
9.74	13942	3414	3684	92.7%	2.7%	2.8%	13718	42.24	3.1%	99.9*	2	0.823	1964
6.96	26903	5942	6107	97.3%	3.8%	3.8%	26601	30.37	4.3%	99.9*	2	0.855	4200
5.70	35109	7617	7767	98.1%	8.7%	8.2%	34792	15.94	9.8%	99.3*	0	0.863	5456
4.94	41600	8992	9100	98.8%	10.7%	10.2%	41302	13.09	12.0%	99.0*	-1	0.828	6407
4.43	47033	10141	10244	99.0%	12.0%	11.7%	46744	11.64	13.5%	99.0*	-1	0.820	7193
4.04	52141	11176	11248	99.4%	18.9%	19.0%	51902	7.64	21.3%	97.5*	1	0.803	7940
3.74	56783	12152	12242	99.3%	35.4%	36.4%	56551	4.35	39.9%	92.2*	-1	0.747	8569
3.50	60907	13010	13096	99.3%	67.2%	70.9%	60691	2.39	75.7%	77.4*	-1	0.693	9139
3.30	57625	13704	13885	98.7%	124.2%	132.5%	57084	1.17	142.0%	42.1*	0	0.650	8177
total	392043	86148	87373	98.6%	12.3%	12.5%	389385	10.07	13.9%	99.8*	0	0.771	59045

(2) Statistics of the best dataset collected for the crystal of tagless AR9 nvRNAP core

```
UNIT_CELL_CONSTANTS=   171.495   231.937   591.908  90.000  90.000  90.000
SPACE_GROUP = P212121

SUBSET OF INTENSITY DATA WITH SIGNAL/NOISE >= -3.0 AS FUNCTION OF RESOLUTION
```

RESOLUTION LIMIT	NUMBER OF REFLECTIONS OBSERVED	UNIQUE	POSSIBLE	COMPLETENESS OF DATA	R-FACTOR observed	R-FACTOR expected	COMPARED	I/SIGMA	R-meas	CC(1/2)	Anomal Corr	SigAno	Nano
11.11	124461	17661	17970	98.3%	5.4%	5.0%	124460	31.60	5.8%	99.8*	74*	1.982	7999
7.96	210322	30891	30923	99.9%	7.0%	6.6%	210319	22.19	7.6%	99.7*	37*	1.328	14615
6.52	274193	39778	39795	100.0%	16.0%	14.9%	274182	10.48	17.4%	98.8*	19*	1.063	19050
5.66	344540	46927	46928	100.0%	27.7%	26.9%	344540	6.39	29.8%	96.6*	8	0.911	22631
5.07	397191	53198	53200	100.0%	34.5%	34.2%	397191	5.14	37.1%	95.2*	4	0.848	25765
4.63	442995	58682	58690	100.0%	38.8%	38.7%	442994	4.60	41.7%	94.2*	1	0.811	28508
4.29	450578	63795	63814	100.0%	51.8%	52.1%	450575	3.25	55.9%	90.0*	-1	0.780	31064
4.02	453211	68482	68537	99.9%	85.3%	87.2%	453195	1.86	92.6%	76.8*	-2	0.722	33393
3.79	492172	70609	72900	96.9%	144.3%	151.4%	491262	1.03	155.8%	55.5*	0	0.669	33831
total	3189663	450023	452757	99.4%	21.4%	21.2%	3188718	6.47	23.1%	99.6*	12*	0.886	216856

(3) Statistics of the best dataset collected for the crystal of AR9 nvRNAP holoenzyme in complex with promoter DNA

```
UNIT_CELL_CONSTANTS=   175.655   108.750   188.839  90.000 109.846  90.000
SPACE_GROUP = C2

SUBSET OF INTENSITY DATA WITH SIGNAL/NOISE >= -3.0 AS FUNCTION OF RESOLUTION
```

RESOLUTION LIMIT	NUMBER OF REFLECTIONS OBSERVED	UNIQUE	POSSIBLE	COMPLETENESS OF DATA	R-FACTOR observed	R-FACTOR expected	COMPARED	I/SIGMA	R-meas	CC(1/2)	Anomal Corr	SigAno	Nano
10.02	11807	3492	3567	97.9%	2.9%	2.8%	11778	37.37	3.4%	99.8*	21*	0.992	1597
7.16	21990	6183	6198	99.8%	4.7%	4.4%	21957	22.93	5.6%	99.8*	5	0.893	2939
5.87	27643	7935	7957	99.7%	12.3%	12.0%	27561	9.74	14.5%	98.5*	5	0.850	3775
5.09	33934	9397	9421	99.7%	14.9%	14.9%	33866	8.33	17.5%	97.9*	3	0.823	4518
4.56	37198	10603	10624	99.8%	13.0%	12.8%	37091	9.33	15.4%	98.4*	2	0.823	5080
4.16	40595	11742	11778	99.7%	16.2%	16.3%	40466	7.53	19.2%	97.6*	1	0.799	5627
3.86	45448	12766	12782	99.9%	25.8%	26.5%	45315	4.95	30.4%	94.8*	1	0.770	6156
3.61	49382	13709	13739	99.8%	41.4%	43.1%	49254	3.12	48.7%	86.8*	1	0.750	6631
3.40	49594	14389	14577	98.7%	69.6%	73.2%	49359	1.80	82.5%	69.9*	2	0.721	6879
total	317591	90216	90643	99.5%	15.3%	15.5%	316647	8.28	18.1%	99.1*	3	0.798	43202

Appendix D

(1) Matthews coefficients calculated for the crystal of AR9 nvRNAP core carrying hexahistidine tag

The coefficients (V_M) were calculated using the online calculator: http://csb.wfu.edu/tools/vmcalc/vm.html

Usual value for V_M: 1.62 < V_M < 3.53 A^3/Da

Mw, Da	**Number of molecules per asymmetric unit**	Volume, A^3	Z	V_M (V/Mw), A^3/Da	Solvent Content, %
259783,44	0,5	5721795,83	2	11,01	88,83
	1	5721795,83	4	5,51	77,66
	2	**5721795,83**	**8**	**2,75**	**55,32**
	3	5721795,83	12	1,84	32,99
	4	5721795,83	16	1,38	10,65

(2) Matthews coefficients calculated for the crystal of tagless AR9 nvRNAP core

The coefficients (V_M) were calculated using the online calculator: http://csb.wfu.edu/tools/vmcalc/vm.html

Usual value for V_M: 1.62 < V_M < 3.53 A^3/Da

Mw, Da	**Number of molecules per asymmetric unit**	Volume, A^3	Z	V_M (V/Mw), A^3/Da	Solvent Content, %
257240,79	1	23543753,8	4	22,88	94,62
	2	23543753,8	8	11,44	89,25
	4	23543753,8	16	5,72	78,5
	5	23543753,8	20	4,58	73,12
	6	23543753,8	24	3,81	67,75
	7	23543753,8	28	3,27	62,37
	8	23543753,8	**32**	**2,86**	**56,99**
	9	23543753,8	36	2,54	51,62
	10	23543753,8	40	2,29	46,24
	11	23543753,8	44	2,08	40,87
	12	23543753,8	48	1,97	35,49

(3) Matthews coefficients calculated for the crystal of AR9 nvRNAP holoenzyme in complex with promoter DNA

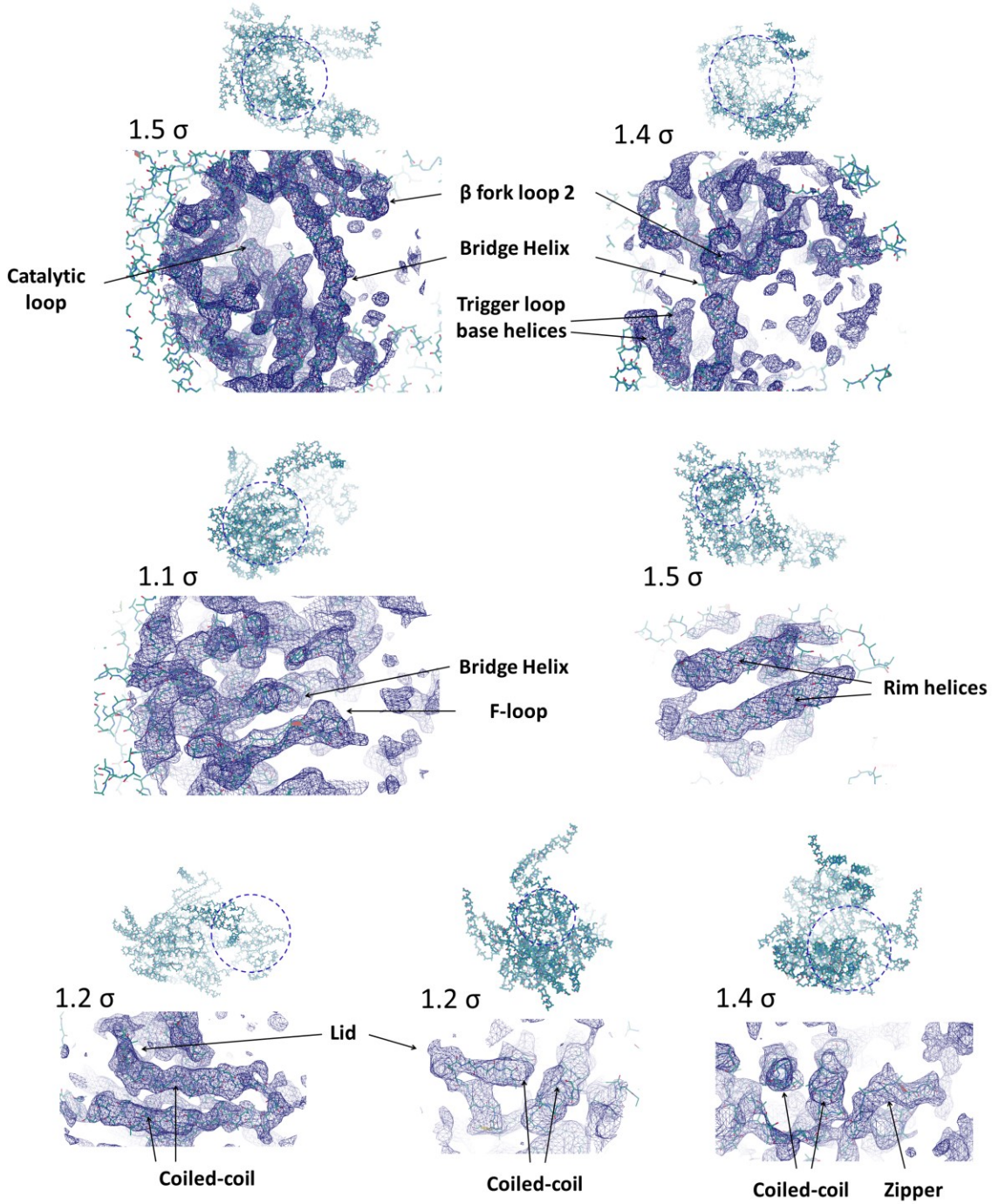
The coefficients (V_M) were calculated using the online calculator: http://csb.wfu.edu/tools/vmcalc/vm.html

Usual value for V_M: 1.62 < V_M < 3.53 A^3/Da

Mw, Da	**Number of molecules per asymmetric unit**	Volume, A^3	Z	V_M (V/Mw), A^3/Da	Solvent Content, %
327353,35	0.5	3393050,92	2	5,18	76,27
	1	**3393050,92**	**4**	**2,59**	**52,53**
	2	3393050,92	8	1,30	5,07

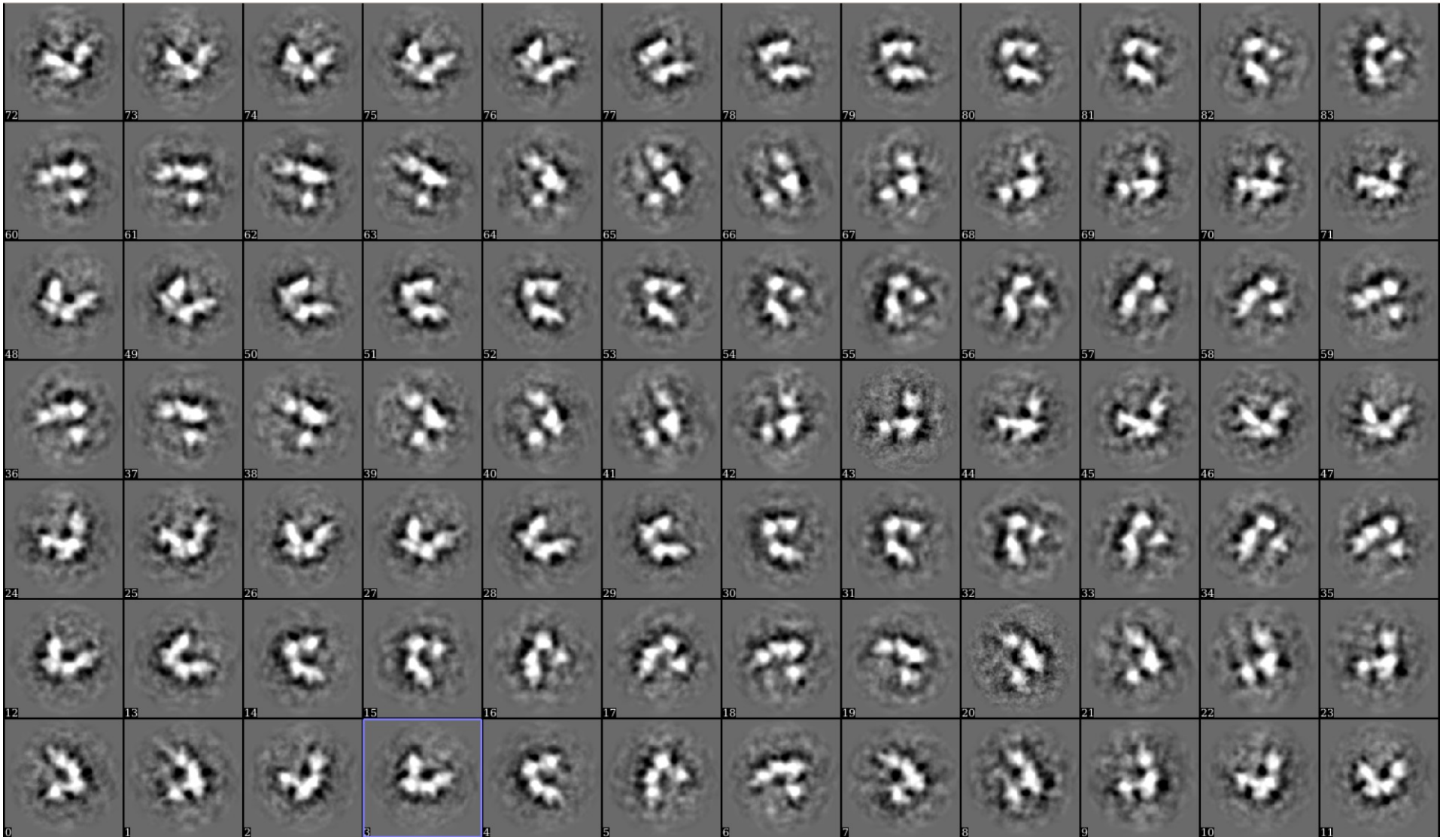
Appendix E

Fragments of the electron density map of AR9 nvRNAP core (shown in Fig. 27) corresponding to some elements common for all multisubunit RNAPs are shown (the elements are indicated by arrows). Contouring levels are indicated above the maps (in standard deviations units (σ) above the mean). Overall views of AR9 nvRNAP core model are shown above the fragments of the map (zoomed regions of the model are indicated by dashed circles). The figure was prepared using the program Coot [132].

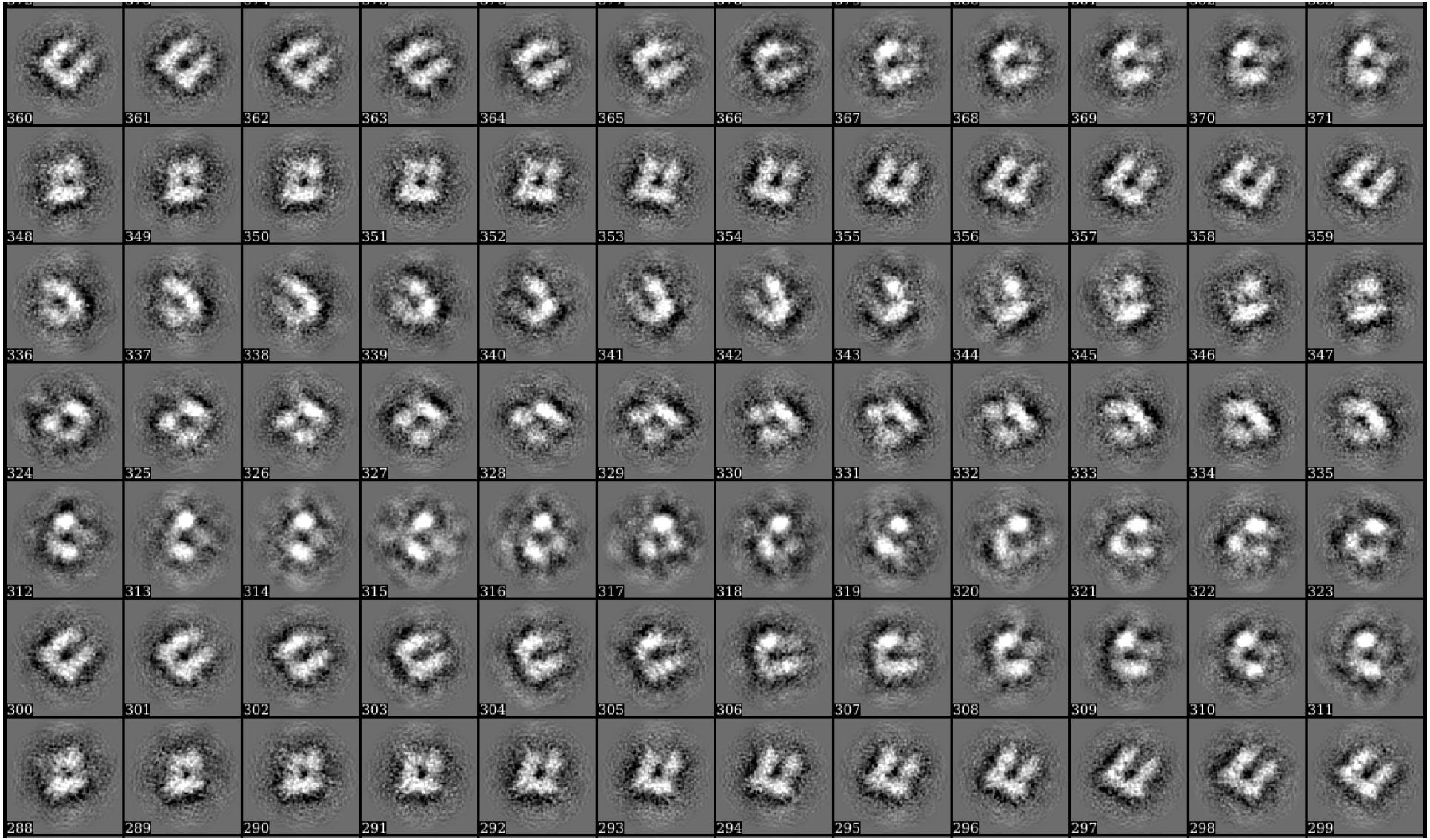


Appendix F

Results of 2D classification of single particles of AR9 nvRNAP core



Results of 2D classification of single particles of AR9 nvRNAP holoenzyme



Bibliography

1. Monttinen, H.A., et al., *Automated structural comparisons clarify the phylogeny of the right-hand-shaped polymerases.* Mol Biol Evol, 2014. **31**(10): p. 2741-52.
2. Fouqueau, T., F. Blombach, and F. Werner, *Evolutionary Origins of Two-Barrel RNA Polymerases and Site-Specific Transcription Initiation.* Annu Rev Microbiol, 2017. **71**: p. 331-348.
3. Iyer, L.M., E.V. Koonin, and L. Aravind, *Evolutionary connection between the catalytic subunits of DNA-dependent RNA polymerases and eukaryotic RNA-dependent RNA polymerases and the origin of RNA polymerases.* BMC Struct Biol, 2003. **3**: p. 1.
4. Sauguet, L., et al., *Shared active site architecture between archaeal PolD and multi-subunit RNA polymerases revealed by X-ray crystallography.* Nat Commun, 2016. **7**: p. 12227.
5. Cermakian, N., et al., *On the evolution of the single-subunit RNA polymerases.* J Mol Evol, 1997. **45**(6): p. 671-81.
6. Werner, F. and D. Grohmann, *Evolution of multisubunit RNA polymerases in the three domains of life.* Nat Rev Microbiol, 2011. **9**(2): p. 85-98.
7. Berdygulova, Z., et al., *Temporal regulation of gene expression of the Thermus thermophilus bacteriophage P23-45.* J Mol Biol, 2011. **405**(1): p. 125-42.
8. Yutin, N., et al., *Discovery of an expansive bacteriophage family that includes the most abundant viruses from the human gut.* Nat Microbiol, 2018. **3**(1): p. 38-46.
9. Ruprich-Robert, G. and P. Thuriaux, *Non-canonical DNA transcription enzymes and the conservation of two-barrel RNA polymerases.* Nucleic Acids Res, 2010. **38**(14): p. 4559-69.
10. Castillo, R.M., et al., *A six-stranded double-psi beta barrel is shared by several protein superfamilies.* Structure, 1999. **7**(2): p. 227-36.
11. Burton, Z.F., *The Old and New Testaments of gene regulation. Evolution of multi-subunit RNA polymerases and co-evolution of eukaryote complexity with the RNAP II CTD.* Transcription, 2014. **5**(3): p. e28674.
12. Sosunov, V., et al., *Unified two-metal mechanism of RNA synthesis and degradation by RNA polymerase.* EMBO J, 2003. **22**(9): p. 2234-44.
13. Vassylyev, D.G., et al., *Structural basis for substrate loading in bacterial RNA polymerase.* Nature, 2007. **448**(7150): p. 163-8.
14. Lane, W.J. and S.A. Darst, *Molecular evolution of multisubunit RNA polymerases: structural analysis.* J Mol Biol, 2010. **395**(4): p. 686-704.
15. Salgado, P.S., et al., *The structure of an RNAi polymerase links RNA silencing and transcription.* PLoS Biol, 2006. **4**(12): p. e434.
16. Wu, S., et al., *Structural comparison of DNA polymerase architecture suggests a nucleotide gateway to the polymerase active site.* Chem Rev, 2014. **114**(5): p. 2759-74.

17. Griesenbeck, J., H. Tschochner, and D. Grohmann, *Structure and Function of RNA Polymerases and the Transcription Machineries.* Subcell Biochem, 2017. **83**: p. 225-270.
18. Zhang, G., et al., *Crystal structure of Thermus aquaticus core RNA polymerase at 3.3 A resolution.* Cell, 1999. **98**(6): p. 811-24.
19. Cramer, P., et al., *Architecture of RNA polymerase II and implications for the transcription mechanism.* Science, 2000. **288**(5466): p. 640-9.
20. Hirata, A., B.J. Klein, and K.S. Murakami, *The X-ray crystal structure of RNA polymerase from Archaea.* Nature, 2008. **451**(7180): p. 851-4.
21. Ishihama, A., *Subunit of assembly of Escherichia coli RNA polymerase.* Adv Biophys, 1981. **14**: p. 1-35.
22. Minakhin, L., et al., *Bacterial RNA polymerase subunit omega and eukaryotic RNA polymerase subunit RPB6 are sequence, structural, and functional homologs and promote RNA polymerase assembly.* Proc Natl Acad Sci U S A, 2001. **98**(3): p. 892-7.
23. Werner, F., *Structural evolution of multisubunit RNA polymerases.* Trends Microbiol, 2008. **16**(6): p. 247-50.
24. Lee, J. and S. Borukhov, *Bacterial RNA Polymerase-DNA Interaction-The Driving Force of Gene Expression and the Target for Drug Action.* Front Mol Biosci, 2016. **3**: p. 73.
25. Gnatt, A.L., et al., *Structural basis of transcription: an RNA polymerase II elongation complex at 3.3 A resolution.* Science, 2001. **292**(5523): p. 1876-82.
26. Vassylyev, D.G., et al., *Structural basis for transcription elongation by bacterial RNA polymerase.* Nature, 2007. **448**(7150): p. 157-62.
27. Korzheva, N. and A. Mustaev, *Transcription elongation complex: structure and function.* Curr Opin Microbiol, 2001. **4**(2): p. 119-25.
28. Nudler, E., et al., *The RNA-DNA hybrid maintains the register of transcription by preventing backtracking of RNA polymerase.* Cell, 1997. **89**(1): p. 33-41.
29. Korzheva, N., et al., *A structural model of transcription elongation.* Science, 2000. **289**(5479): p. 619-25.
30. Mukhopadhyay, J., et al., *Antibacterial peptide microcin J25 inhibits transcription by binding within and obstructing the RNA polymerase secondary channel.* Mol Cell, 2004. **14**(6): p. 739-51.
31. Batada, N.N., et al., *Diffusion of nucleoside triphosphates and role of the entry site to the RNA polymerase II active center.* Proc Natl Acad Sci U S A, 2004. **101**(50): p. 17361-4.
32. Kettenberger, H., K.J. Armache, and P. Cramer, *Architecture of the RNA polymerase II-TFIIS complex and implications for mRNA cleavage.* Cell, 2003. **114**(3): p. 347-57.
33. Parshin, A., et al., *DksA regulates RNA polymerase in Escherichia coli through a network of interactions in the secondary channel that includes Sequence Insertion 1.* Proc Natl Acad Sci U S A, 2015. **112**(50): p. E6862-71.
34. Laptenko, O., et al., *Transcript cleavage factors GreA and GreB act as transient catalytic components of RNA polymerase.* EMBO J, 2003. **22**(23): p. 6322-34.
35. Tagami, S., et al., *Crystal structure of bacterial RNA polymerase bound with a transcription inhibitor protein.* Nature, 2010. **468**(7326): p. 978-82.

36. Chakraborty, A., et al., *Opening and closing of the bacterial RNA polymerase clamp.* Science, 2012. **337**(6094): p. 591-5.
37. 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.
38. Schulz, S., et al., *TFE and Spt4/5 open and close the RNA polymerase clamp during the transcription cycle.* Proc Natl Acad Sci U S A, 2016. **113**(13): p. E1816-25.
39. 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.
40. Weinzierl, R.O., *The Bridge Helix of RNA polymerase acts as a central nanomechanical switchboard for coordinating catalysis and substrate movement.* Archaea, 2011. **2011**: p. 608385.
41. Mejia, Y.X., E. Nudler, and C. Bustamante, *Trigger loop folding determines transcription rate of Escherichia coli's RNA polymerase.* Proc Natl Acad Sci U S A, 2015. **112**(3): p. 743-8.
42. Miropolskaya, N., et al., *Allosteric control of catalysis by the F loop of RNA polymerase.* Proc Natl Acad Sci U S A, 2009. **106**(45): p. 18942-7.
43. Sidorenkov, I., N. Komissarova, and M. Kashlev, *Crucial role of the RNA:DNA hybrid in the processivity of transcription.* Mol Cell, 1998. **2**(1): p. 55-64.
44. Kireeva, M.L., N. Komissarova, and M. Kashlev, *Overextended RNA:DNA hybrid as a negative regulator of RNA polymerase II processivity.* J Mol Biol, 2000. **299**(2): p. 325-35.
45. 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.
46. 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.
47. 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.
48. Barnes, C.O., et al., *Crystal Structure of a Transcribing RNA Polymerase II Complex Reveals a Complete Transcription Bubble.* Mol Cell, 2015. **59**(2): p. 258-69.
49. Naji, S., et al., *Structure-function analysis of the RNA polymerase cleft loops elucidates initial transcription, DNA unwinding and RNA displacement.* Nucleic Acids Res, 2008. **36**(2): p. 676-87.
50. Paget, M.S., *Bacterial Sigma Factors and Anti-Sigma Factors: Structure, Function and Distribution.* Biomolecules, 2015. **5**(3): p. 1245-65.
51. Blombach, F., et al., *Molecular Mechanisms of Transcription Initiation-Structure, Function, and Evolution of TFE/TFIIE-Like Factors and Open Complex Formation.* J Mol Biol, 2016. **428**(12): p. 2592-2606.
52. Hantsche, M. and P. Cramer, *Conserved RNA polymerase II initiation complex structure.* Curr Opin Struct Biol, 2017. **47**: p. 17-22.

53. Burgess, R.R., et al., *Factor stimulating transcription by RNA polymerase.* Nature, 1969. **221**(5175): p. 43-6.
54. Staron, A., et al., *The third pillar of bacterial signal transduction: classification of the extracytoplasmic function (ECF) sigma factor protein family.* Mol Microbiol, 2009. **74**(3): p. 557-81.
55. Mittenhuber, G., *An inventory of genes encoding RNA polymerase sigma factors in 31 completely sequenced eubacterial genomes.* J Mol Microbiol Biotechnol, 2002. **4**(1): p. 77-91.
56. Rappas, M., D. Bose, and X. Zhang, *Bacterial enhancer-binding proteins: unlocking sigma54-dependent gene transcription.* Curr Opin Struct Biol, 2007. **17**(1): p. 110-6.
57. Feklistov, A., et al., *Bacterial sigma factors: a historical, structural, and genomic perspective.* Annu Rev Microbiol, 2014. **68**: p. 357-76.
58. Mekler, V., et al., *Structural organization of bacterial RNA polymerase holoenzyme and the RNA polymerase-promoter open complex.* Cell, 2002. **108**(5): p. 599-614.
59. Feklistov, A., *RNA polymerase: in search of promoters.* Ann N Y Acad Sci, 2013. **1293**: p. 25-32.
60. Blatter, E.E., et al., *Domain organization of RNA polymerase alpha subunit: C-terminal 85 amino acids constitute a domain capable of dimerization and DNA binding.* Cell, 1994. **78**(5): p. 889-96.
61. Ross, W., et al., *A third recognition element in bacterial promoters: DNA binding by the alpha subunit of RNA polymerase.* Science, 1993. **262**(5138): p. 1407-13.
62. Feklistov, A. and S.A. Darst, *Structural basis for promoter-10 element recognition by the bacterial RNA polymerase sigma subunit.* Cell, 2011. **147**(6): p. 1257-69.
63. Newlands, J.T., et al., *Both fis-dependent and factor-independent upstream activation of the rrnB P1 promoter are face of the helix dependent.* Nucleic Acids Res, 1992. **20**(4): p. 719-26.
64. Leirmo, S. and R.L. Gourse, *Factor-independent activation of Escherichia coli rRNA transcription. I. Kinetic analysis of the roles of the upstream activator region and supercoiling on transcription of the rrnB P1 promoter in vitro.* J Mol Biol, 1991. **220**(3): p. 555-68.
65. McAllister, C.F. and E.C. Achberger, *Effect of polyadenine-containing curved DNA on promoter utilization in Bacillus subtilis.* J Biol Chem, 1988. **263**(24): p. 11743-9.
66. Hook-Barnard, I.G. and D.M. Hinton, *Transcription initiation by mix and match elements: flexibility for polymerase binding to bacterial promoters.* Gene Regul Syst Bio, 2007. **1**: p. 275-93.
67. Chang, B.Y., Y.T. Shyu, and R.H. Doi, *The interaction between Bacillus subtilis sigma-A (sigma A) factor and RNA polymerase with promoters.* Biochimie, 1992. **74**(7-8): p. 601-12.
68. Dombroski, A.J., W.A. Walter, and C.A. Gross, *The role of the sigma subunit in promoter recognition by RNA polymerase.* Cell Mol Biol Res, 1993. **39**(4): p. 311-7.

69. Callaci, S., E. Heyduk, and T. Heyduk, *Core RNA polymerase from E. coli induces a major change in the domain arrangement of the sigma 70 subunit.* Mol Cell, 1999. **3**(2): p. 229-38.
70. Kuznedelov, K., et al., *A role for interaction of the RNA polymerase flap domain with the sigma subunit in promoter recognition.* Science, 2002. **295**(5556): p. 855-7.
71. Campbell, E.A., et al., *Structure of the bacterial RNA polymerase promoter specificity sigma subunit.* Mol Cell, 2002. **9**(3): p. 527-39.
72. Barne, K.A., et al., *Region 2.5 of the Escherichia coli RNA polymerase sigma70 subunit is responsible for the recognition of the 'extended-10' motif at promoters.* EMBO J, 1997. **16**(13): p. 4034-40.
73. Liu, X., D.A. Bushnell, and R.D. Kornberg, *Lock and key to transcription: sigma-DNA interaction.* Cell, 2011. **147**(6): p. 1218-9.
74. Severinov, K. and S.A. Darst, *A mutant RNA polymerase that forms unusual open promoter complexes.* Proc Natl Acad Sci U S A, 1997. **94**(25): p. 13481-6.
75. Heyduk, E., et al., *A consensus adenine at position -11 of the nontemplate strand of bacterial promoter is important for nucleation of promoter melting.* J Biol Chem, 2006. **281**(18): p. 12362-9.
76. Bae, B., et al., *Structure of a bacterial RNA polymerase holoenzyme open promoter complex.* Elife, 2015. **4**.
77. Tomsic, M., et al., *Different roles for basic and aromatic amino acids in conserved region 2 of Escherichia coli sigma(70) in the nucleation and maintenance of the single-stranded DNA bubble in open RNA polymerase-promoter complexes.* J Biol Chem, 2001. **276**(34): p. 31891-6.
78. 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.
79. Juang, Y.L. and J.D. Helmann, *A promoter melting region in the primary sigma factor of Bacillus subtilis. Identification of functionally important aromatic amino acids.* J Mol Biol, 1994. **235**(5): p. 1470-88.
80. Campagne, S., et al., *Structural basis for -10 promoter element melting by environmentally induced sigma factors.* Nat Struct Mol Biol, 2014. **21**(3): p. 269-76.
81. Murakami, K.S. and S.A. Darst, *Bacterial RNA polymerases: the wholo story.* Curr Opin Struct Biol, 2003. **13**(1): p. 31-9.
82. Kulbachinskiy, A. and A. Mustaev, *Region 3.2 of the sigma subunit contributes to the binding of the 3'-initiating nucleotide in the RNA polymerase active center and facilitates promoter clearance during initiation.* J Biol Chem, 2006. **281**(27): p. 18273-6.
83. Basu, R.S., et al., *Structural basis of transcription initiation by bacterial RNA polymerase holoenzyme.* J Biol Chem, 2014. **289**(35): p. 24549-59.
84. Pupov, D., et al., *Distinct functions of the RNA polymerase sigma subunit region 3.2 in RNA priming and promoter escape.* Nucleic Acids Res, 2014. **42**(7): p. 4494-504.

85. Nickels, B.E., et al., *The interaction between sigma70 and the beta-flap of Escherichia coli RNA polymerase inhibits extension of nascent RNA during early elongation.* Proc Natl Acad Sci U S A, 2005. **102**(12): p. 4488-93.
86. Liu, B., Y. Zuo, and T.A. Steitz, *Structures of E. coli sigmaS-transcription initiation complexes provide new insights into polymerase mechanism.* Proc Natl Acad Sci U S A, 2016. **113**(15): p. 4051-6.
87. Iyer, L.M. and L. Aravind, *Insights from the architecture of the bacterial transcription apparatus.* J Struct Biol, 2012. **179**(3): p. 299-319.
88. Lavysh, D., et al., *The genome of AR9, a giant transducing Bacillus phage encoding two multisubunit RNA polymerases.* Virology, 2016. **495**: p. 185-96.
89. Mirzakhanyan, Y. and P.D. Gershon, *Multisubunit DNA-Dependent RNA Polymerases from Vaccinia Virus and Other Nucleocytoplasmic Large-DNA Viruses: Impressions from the Age of Structure.* Microbiol Mol Biol Rev, 2017. **81**(3).
90. Forrest, D., et al., *Single-peptide DNA-dependent RNA polymerase homologous to multi-subunit RNA polymerase.* Nat Commun, 2017. **8**: p. 15774.
91. Lane, W.J. and S.A. Darst, *Molecular evolution of multisubunit RNA polymerases: sequence analysis.* J Mol Biol, 2010. **395**(4): p. 671-85.
92. Zakharova, N., et al., *The largest subunits of RNA polymerase from gastric helicobacters are tethered.* J Biol Chem, 1998. **273**(31): p. 19371-4.
93. Shabalina, S.A. and E.V. Koonin, *Origins and evolution of eukaryotic RNA interference.* Trends Ecol Evol, 2008. **23**(10): p. 578-87.
94. Spencer, E., S. Shuman, and J. Hurwitz, *Purification and properties of vaccinia virus DNA-dependent RNA polymerase.* J Biol Chem, 1980. **255**(11): p. 5388-95.
95. Ahn, B.Y. and B. Moss, *RNA polymerase-associated transcription specificity factor encoded by vaccinia virus.* Proc Natl Acad Sci U S A, 1992. **89**(8): p. 3536-40.
96. Guarino, L.A., et al., *A virus-encoded RNA polymerase purified from baculovirus-infected cells.* J Virol, 1998. **72**(10): p. 7985-91.
97. Rohrmann, G.F., in *Baculovirus Molecular Biology*, rd, Editor. 2013: Bethesda (MD).
98. Knebel-Morsdorf, D., et al., *Expression of baculovirus late and very late genes depends on LEF-4, a component of the viral RNA polymerase whose guanyltransferase function is essential.* J Virol, 2006. **80**(8): p. 4168-73.
99. Su, J., O. Lung, and G.W. Blissard, *The Autographa californica multiple nucleopolyhedrovirus lef-5 gene is required for productive infection.* Virology, 2011. **416**(1-2): p. 54-64.
100. Guarino, L.A., W. Dong, and J. Jin, *In vitro activity of the baculovirus late expression factor LEF-5.* J Virol, 2002. **76**(24): p. 12663-75.
101. Ceyssens, P.J., et al., *Development of giant bacteriophage varphiKZ is independent of the host transcription apparatus.* J Virol, 2014. **88**(18): p. 10501-10.
102. Lavysh, D., et al., *Transcription Profiling of Bacillus subtilis Cells Infected with AR9, a Giant Phage Encoding Two Multisubunit RNA Polymerases.* MBio, 2017. **8**(1).

103. Hendrix, R.W., *Jumbo bacteriophages.* Curr Top Microbiol Immunol, 2009. **328**: p. 229-40.
104. Bhunchoth, A., et al., *Two asian jumbo phages, varphiRSL2 and varphiRSF1, infect Ralstonia solanacearum and show common features of varphiKZ-related phages.* Virology, 2016. **494**: p. 56-66.
105. Lecoutere, E., et al., *Identification and comparative analysis of the structural proteomes of phiKZ and EL, two giant Pseudomonas aeruginosa bacteriophages.* Proteomics, 2009. **9**(11): p. 3215-9.
106. Skurnik, M., et al., *Characterization of the genome, proteome, and structure of yersiniophage varphiR1-37.* J Virol, 2012. **86**(23): p. 12625-42.
107. Thomas, J.A., et al., *Proteome of the large Pseudomonas myovirus 201 phi 2-1: delineation of proteolytically processed virion proteins.* Mol Cell Proteomics, 2010. **9**(5): p. 940-51.
108. Yakunina, M., et al., *A non-canonical multisubunit RNA polymerase encoded by a giant bacteriophage.* Nucleic Acids Res, 2015. **43**(21): p. 10411-20.
109. Clark, S., R. Losick, and J. Pero, *New RNA polymerase from Bacillus subtilis infected with phage PBS2.* Nature, 1974. **252**(5478): p. 21-4.
110. Clark, S., *Transcriptional specificity of a multisubunit RNA polymerase induced by Bacillus subtilis bacteriophage PBS2.* J Virol, 1978. **25**(1): p. 224-37.
111. Kiljunen, S., et al., *Yersiniophage phiR1-37 is a tailed bacteriophage having a 270 kb DNA genome with thymidine replaced by deoxyuridine.* Microbiology, 2005. **151**(Pt 12): p. 4093-102.
112. Bickle, T.A. and D.H. Kruger, *Biology of DNA restriction.* Microbiol Rev, 1993. **57**(2): p. 434-50.
113. Choy, H.A., J.M. Romeo, and E.P. Geiduschek, *Activity of a phage-modified RNA polymerase at hybrid promoters. Effects of substituting thymine for hydroxymethyluracil in a phage SP01 middle promoter.* J Mol Biol, 1986. **191**(1): p. 59-73.
114. Kashlev, M., et al., *Bacteriophage T4 Alc protein: a transcription termination factor sensing local modification of DNA.* Cell, 1993. **75**(1): p. 147-54.
115. Zhilina, E., et al., *Structural transitions in the transcription elongation complexes of bacterial RNA polymerase during sigma-dependent pausing.* Nucleic Acids Res, 2012. **40**(7): p. 3078-91.
116. Soding, J., A. Biegert, and A.N. Lupas, *The HHpred interactive server for protein homology detection and structure prediction.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W244-8.
117. Alva, V., et al., *The MPI bioinformatics Toolkit as an integrative platform for advanced protein sequence and structure analysis.* Nucleic Acids Res, 2016. **44**(W1): p. W410-5.
118. Minakhin, L. and K. Severinov, *On the role of the Escherichia coli RNA polymerase sigma 70 region 4.2 and alpha-subunit C-terminal domains in promoter complex formation on the extended -10 galP1 promoter.* J Biol Chem, 2003. **278**(32): p. 29710-8.

119. Minakhin, L., et al., *Recombinant Thermus aquaticus RNA polymerase, a new tool for structure-based analysis of transcription.* J Bacteriol, 2001. **183**(1): p. 71-6.
120. Matthews, B.W., *Solvent content of protein crystals.* J Mol Biol, 1968. **33**(2): p. 491-7.
121. Scapin, G., *Molecular replacement then and now.* Acta Crystallogr D Biol Crystallogr, 2013. **69**(Pt 11): p. 2266-75.
122. Pike, A.C., et al., *An overview of heavy-atom derivatization of protein crystals.* Acta Crystallogr D Struct Biol, 2016. **72**(Pt 3): p. 303-18.
123. Taylor, G., *The phase problem.* Acta Crystallogr D Biol Crystallogr, 2003. **59**(Pt 11): p. 1881-90.
124. Walden, H., *Selenium incorporation using recombinant techniques.* Acta Crystallogr D Biol Crystallogr, 2010. **66**(Pt 4): p. 352-7.
125. de La Fortelle, E.B., Gérard., *Maximum-Likelihood Heavy-Atom Parameter Refinement for Multiple Isomorphous Replacement and Multiwavelength Anomalous Diffraction Methods. Methods in enzymology.* Methods in enzymology, 1997. **276C(part A)**: p. 472-494.
126. Cowtan, K., *Recent developments in classical density modification.* Acta Crystallogr D Biol Crystallogr, 2010. **66**(Pt 4): p. 470-8.
127. Cowtan, K., *'dm': An automated procedure for phase improvement by density modification.* Joint CCP4 and ESF-EACBM Newsletter on Protein Crystallography, 1994. **31**: p. 34-38.
128. Pettersen, E.F., et al., *UCSF Chimera--a visualization system for exploratory research and analysis.* J Comput Chem, 2004. **25**(13): p. 1605-12.
129. Chandrasekar, V., S. Munshi, and J.E. Johnson, *Crystallization and preliminary X-ray analysis of tobacco ringspot virus.* Acta Crystallogr D Biol Crystallogr, 1997. **53**(Pt 1): p. 125-8.
130. Duyvesteyn, H.M.E., et al., *Towards in cellulo virus crystallography.* Sci Rep, 2018. **8**(1): p. 3771.
131. McCoy, A.J., et al., *Phaser crystallographic software.* J Appl Crystallogr, 2007. **40**(Pt 4): p. 658-674.
132. Emsley, P. and K. Cowtan, *Coot: model-building tools for molecular graphics.* Acta Crystallogr D Biol Crystallogr, 2004. **60**(Pt 12 Pt 1): p. 2126-32.