UNIVERSITÉ DE STRASBOURG

ÉCOLE DOCTORALE des Sciences de la Vie et de la Santé

IGBMC - CNRS UMR 7104 - Inserm U 964

THÈSE

présentée par:

Ayesha Dinshaw EDULJEE

soutenue le : **03 Mai 2021**

pour obtenir le grade de : **Docteur de l'Université de Strasbourg**

Discipline / Spécialité : Biophysique et Biologie Structurale

Structural and functional studies on transcriptional proofreading by GreA

Étude Structurale et fonctionnelle de l'activité de rélecture par GreA au cours de la transcription

THÈSE dirigée par :

Dr. WEIXLBAUMER Albert	Directeur de recherches, IGBMC, France

RAPPORTEURS :

Prof. ARTSIMOVITCH Irina	Professeur, Ohio State University, Etats-Unis
Prof. GROHMANN Dina	Professeur, University of Regensburg, Allemagne

AUTRES MEMBRES DU JURY :

Dr. KLAHOLZ Bruno	Directeur de recherches, IGBMC, France

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my thesis supervisor, Albert Weixlbaumer, for his guidance and trust. I truly could not have asked for a better teacher or mentor during my thesis.

I wish to thank the members of my thesis jury – Irina Artsimovitch, Dina Grohmann, and Bruno Klaholz – for their valuable time and insights in evaluating my work.

I am extremely grateful to every member of the lab, past and present, all of whom I've learnt so much from. Thank you to Claire Batisse and Charlotte Saint-André for teaching me practically everything I needed to know when I first started. Thank you to Maria Takacs and Michael Webster for the opportunity to work on a small part of their project, which was such a valuable learning experience. Thank you to Xieyang Guo for helping me out with collecting one of my first test datasets on the microscope. Thanks also to Sanjay Dey, Mo'men Abdelkareem, Chengjin Zhu, and Vita Vidmar for all of the support and helpful discussions over the last few years.

I very much appreciate all of the help and training I've received over the years from everyone on the EM platform at the IGBMC: Corinne Crucifix, Alexandre Durand, Gabor Papai, Tat Cheng, and Julio Ortiz. I also wish to thank the administrative and technical staff at the IGBMC, especially for keeping everything running through a particularly difficult year.

The last three years have been anything but dull, all thanks to the friends and colleagues I was lucky enough to be surrounded by. Special thanks to Robert Fagiewicz and Pernelle Klein, who have been my pillars of support every step of the way. Many thanks also to Beatriz Germán, Pau Jane, Torben Klos, Alastair McEwen, Christophe Lotz, and everyone from every iteration of our lunch group, for the countless laughs.

None of this would have been possible without my family. My deepest thanks and appreciation to my parents, Sravani and Dinshaw Eduljee, for their constant encouragement and support, which has always been and continues to be the greatest driving force in every single thing that I do. I am so incredibly grateful to my sister, Trisha Eduljee, for never failing to inspire me and for always having my back, even if its from 6600 km away.

Finally, I could not begin to express how immensely grateful I am to my best friend and partner, Yash Deshpande, for his unwavering support and belief in me, and for always pushing me to be better.

Abstract

In all three kingdoms of life, gene expression begins with the transcription of DNA into messenger RNA (mRNA) by multi-subunit DNA-dependent RNA polymerases (RNAP). In spite of their varied subunit compositions, the core architecture and catalytic mechanisms are conserved in RNAPs in all organisms. The RNAP active site catalyses two reactions – substrate addition and hydrolysis of phosphodiester bonds. Of these, the latter is required for rescuing RNAP elongation complexes that have undergone backtracking, resulting in a complex incapable of extending the nascent RNA. In a backtracked complex, the 3'-end of the RNA is pushed out of the active site and into a region of the enzyme known as the secondary channel. Apart from its role in regulatory pauses, backtracking is also necessary for proofreading: misincorporated NTPs result in the elongation complex favouring backtracking by 1-2 nucleotides. In *E. coli*, cleavage of shorter RNA fragments is increased in the presence of the transcription factor GreA (functionally analogous to eukaryotic TFIIS), which is known to assist in cleavage through two highly conserved acidic residues.

Cryo-EM structures of a complex backtracked by 1 nucleotide with and without the cleavage factor GreA bound to the secondary channel were used to address questions related to the process of proofreading in *E. coli* RNAP, specifically those of the importance of the RNAP structural motif known as the trigger loop, and the process of selection for GreA amongst other structurally similar transcription factors. In addition to this, the structural data along with results from *in vitro* transcription assays highlighted the role of GreA in proofreading outside of the interactions involving its acidic residues.

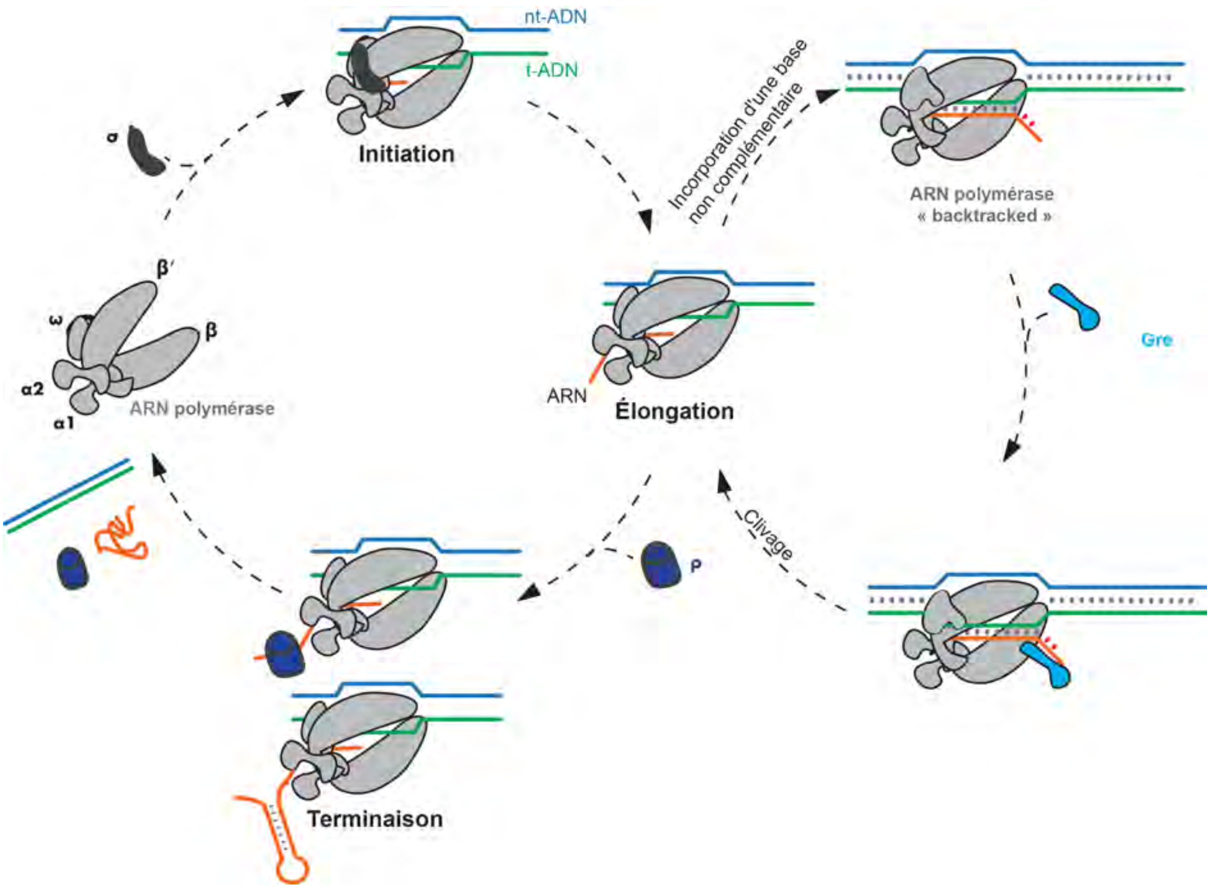
Résumé de Thèse

Dans les trois règnes de la vie, l'expression des gènes commence par la transcription de l'ADN en ARN messager (ARNm) par des ARN polymérases ADN-dépendantes (RNAP) à plusieurs sous-unités. Malgré la diversité de leurs compositions en sous-unités, l'architecture centrale et les mécanismes catalytiques sont conservés dans les RNAP de tous les organismes. Le site actif de la RNAP catalyse deux réactions : l'addition du substrat et l'hydrolyse des liaisons phosphodiester. Cette dernière est nécessaire pour sauver les complexes d'élongation de la RNAP qui ont subi un retour en arrière, résultant en un complexe incapable d'étendre l'ARN naissant. Dans un complexe qui a fait marche arrière, l'extrémité 3' de l'ARN est poussée hors du site actif et dans une région de l'enzyme appelée canal secondaire. Outre son rôle dans les pauses régulatrices, le backtracking est également nécessaire pour la relecture : des NTP mal incorporés font que le complexe d'élongation favorise le backtracking de 1 à 2 nucléotides. Chez E. coli*, le clivage de fragments d'ARN plus courts est accru en présence du facteur de transcription GreA (fonctionnellement analogue au TFIIS eucaryote), qui est connu pour aider au clivage grâce à deux résidus acides hautement conservés. Les structures cryo-EM d'un complexe rétrogradé d'un nucléotide avec et sans le facteur de clivage GreA lié au canal secondaire ont été utilisées pour répondre aux questions liées au processus de relecture dans la RNAP de* E. coli. *RNAP, en particulier celles de l'importance du motif structurel de la RNAP connu sous le nom de boucle de déclenchement, et du processus de sélection de GreA parmi d'autres facteurs de transcription structurellement similaires. En outre, les résultats des essais de transcription in vitro ont également mis en évidence le rôle de GreA dans la relecture, en dehors des interactions impliquant ses résidus acides.*

Introduction

L'expression de l'information génétique chez tous les êtres vivants commence par la transcription, le processus de copie de l'ADN en ARN messager simple brin complémentaire (ARNm). La transcription est universellement réalisée par les ARN polymérases ADN-dépendantes (RNAP). Ce sont des enzymes à sous-unités multiples chez tous les organismes mono et pluricellulaires et elles existent également en tant qu'enzymes à sous-unités uniques chez les bactériophages, les mitochondries et les chloroplastes. Chez la bactérie *Escherichia coli*, la RNAP est composée de cinq sous-unités : $\alpha 1$, $\alpha 2$, β, β', et ω. β et β' sont les deux plus grandes sous-unités et forment les pinces qui donnent à la RNAP son aspect caractéristique de pince de crabe. La composition des sous-unités de la RNAP dans les trois règnes de la vie - bactéries, archées et eucaryotes - varie, mais l'architecture de base et les mécanismes fonctionnels restent conservés (Werner et Grohmann 2011). Cela est particulièrement vrai pour la plus grande sous-unité (β' chez les bactéries, Rpo1 chez les archées et Rpb1 pour l'ARN Pol-II des eucaryotes). La compréhension de ces mécanismes communs chez *E. coli* a donc des implications sur la compréhension des processus équivalents chez d'autres organismes, y compris chez l'homme. Ceci est extrêmement important puisque de nombreuses maladies humaines sont causées par une régulation défectueuse de la transcription.

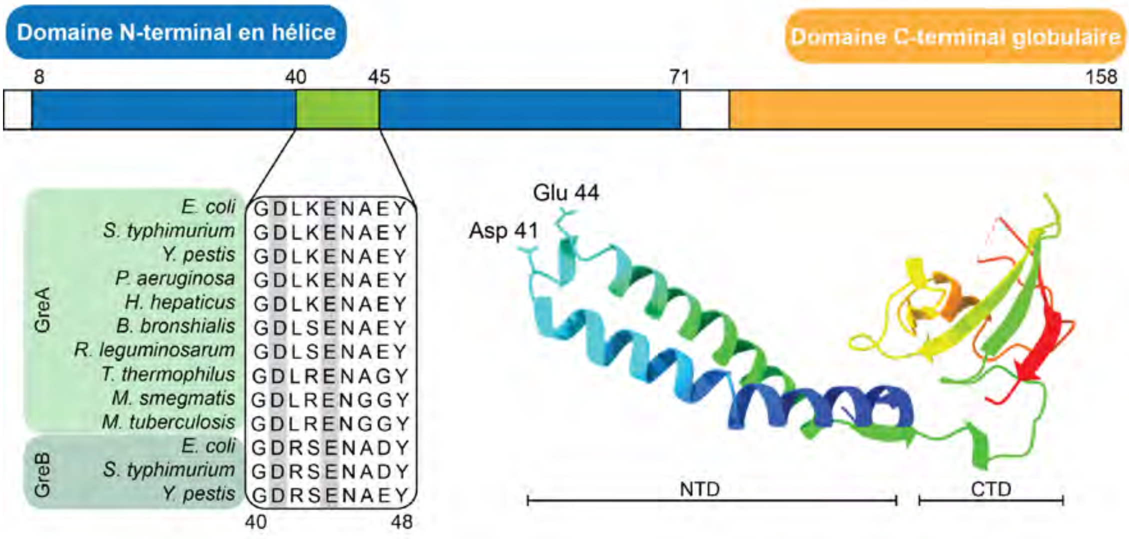
La transcription s'effectue selon un processus cyclique, commençant par l'initiation et se terminant par la terminaison (Figure (1)). La RNAP contient trois canaux qui servent de points d'entrée et de sortie au noyau de l'enzyme. Il s'agit du canal principal, du canal secondaire et du canal de sortie de l'ARN. Au cours de la transcription, l'ADN double brin pénètre dans la RNAP par le canal principal et sa double hélice est ouverte pour faciliter l'entrée du brin matrice dans le site actif de l'enzyme. Des nucléosides triphosphate (NTP) complémentaires du brin d'ADN matrice sont utilisés pour étendre l'ARN naissant, un nucléotide à la fois. Les NTP entrent dans le noyau de la RNAP par le canal secondaire. L'extension de l'ARN se déroule comme suit : un NTP complémentaire de la base d'ADN se lie dans le site actif (site A), et une nouvelle liaison est formée par



(1) Illustration du cycle de transcription chez les bactéries, de l'initiation à l'élongation et enfin à la terminaison. L'initiation a lieu avec l'aide du facteur "sigma", et la terminaison peut se produire avec ou sans le facteur "rho". Dans cette thèse, l'accent est mis sur la phase d'élongation, où l'ARN est étendu d'une base à la fois. En cas d'erreur, la rétrogradation est privilégiée. Pour poursuivre l'élongation à partir de cet état, l'ARN polymérase doit cliver l'ARN extrudé. Cette réaction est plus efficace en présence du facteur GreA.

une réaction de substitution nucléophile médiée par Mg^{2+}. Grâce à la translocation vers l'avant de la RNAP par rapport à l'ADN, la nouvelle extrémité 3' de l'ARN est ensuite déplacée vers le site P, adjacent au site A. Au fur et à mesure que l'ARN s'allonge, il est guidé hors de l'enzyme par le canal de sortie de l'ARN. Les mécanismes de criblage dans le noyau de la RNAP assurent généralement la liaison du NTP correct en éliminant à la fois les désoxy-nucléosides triphosphates (dNTP) et les NTP non complémentaires. Bien qu'il existe un processus de sélection approfondi pour sélectionner les NTP complémentaires, il arrive qu'une base non complémentaire soit ajoutée à l'ARN. Un résultat direct de ce type d'incorporation erronée est généralement la translocation inverse de la RNAP, qui pousse l'extrémité 3' de l'ARN dans la direction opposée. C'est ce qu'on appelle le « backtracking

», un processus qui sert également d'autres objectifs en stabilisant la RNAP sur les sites de pause permettant la régulation de la transcription (Nudler et al. 1997 ; Komissarova et Kashlev 1997). Au lieu d'être positionnée sur le site P, l'extrémité 3' de l'ARN naissant est poussée vers une région de l'enzyme connue sous le nom de canal secondaire (SC), qui sert également de voie d'entrée pour les NTP et les ions Mg^{2+} entrants. La RNAP bloquée dans son état de recul, ou rétrogradée, ne dispose pas de son site A pour la liaison de nouveaux substrats NTP, ce qui bloque la transcription.



(2) La figure montre la structure de GreA résolue par cristallographie. La barre en haut donne les numéros de résidus pour les domaines N-terminal et C-terminal, qui sont également indiqués dans la figure de la structure de GreA. La section en vert au milieu du NTD est la pointe de GreA qui entre en contact avec le site actif. Les alignements de séquences sur la gauche concernent cette région, à la fois dans GreA et GreB.

Pour que la RNAP bloquée dans son état de recul puisse reprendre la transcription, la partie extrudée de l'ARN doit être coupée. Le clivage de l'ARN peut être effectué par la RNAP elle-même, grâce à un mécanisme endo-nucléolytique intrinsèque (Orlova et al. 1995). Cependant, ce processus est accéléré en présence de certains facteurs de transcription se liant au SC. Chez les bactéries, il s'agit des facteurs Gre (Borukhov, Bagitov et Goldfarb 1993), et ils sont fonctionnellement analogues au facteur TFIIS chez les eucaryotes et TFS chez les archées. TFIIS n'influence que le clivage de l'ARN Pol-II. Dans Pol-I et Pol-III, les enzymes utilisent des sous-unités spécifiques pour effectuer le clivage. *E. coli* ainsi que d'autres espèces de bactéries ont deux formes du facteur Gre, GreA et GreB. Alors que GreA participe au clivage des di- et tri-nucléotides, GreB participe au clivage d'ARN plus longs. Cela a été prouvé par des expériences qui ont comparé la préférence de chaque facteur de transcription pour différentes longueurs d'ARN (Koulich et al. 1997). En raison de sa préférence pour les produits de clivage plus courts, GreA est impliqué dans la fidélité de la transcription. Structurellement, GreA et GreB sont tous deux similaires à d'autres facteurs de liaison du SC bactériens comme DksA et Gfh1, en ce sens qu'ils possèdent un domaine C-terminal globulaire et un domaine N-terminal en hélice (Figure (2)). C'est ce dernier qui entre dans le SC de la RNAP, permettant à la boucle en épingle à cheveux à son extrémité d'entrer en contact avec le site actif. On sait que les domaines N-terminaux de GreA et GreB participent au clivage de l'ARN extrudé grâce à des interactions avec le site A impliquant deux résidus acides universellement conservés – un aspartate et un glutamate (Asp 41 et Glu 44 dans *E. coli* GreA). En dehors des deux résidus catalytiques, de nombreux autres résidus de la pointe sont également conservés dans GreA et GreB. La principale différence dans les séquences de ces protéines se situe dans la partie restante du domaine N-terminal. Le domaine N-terminal de GreB contient un grand patch basique, qui est important pour stabiliser l'ARN le plus long dans le canal secondaire. Étant donné que GreA se lie à des complexes rétrocontrôlés par une courte longueur, ils n'ont pas besoin d'un patch basique pour stabiliser l'ARN dans le canal secondaire. Des différences supplémentaires entre les deux sont également observées dans leur affinité avec l'ARN polymérase et leur abondance dans la cellule. Alors que GreA est plus abondant mais avec une affinité plus faible, GreB est moins abondant avec une affinité plus élevée. Il est intéressant de noter que des expériences ont montré qu'il existe un certain degré de similarité fonctionnelle entre GreA, GreB et une autre protéine de liaison au SC, DksA. En effet, la surexpression d'une protéine dans des souches contenant des délétions de gènes codant pour une autre peut compenser cette perte. Cependant, le point le plus important à noter est que toutes les bactéries ne contiennent pas à la fois GreA et GreB. Des deux, GreA est trouvé dans toutes les espèces de bactéries. Cela s'explique par l'importance de ce facteur dans le maintien de la fidélité de la transcription. D'autre part, GreB n'est présent que dans certaines espèces. Ceci, ainsi que les différences d'affinité de liaison et de concentration dans la cellule, conduisent à quelques questions. Comment les ARN polymérases de certaines espèces font-elles la distinction entre GreA et GreB lorsqu'ils sont tous deux présents, et quelle influence spécifique GreA exerce-t-il sur les complexes de transcription ?

La sous-unité β' de la RNAP contient de nombreux modules structurels qui sont largement conservés chez de nombreuses espèces. Parmi ceux-ci, deux des plus pertinents pour ce travail sont la « Bridge Helix » (BH) et la « Trigger Loop » (TL). La BH est connue pour faciliter la translocation de la RNAP le long du brin d'ADN matrice via un mécanisme de cliquet provoqué par la "torsion" de l'hélice près du site actif (Bar-Nahum et al. 2005). Ce changement de conformation du BH est observé pour un certain nombre d'états du complexe, y compris un état de pause (Sekine et al. 2015). Lors de l'ajout de nucléotides, la TL se structure en un faisceau connu sous le nom de « Trigger Helices ». Le TL contient un résidu histidine universellement conservé, ou invariant (His 936 dans l'ARN polymérase d'*E. coli*, et His 1242 dans l'ARN polymérase d'*Thermus aquaticus*). Lorsqu'elle se replie dans les hélices, cette histidine peut entrer en contact avec le site actif et participer aux réactions catalytiques. Il a été démontré que la TL est un élément très dynamique de la RNAP, qui passe de sa conformation structurée à sa conformation non structurée à chaque cycle d'ajout de nucléotides (Mazumder et al. 2020 ; Vassylyev et al. 2007 ; Wang et al. 2006). Cependant, le rôle exact de la TL pendant le clivage, n'est pas clair. Certaines études ont montré que les délétions et les mutations de la TL n'affectent pas de manière significative le clivage intrinsèque de la RNAP, tandis que d'autres ont montré que ces changements affectent la capacité de l'enzyme à cliver l'ARN. Grâce à une analyse biochimique *in vitro* de l'hydrolyse de l'ARN dans les ARN polymérases de *T. aquaticus* et de *E. coli*, Yuzenkova et Zenkin (2010) a montré que la participation du TL est essentielle. En particulier, ils ont également montré que le résidu histidine invariant est essentiel pour le clivage. Cependant, la même année, Zhang, Palangat et Landick (2010) a montré que le repliement du TL n'est pas essentiel pour le clivage. Ils l'ont démontré en effectuant des essais *in vitro* avec l'ARN polymérase d'*E. coli* contenant des mutations qui empêchent le repliement du TL. Les données provenant des structures de Pol-II suggèrent également que l'histidine n'entre probablement pas en contact avec le site actif, puisque le TL est dans la conformation ouverte (Wang et al. 2009). Plus récemment, les travaux de Mishanina et al. (Mishanina et al. 2017) ont mis en évidence le rôle de la TL comme catalyseur positionnel dans le clivage. L'ensemble de ces résultats soulève quelques questions sur le rôle spécifique du TL dans le clivage. Si la formation des hélices et la mise en contact de l'histidine avec le site actif ne sont pas importantes, alors comment le TL participe-t-il au clivage hydrolytique ?

Le travail décrit dans cette thèse a été planifié pour répondre à certaines questions clés concernant la fidélité de la transcription et le clivage :

1. Pouvons-nous capturer un complexe RNAP rétrogradé dans un état pré-catalytique, avec le facteur GreA de type sauvage présent dans le canal secondaire ? Ceci est dû au fait que parmi les diverses structures de facteurs de clivage liés à l'ARN polymérase publiées jusqu'à présent, aucune ne contient GreA non modifié. Chez les bactéries, on a toujours préféré utiliser GreB en raison de sa plus grande affinité, mais GreB n'est pas impliqué dans la relecture. La cryo-EM nous permet de capturer le complexe avec GreA, même sans aucune mutation.

2. De quelle manière précise GreA influence-t-il le clivage des ARNm rétrogradés ? Cela pourrait également aider à répondre aux questions relatives à l'existence de deux facteurs de clivage distincts chez certaines espèces de bactéries, par exemple comment la polymérase sélectionne GreA et non GreB en présence d'un ARN rétrogradé court.

3. La TL affecte-t-elle ce clivage, et si oui, de quelle manière ? Une structure à haute résolution pourrait répondre à la question jusqu'ici sans réponse du rôle du TL . Si elle n'entre pas en contact avec le site actif pendant le clivage, comment influence-

t-elle exactement la réaction ?

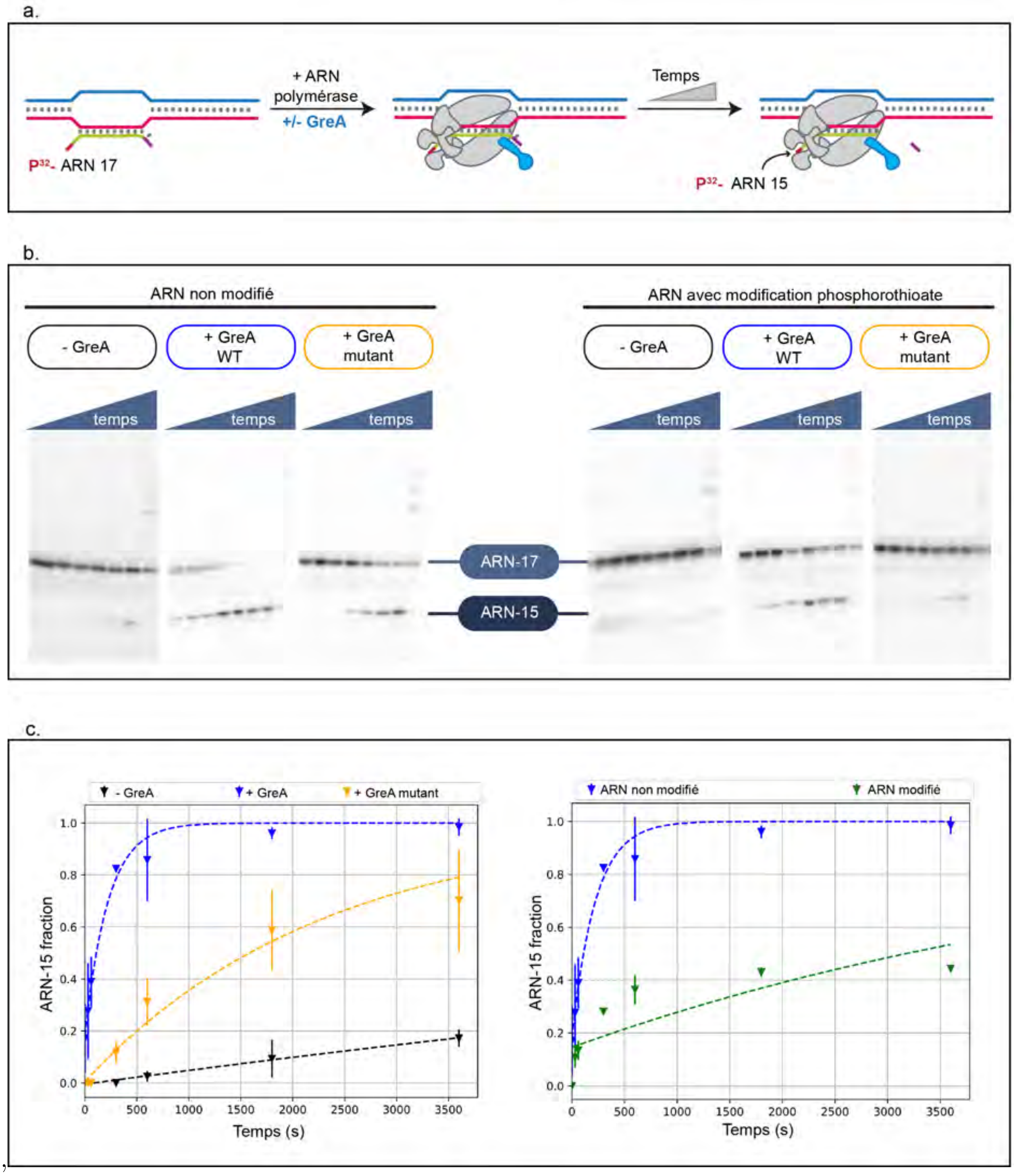
Résultats

Des essais de transcription in vitro ont été réalisés pour tester à la fois la fonctionnalité et l'efficacité de clivage de complexes d'élongation de la transcription rétrogradés. Ces complexes sont constitués de la RNAP, d'un « scaffold » mimant une bulle de transcription (un fragment d'ADN double brin auquel vient s'hybrider un fragment d'ARN complémentaire au brin d'ADN matrice), et du facteur GreA. Les « scaffolds » ont été conçus de manière à ce que l'ADN présente un décalage de 10 paires de bases en son centre, imitant ainsi la bulle centrale qui se forme dans tous les complexes de transcription *in vivo*. Le brin d'ARN contient une section qui est complémentaire à l'un des brins d'ADN dans cette région. Les principaux résultats sont présentés dans la figure (3). Les premiers tests de complexes d'élongation rétrogradés par un et deux nucléotides ont montré que les complexes rétrogradés par un nucléotide donnaient un produit de clivage di-nucléotidique plus homogène. À cette époque, il a également été observé que le clivage assisté par GreA de type sauvage (WT) était trop rapide pour capturer le complexe dans son état pré-catalytique.En contraste avec cela, le clivage de l'ARN sans le GreA WT était plus faible d'environ 100x. La réaction a également été réalisée en présence d'un mutant de GreA dans lequel les deux résidus acides catalytiques ont été mutés en alanines. Le clivage en présence du mutant GreA était plus rapide que le clivage sans aucun GreA, mais plus lent que le clivage avec le GreA WT. Il a donc été décidé que la meilleure stratégie serait d'assembler des « scaffolds » à l'aide d'ARN contenant des modifications de phosphorothioate. Une modification phosphorothioate dans un acide nucléique consiste à remplacer un oxygène non pontant du squelette phosphate par un atome de soufre. Puisque les oxygènes du squelette phosphate de l'ARN dans le site actif contribuent à la coordination des ions Mg^{2+} nécessaires au clivage, les modifications phosphorothioate de l'ARN dans cette région réduiraient la coordination des ions Mg^{2+}. Par conséquent, la vitesse de la réaction de clivage serait réduite sans modifier les propriétés structurelles des acides nucléiques dans la structure. Ce site permettrait donc la reconstruction en une seule partie d'un complexe de pré-clivage avec GreA WT lié.

Des essais comparant les taux de clivage du même complexe avec l'ARN modifié et non modifié ont en effet montré que le taux de clivage avait suffisamment ralenti laissant ainsi le temps de congeler un échantillon sur une grille de cryo-microscopie électronique (cryoEM) et donc de capturer le complexe souhaité avant que le clivage n'ait lieu. Avec la cryo-EM, la première étape est l'optimisation des conditions pour obtenir une épaisseur de glace optimale pour le spécimen et une contamination minimale. Après avoir recherché les meilleures conditions de préparation de l'échantillon pour la cryoEM, une première reconstruction a révélé un complexe avec GreA lié au SC. Cette reconstruction ayant permis de confirmer la présence de tous les partenaires, nous avons par la suite procédé à la collecte d'un ensemble d'images à haute résolution.

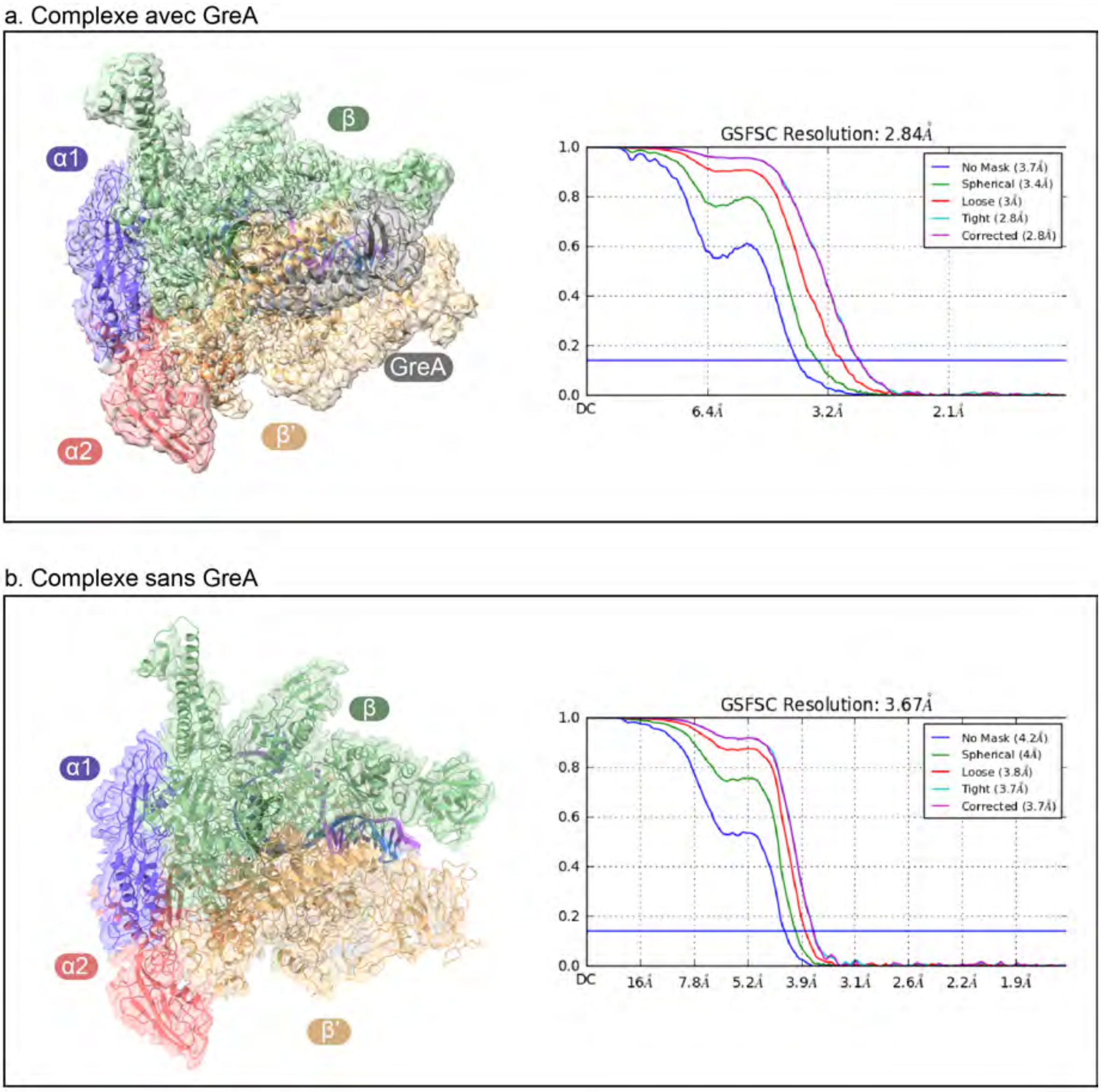
Ces données ont été affinées à une résolution de 3 Å, et ont permis d'obtenir un modèle tridimensionnel (3D) montrant la base rétrogradée. Par la suite, trois autres jeux de données à haute résolution ont été collectés, chacun avec une base rétrogradée différente, afin d'observer d'éventuelles différences structurelles entre les complexes. Ces 3 complexes ont été affinés à des résolutions de 3,7, 3,8 et 4,2 Å. Bien que les résolutions n'aient pas été aussi élevées que celles du premier complexe, la densité électronique des chaines polypeptidiques a montré que les conformations générales des domaines individuels au cœur de la RNAP étaient les mêmes. L'analyse ultérieure du complexe de pré-clivage lié à GreA (GreA-BC) a ensuite été effectuée uniquement en se basant sur la reconstruction à la plus haute résolution (Figure (4)a.).

L'analyse de la structure de pré-clivage a montré la TL dans une conformation intéressante et nouveau. Alors qu'on s'attendait à ce que la TL soit trouvée dans une conformation ouverte non structurée avec l'histidine conservée dirigée loin du site A, l'état dans lequel elle a été observée était différent de ce qui avait été observé avec un complexe similaire lié à GreB (Figure (5)). Les différences entre les structures du complexe de pré-clivage GreB et du complexe avec GreA sont les longueurs de l'ARN extrudé (le complexe GreB contenait un ARN rétrograde plus long) et les facteurs de transcription eux-mêmes. La nouvelle conformation de la TL suggère que la liaison de GreA elle-même



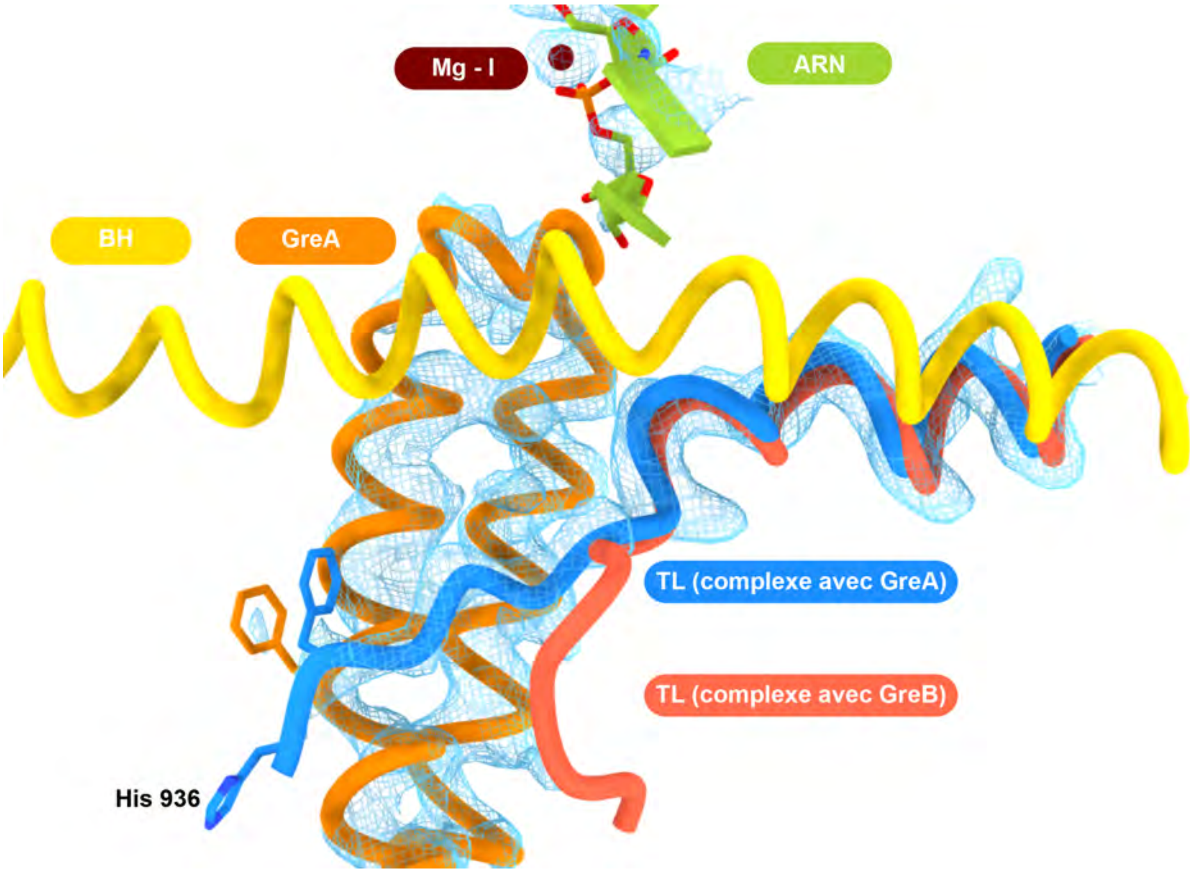
,

(3) a. Schéma de la procédure expérimentale pour les essais de transcription *in vitro*. b. Gels montrant la progression du clivage (de l'ARN-17 à l'ARN-15) dans différentes conditions de réaction pour un temps de 60 minutes chacun. L'ajout d'une modification phosphorothioate (gels de droite) à l'ARN ralentit la réaction. c. Graphiques représentant les résultats ci-dessus. Les demi-vies pour les graphiques de gauche sont de 212,8 minutes, 26,6 minutes, et 2,6 minutes pour les conditions sans GreA, avec GreA, et avec le mutant GreA. Le graphique de droite illustre la différence des taux de clivage avec et sans modification de l'ARN.



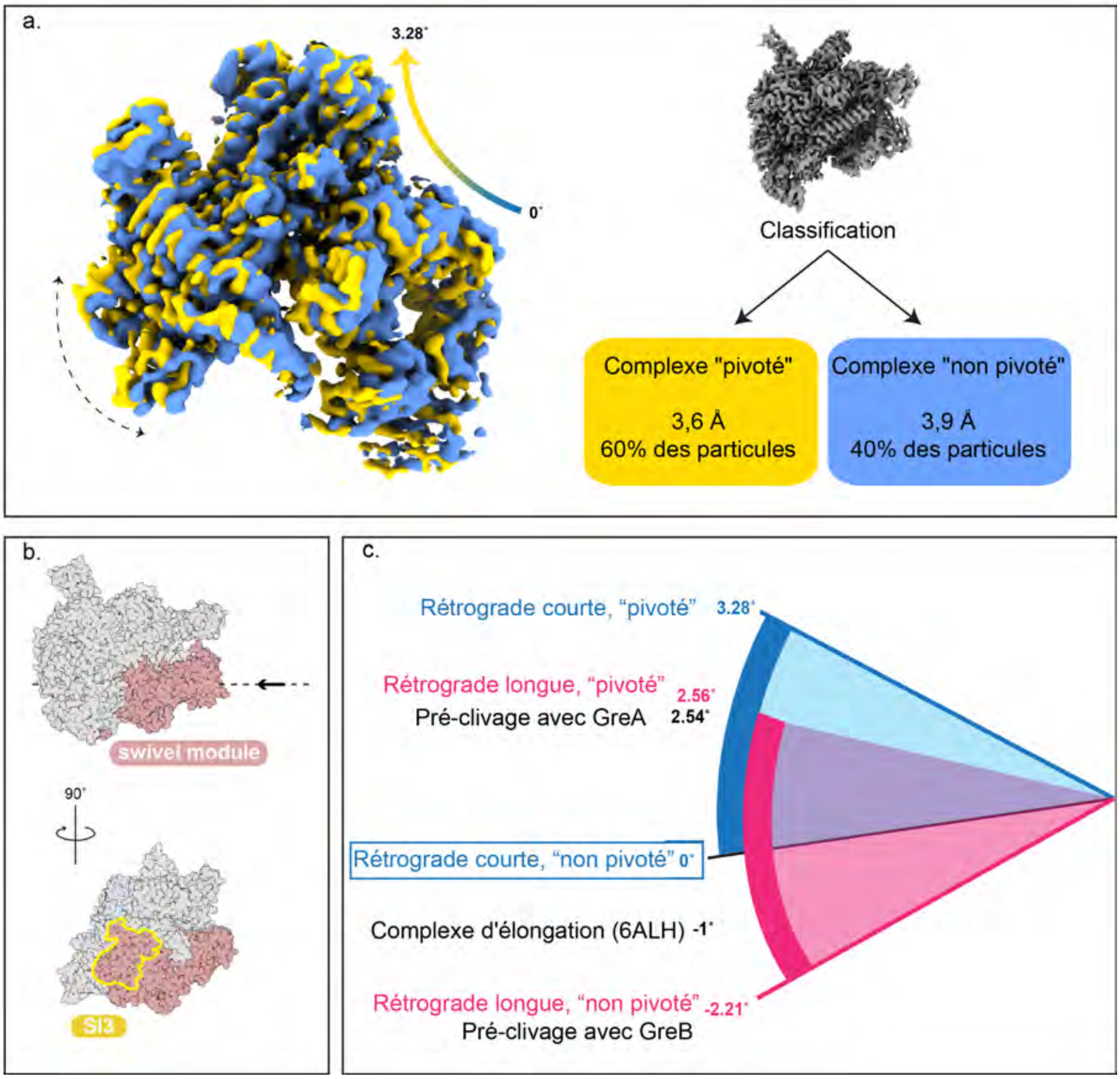
(4) Deux structures principales résolues par cryo-EM : a. Complexe de pré-clivage avec GreA, et b. Complexe d'élongation rétrogradé d'un nucléotide sans GreA. Les figures à gauche des deux panneaux montrent les reconstructions (volumes transparents) ainsi que les structures modélisées (rubans solides). A droite se trouvent les courbes FSC montrant les résolutions pour les reconstructions.

pourrait influencer la conformation ouverte spécifique que la TL adopte. Pour approfondir cette question, nous avons collecté un ensemble de données supplémentaires avec le même échantillon mais sans GreA (Figure (4)b.). Cela nous permettrait de voir l'influence que la liaison à GreA a sur la conformation de la TL ainsi que sur d'autres éléments du noyau enzymatique. Comme pour les ensembles de données précédents, les grilles de cryo-EM préparées avec l'échantillon ont été passées au crible pour trouver les meilleures conditions afin d'obtenir le meilleur ensemble de données possible. La classification 3D des particules pour le complexe rétrogradé (BC) a révélé deux classes : la première classe, comprenant environ 40% des particules, permet d'obtenir un modèle affiné à 3,9 Å, tandis que les 60% restants de la seconde classe ont atteint 3,6 Å (Figure (5)). La principale différence de conformation entre les deux classes se situe au niveau du motif structurel appelé « shelf » et des pinces de la RNAP. Les classes 1 et 2, appelées respectivement "pivotées" et "non pivotées", ressemblent aux classes pivotées et non pivotées observées dans des travaux publiés précédemment (Abdelkareem et al. 2019). La TL n'est pas aussi bien résolue lorsqu'on la compare à d'autres éléments dans la même région, indiquant qu'il pourrait ne pas être maintenu en place dans une conformation stable comme c'était le cas dans le complexe GreA. Cependant, on a pu constater que dans le complexe non pivoté, la TL est orientée légèrement différemment de ce qui a été observé dans le modèle GreA-BC. Les observations les plus importantes pour le complexe rétrogradé sans GreA se trouvent dans les mesures des degrés de rotation du « swivel module ». Ces mesures sont illustrées dans la Figure (6)c, les degrés de rotation du même module dans d'autres complexes apparentés sont également indiqués. La première observation est que le degré de rotation est plus pivotant lorsque le complexe a un ARN rétrogradé plus court. Pour un ARN rétrogradé long (plus long de seulement 2 nucléotides), il y a un changement clair vers plus de "non-pivot". La deuxième observation concerne les conformations adoptées par les complexes contenant GreA et GreB. Comme mentionné précédemment, GreA et GreB ont des préférences pour les complexes rétrogradés par différentes longueurs d'ARN. Dans la figure, nous voyons que le complexe avec GreA adopte une conformation pivotée, en comparaison avec la conformation non pivotée adoptée par le complexe avec GreB. Ceci



(5) La figure montre la différence dans l'élément du site actif « trigger loop » (TL) entre les complexes contenant GreA et GreB (TL colorés en bleu et rouge, respectivement). Les éléments supplémentaires montrés sont : « bridge helix » (jaune), l'ion magnésium (marron), l'ARN (vert clair), et GreA (orange). La reconstruction de la carte pour le complexe GreA est également représentée sous la forme d'une maille bleu clair. L'image montre clairement que la TL dans le complexe avec GreA est orientée vers le facteur de transcription, indiquant une interaction spécifique ayant lieu entre eux.

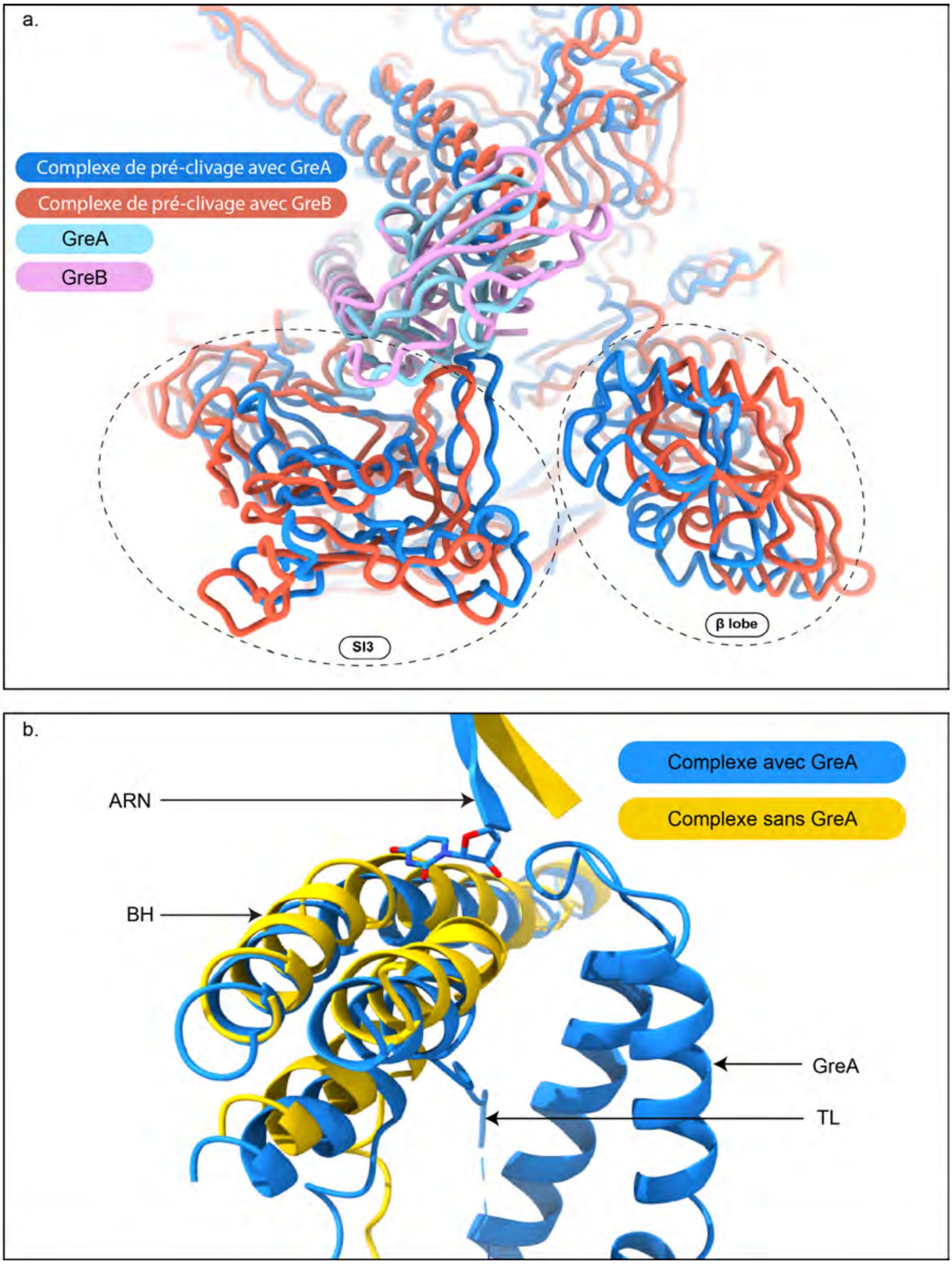
indique clairement un mécanisme dans lequel le *E. coli* ARN polymérase peut sélectionner soit GreA soit GreB, en fonction de l'étendue de la rotation. Pour un ARN rétrogradé court, le TL a plus de liberté de mouvement, permettant au complexe de se déplacer vers une gamme pivotée. Lorsque l'ARN est plus long, il est poussé plus loin dans le canal secondaire, limitant le mouvement du TL. Une partie importante du « swivel module » est le SI3 spécifique à *E. coli*, qui est également une extension du TL. Il est donc logique que les conformations autorisées du TL se reflètent dans les états autorisés du « swivel module », ce qui entraîne la sélection du complexe pour GreA ou GreB en fonction de la longueur de l'ARN dans le canal secondaire.


a.
3.28°
0°
Classification
Complexe "pivoté"
3,6 Å
60% des particules
Complexe "non pivoté"
3,9 Å
40% des particules
b.
swivel module
90°
SI3
c.
Rétrograde courte, "pivoté" 3.28°
Rétrograde longue, "pivoté" 2.56°
Pré-clivage avec GreA 2.54°
Rétrograde courte, "non pivoté" 0°
Complexe d'élongation (6ALH) -1°
Rétrograde longue, "non pivoté" -2.21°
Pré-clivage avec GreB

(6) a. La classification du complexe rétrogradé a révélé deux classes - "pivoté" et "non pivoté", colorées respectivement en jaune et en bleu. La différence entre ces classes réside dans la rotation d'une région appelée « swivel module », représentée sur la gauche. b. Schéma montrant la position du « swivel module » dans l'ARN polymérase. L'insertion SI3 spécifique à *E. coli* est marquée en jaune. c. Gamme d'angles de rotation adoptés dans différentes conditions.Tous les angles ont été mesurés par rapport au court rétrogradé "non pivoté". La plage de rotation pour les complexes rétrogradés longs et courts est représentée en rose et en bleu, marquant la nette différence entre eux.

Dans le modèle GreA-BC, nous avons vu que les deux résidus acides conservés du domaine N-terminal (Aspartate 41 et Glutamate 44 dans *E. coli*) sont positionnés près du site A. En raison des interactions avec le faisceau d'électrons dans le microscope avec les chaînes latérales de ces résidus, il n'y a pas beaucoup de densité pour eux dans les cartes de densités électroniques. Les deux résidus acides sont conservés dans presque toutes les variantes bactériennes des facteurs Gre, et certains des résidus voisins sont également conservés dans différentes espèces. Lors des premiers tests visant à trouver le complexe optimal pour la cryoEM, une version mutante de GreA a été utilisée comme témoin, dans laquelle chacun des deux résidus acides a été muté en une alanine. On espérait que le taux de clivage en présence du mutant GreA serait comparable au taux de clivage en l'absence de GreA. Cependant, cela n'a pas été le cas. Au contraire, nous avons vu que le taux de clivage de l'ARN rétrogradé en présence du mutant GreA était plus proche de la réaction observée avec GreA WT. Les observations faites concernant le mutant GreA deviennent plus claires lorsque nous prenons en compte les données structurelles. La nouvelle forme du TL qui le montre ouvert mais dirigé vers GreA explique pourquoi, même lorsque les résidus catalytiques à l'extrémité sont mutés, GreA conserve une partie de sa fonction de facteur de clivage. Les deux éléments semblent interagir l'un avec l'autre, spécifiquement sous la forme d'une interaction phénylalanine-phénylalanine (vue dans la figure (5)). L'explication de la raison pour laquelle cette nouvelle conformation n'est observée qu'avec GreA et non GreB est également simple : les résidus de la NTD de GreA qui sont à proximité du TL ne se trouvent pas dans GreB.

Un autre résidu important dans la pointe du domaine N-terminal de GreA est la Lysine 43, qui est conservée chez de nombreuses espèces. L'analyse de la structure GreA-BC montre que cette lysine est dirigée vers la chaine phosphatée de l'ARN rétrogradé. Les interactions entre la lysine et le phosphate de l'ARN pourraient aider à le stabiliser au sein du site A et à le cliver. Pour approfondir à la question de savoir si et comment GreA influence la conformation de la TL, nous avons analysé les complexes avec et sans GreA. Le modèle GreA-BC a montré des interactions stacking de cycles aromatiques entre deux chaînes latérales de phénylalanine - l'une dans la TL et l'autre dans la pointe du domaine



(7) a. La différence de positionnement du SI3 en fonction de la longueur de l'ARN extrudé se reflète également dans le positionnement du SI3 dans les complexes avec GreA et GreB liés. Pour une longueur d'ARN courte, le complexe est plus pivoté, ce qui rapproche le SI3 et le lobe β l'un de l'autre (en bleu). b. La liaison de GreA affecte également les éléments du site actif. Les deux structures présentées ici, avec et sans GreA en bleu et jaune respectivement, montrent que GreA positionne la dernière base de l'ARN et la TL et permet à la réaction de se dérouler efficacement. Sans GreA, le BH peut entrer en conflit avec la dernière base de l'ARN.

N-terminal de GreA. Nous avons également remarqué que dans la classe rétrogradée à plus haute résolution, la conformation ouverte adoptée par la TL semble se structurer vers le site de liaison de GreA. Enfin, il y a une nette différence dans le positionnement de la BH dans les structures BC et GreA-BC. Dans les trois structures, on observe une courbure de la BH, comme prévu pour un complexe d'élongation dans cet état. Cependant, le positionnement exact de la BH dans le modèle GreA-BC est distinct de celui du modèle BC, et une superposition des trois a révélé que les chaînes latérales du BH dans le modèle BC se heurtent à la position de l'ARN rétrogradé dans le modèle GreA-BC (Figure (7)b.). Cela semble montrer qu'en plus de se comporter comme un contributeur catalytique au clivage (par le biais des deux résidus acides), GreA fonctionne également en positionnant les éléments nécessaires au bon déroulement de la réaction. Ces résultats sont en accord avec l'article de Mishanina et al. (2017) dans lequel, grâce à des expériences biochimiques, ils ont suggéré le même mécanisme.

Conclusion

Les structures cryo-EM des complexes d'élongation de transcription rétrogradés avec et sans le facteur de clivage GreA ont été résolues, ce qui permet de mieux comprendre les mécanismes impliqués dans la correction de la transcription. Une compréhension complète des mécanismes de correction de transcription est essentielle, car elle permet de comprendre comment l'ARN polymérase bactérienne corrige les erreurs d'incorporation de nouveaux substrats NTP dans l'ARN et, si nécessaire, sélectionne un facteur de transcription spécifique pour effectuer la réaction plus efficacement.

Quatre complexes de pré-clivage avec GreA lié au canal secondaire de l'ARN polymérase, chacun rétrogradé par une base différente, ont montré que la mauvaise incorporation spécifique, ou base rétrogradée, n'entraîne pas de différences structurelles dans le site actif et n'influence pas le mécanisme de réaction. Parmi ces quatre structures, la carte reconstruite du complexe de pré-clivage avec la base rétrogradée 'U' a atteint une résolution de 2,8AA, révélant un élément particulier du noyau catalytique de l'enzyme, la « trigger loop », dans une nouvelle conformation. Pour évaluer pleinement l'impact de GreA sur le clivage, un cinquième ensemble de données cryo-EM a été collecté sur un complexe de 1-nucléotide retourné à l'envers en l'absence de GreA. Les données de ce complexe ont révélé une gamme continue d'angles de rotation d'un motif structurel connu sous le nom de « swivel module », les particules de l'ensemble de données convergeant vers deux états : "pivoté" et "non pivoté".

Des comparaisons des structures du complexe de pré-clivage de GreA et des deux états du complexe rétrogradé avec des structures similaires de complexes rétrogradés à 3 nucléotides de Abdelkareem et al. (2019) ont révélé que l'étendue du rétrogradage pouvait influencer la plage dans laquelle le module pivotant peut tourner. En outre, la liaison de GreA ou GreB au secondary channel (SC) maintient également le module pivotant dans un état spécifique, le complexe de préliaison de GreA adoptant une conformation plus pivotante. Les états autorisés de ce module, qui dépendent de la longueur de l'ARN extrudé dans le canal secondaire, permettent de comprendre comment, chez une espèce comme le *E. coli* qui contient deux facteurs de clivage, elle peut choisir l'un ou l'autre en fonction de la longueur de l'ARN. Les observations relatives au "module pivotant" et au domaine SI3 ne concernent que le *E. coli* dans ce travail, mais elles nous permettent de comprendre comment d'autres espèces bactériennes peuvent également sélectionner un facteur de transcription lorsque plus d'un est présent. Il est important d'indiquer très clairement que chez les eucaryotes et les archées, il n'y a pas de SI3 et que les facteurs de clivage TFIIS et TFS sont structurellement très différents de GreA.

GreA semble également positionner directement et indirectement les éléments autour du site actif pour qu'un clivage efficace ait lieu. On pensait auparavant que GreA fonctionnait uniquement grâce aux résidus acides conservés à l'extrémité de sa bobine NTD en contact avec le site actif. Les données structurelles et biochimiques décrites ici suggèrent plutôt un modèle de clivage assisté par GreA dans lequel le facteur de transcription ne participe pas au clivage uniquement via ses résidus acides. Au contraire, il le fait par le biais d'interactions supplémentaires impliquant à la fois la NTD de GreA (avec les éléments du noyau enzymatique) et sa CTD (avec la surface). Cette observation pourrait avoir des implications dans la compréhension du rôle des facteurs de clivage dans d'autres domaines de la vie. Alors que TFIIS et TFS sont structurellement différents de GreA, ils participent également par le biais de résidus à proximité du site actif. Une analyse plus approfondie des résidus dans ces espèces pourrait très probablement montrer que ces autres facteurs de clivage fonctionnent également d'une manière similaire à GreA – en tant que participant à la fois catalytique et fonctionnel dans la réaction de clivage.

La compréhension de la transcription et le maintien de la fidélité de la transcription sont extrêmement cruciaux pour la compréhension globale de l'expression des gènes. Le travail décrit dans la thèse n'offre qu'une pièce du puzzle, mais avec l'utilisation plus répandue de la cryo-EM pour déterminer la structure des complexes biomoléculaires, nous sommes plus près d'obtenir une image universelle complète des mécanismes de fidélité dans tous les domaines de la vie.

Publications

- M. W. Webster, M. Takacs, C. Zhu, V. Vidmar, **A. Eduljee**, M. Abdelkareem, and A. Weixlbaumer, "Structural basis of transcription-translation coupling and collision in bacteria", Science, Aug. 2020, issn : 0036-8075
- (En cours) **A. Eduljee**, C. Saint-André, M. Takacs, A. Durand, A. Weixlbaumer, "Structural and Biochemical Insights into Transcriptional Proofreading"

Présentations

- Poster : "Structural and Biochemical Studies on the Fidelity of Transcription" **A. Eduljee**, C. Saint-André, M. Takacs, T. Cheng, X. Guo, J. Ortiz, A. Weixlbaumer. *Cryo 3D Electron Microscopy Course, International School of Crystallography Juillet 2019 à Erice, Italie*
- Poster : "Structural and Biochemical Studies on the Fidelity of Transcription" **A. Eduljee**, C. Saint-André, M. Takacs, T. Cheng, X. Guo, J. Ortiz, A. Weixlbaumer. *Poster Session Interne, Février 2020 à l'IGBMC*

Table of Contents

List of Tables

List of Figures

List of Abbreviations

A-site	acceptor site
ATP	adenosine triphosphate
BC	backtracked complex
BH	bridge helix
cryo-EM	cryogenic electron microscopy
CTD	carboxyl-terminal domain
DNA	deoxyribonucleic acid
dNTP	deoxynucleoside triphosphate
DPBB	double-psi-beta-barrel
DSB	double strand break
dsDNA	double-stranded DNA
EC	elongation complex
GF	gel filtration
GO	graphene oxide
GreA-PC	GreA pre-cleavage complex
GreB-PC	GreB pre-cleavage complex
lncRNA	long non-coding RNA
LSU	large subunit
miRNA	micro RNA
mRNA	messenger RNA
NAC	nucleotide addition cycle
ncRNA	non-coding RNA
NMP	nucleoside monophosphate
NTD	amino-terminal domain
ntDNA	non-template DNA strand
NTP	nucleoside triphosphate
PEC	paused elongation complex
ppGpp	guanosine tetraphosphate
P-site	product site
RNA	ribonucleic acid
RNAP	RNA Polymerase
rRNA	ribosomal RNA
SC	secondary channel
SCBF	secondary channel-binding factors
siRNA	short interfering RNA
SPA	single-particle analysis
SSU	small subunit

TCR	transcription-coupled repair
tDNA	template DNA strand
TEC	transcription elongation complex
TEM	transmission electron microscopy
TF	transcription factor
TH	trigger helices
TL	trigger loop
tRNA	transfer RNA

Chapter 1

Introduction

1.1 Gene Expression

The central dogma of molecular biology, described by Francis Crick in 1958 (Crick 1958), centers around the flow of information in biological systems. Information is stored in the genetic sequence of double-stranded deoxyribonucleic acid (DNA) molecules. This can then be copied out into single-stranded ribonucleic acid (RNA), one form of which is capable of being read out to synthesise proteins. In 1970, electron micrographs directly visualized transcribed *E. coli* genes with ribosomes bound to them for the first time (Miller, Hamkalo, and Thomas 1970)(Figure 1.1). In the decades that followed, there have been truly incredible strides made in the understanding of these processes captured in those images. Structures of the particles which initially could only be observed as dark featureless blobs, are now known at resolutions of a few angstroms. And yet, there are still questions that are left to be addressed, each of which is required to fully understand the flow of genetic information. The main question explored in this thesis does this exactly: to add one more piece of the jigsaw in piecing together the overall picture of the core mechanisms that govern gene expression.

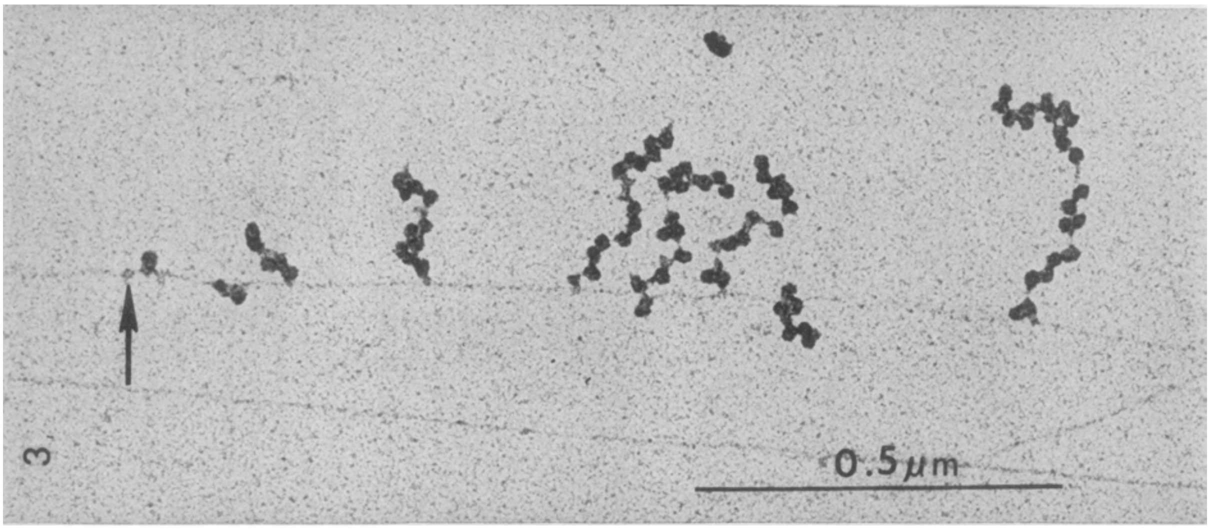


Figure 1.1 – Electron micrograph of an *E. coli* gene being transcribed by RNA polymerases (marked by arrow), with the transcripts appearing as branches along the DNA –the longer transcripts being further away from the transcription start sites. The darker spots are ribosomes bound to the nascent RNA. (Miller, Hamkalo, and Thomas 1970)

At any point of time within an individual living organism, there are a vast number of processes being carried out on the molecular level. Of these, some of the most fundamentally significant are those responsible for the processing of genetic information (Figure 1.2). DNA-dependent RNA Polymerase (RNAP) transcribes DNA to produce single stranded messenger RNA (mRNA) containing the same genetic sequence. mRNA can then be translated by the ribosome to produce the protein whose amino acid sequence was encoded within the DNA strand first transcribed. However, the entire process of DNA → RNA → protein is far more complex than this, with a host of regulatory mechanisms and the influence of external enzyme-binding proteins. Since the discovery of ribosomes and RNAPs in the mid-20th century, the pool of our collective knowledge of what goes on during gene expression has only deepened, culminating in the advent of what has been dubbed the 'resolution revolution' in cryo-electron microscopy (Kühlbrandt 2014).

Before taking a closer look at transcription and translation, it is worthwhile to briefly mention the different forms of RNA that exist within the cell:

- mRNA, which carries the encoded genetic information that is read out by ribosomes for protein synthesis.
- Transfer RNA (tRNA), responsible for transferring amino acids to the ribosome during protein synthesis.
- Ribosomal RNA (rRNA), which assemble along with ribosomal proteins to form functional ribosomes. rRNA are essential functional components of the ribosome.
- Non-coding RNA (ncRNA), which are RNA molecules synthesised from DNA but do not get translated into proteins. These are functionally active and regulate gene expression, and include micro RNA (miRNA), short interfering RNA (siRNA), and long non-coding RNA (lncRNA)

1.1.1 Transcription

The genetic sequence of DNA in an organism encodes for a whole host of proteins required for its growth and survival. In order for the sequence of a particular gene to be

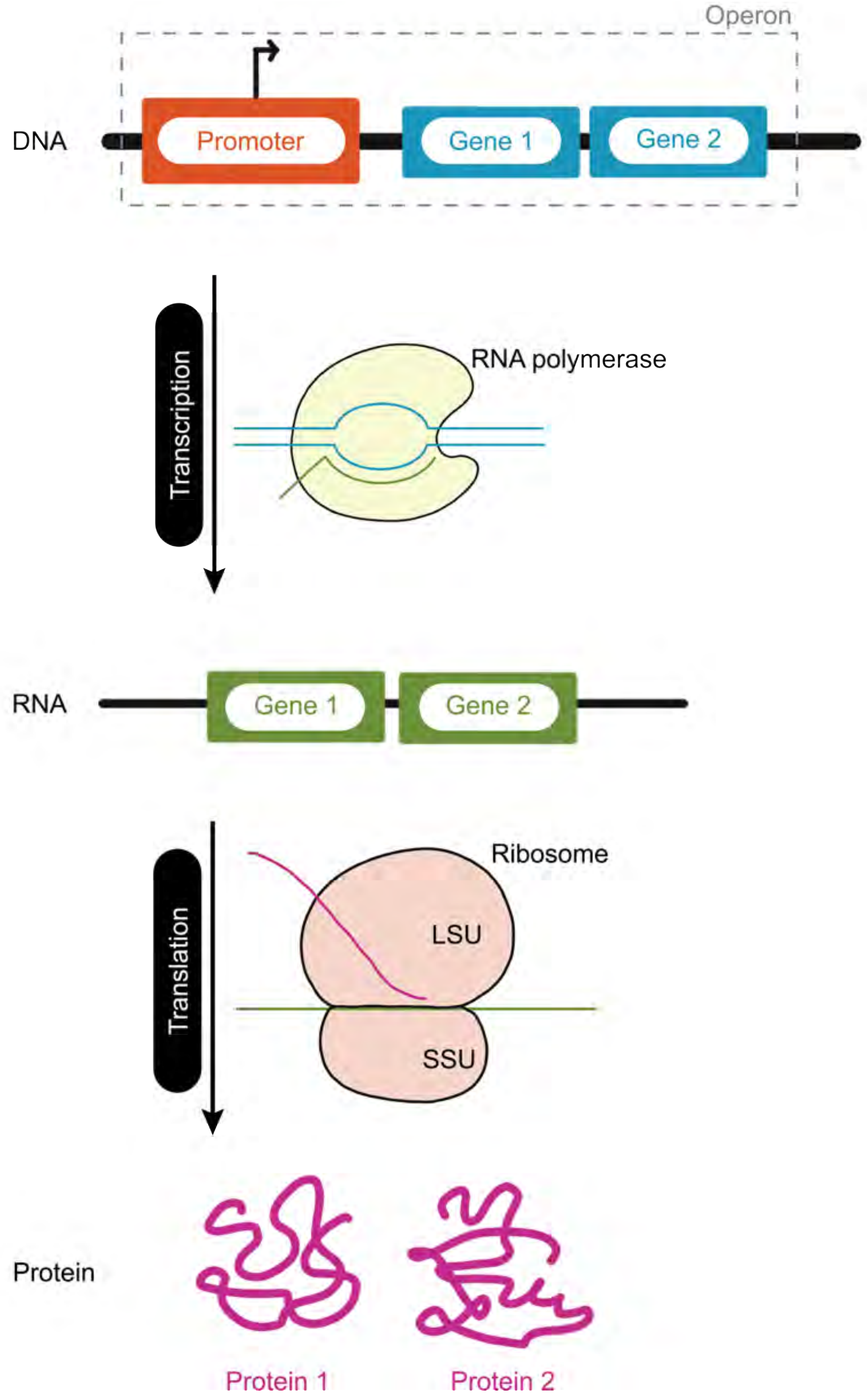


Figure 1.2 – Gene expression in bacteria: starting at a DNA promoter, transcription of one strand takes place to produce an RNA containing the same sequence of the transcribed DNA. This is followed by translation by by the ribosome, in which each gene is read out to form the protein that it encodes for.

read out by the ribosome to synthesise the corresponding polypeptide chain, the same sequence first needs to be copied in the form of single-stranded mRNA by RNAP. DNA-dependent RNAPs, which carry out transcription, exist as multi-subunit enzymes in bacteria, archaea, and eukaryotes, and are single-subunit enzymes in bacteriophages and mitochondria. Bacteria and archaea each have only one form of RNAP that synthesises RNA, while eukaryotes have up to five types of RNAP, each responsible for transcribing specific types of RNA.

Transcription takes place in three phases in a cyclic manner : initiation, elongation, and termination. While Section 1.8 will go into more details regarding the transcription cycle, the overview of the entire process is as follows: RNAP, with help from an initiation factor, locates and binds to a promoter sequence on the DNA. It begins to unwind the double-stranded DNA (dsDNA) and through polymerisation of nucleoside triphosphate (NTP) substrates, forms a DNA-RNA hybrid. After escaping the promoter site, RNAP extends the newly formed, or nascent, RNA one base at a time until it finally reaches the end of the gene and terminates the cycle either through formation of an RNA hairpin or through binding of a termination factor. Each stage of the transcription cycle is controlled by a number of regulatory factors which bind to RNAP itself or to the DNA being transcribed (Cramer 2019). RNAP is responsible for transcribing genes related to the formation of the different types of RNA required by an individual organism. Most relevant to this work is the transcription of mRNA, which encode for proteins.

The first structures of RNAP were the 3.3Å *Thermus aquaticus* core RNAP (Zhang et al. 1999) and the 2.8Å yeast Pol-II (Cramer, Bushnell, and Kornberg 2001), both solved through x-ray crystallography. Advancements in cryogenic electron microscopy (cryo-EM) have led to structures of more short-lived, conformationally diverse states of different RNAP complexes being solved.

1.1.2 Translation

Once an mRNA has been synthesised by RNAP, the next step is for the protein encoded by that genetic sequence to be synthesised. Protein synthesis is carried out by the ribosome, a molecular machine made up of a network of protein and rRNA subunits (Ramakrishnan 2002). Ribosomes in all species are comprised of one large subunit (LSU) and one small subunit (SSU). In *E. coli* the 70S ribosome is comprised of the 50S LSU and the 30S SSU ('S' denoting a non-SI unit known as a Svedberg unit, which relates to the rate of sedimentation in a centrifuge). The LSU consists of a 23S and a 5S rRNA in addition to 33 protein chains, while the SSU is made of a single 16S rRNA along with 21 proteins. Crystallographic structures for both ribosomal subunits were published in 2000: 30S from *Thermus thermophilus* (Wimberly et al. 2000; Schluenzen et al. 2000) and 50S from *Haloarcula marismortui* (Ban et al. 2000). A high resolution 2.4Å structure of the *E. coli* 70S ribosome was published in 2015 (Noeske et al. 2015).

Within the ribosome are three RNA-binding sites –the A (aminoacyl), P (peptidyl) and E (exit) sites. The genetic sequence in an mRNA is broken up into sets of three base and each triplet is known as a codon. Within the cell, the two ribosomal subunits are dissociated from one another. At initiation, a SSU binds to an mRNA at a start codon (AUG) which encodes for a methionine, and recruits a LSU. At the start of the elongation cycle in which single amino acids are added to the growing peptide chain, the P-site contains a peptidyl tRNA, the E-site contains a deacylated tRNA, and the A-site is free for binding of an aminoacyl tRNA. Entry of an aminoacylated tRNA containing an anticodon complementary to the mRNA codon in the A-site takes place along with the release of the tRNA in the E-site. A sequence of reactions elongates the peptide chain by one amino acid, ending with the deacylated tRNA in the P-site moving to the E-site and the peptidyl tRNA in the A-site moving to the P-site, once again freeing up the A-site for binding of the tRNA complementary to the next codon. This cycle continues until the ribosome encounters a stop codon (UAG, UAA, UGA).

1.1.3 Transcription-Translation Coupling

The idea that the processes of transcription and translation could be coupled to any degree was first discussed in the mid-1960s (Byrne et al. 1964). Electron micrographs published in 1970 of genes within *E. coli* chromosomes being expressed showed these two processes working in sync with one another (Figure 1.1) (Miller, Hamkalo, and Thomas 1970). In the decades following this, more and more evidence pointed towards the two processes being associated with one another in bacteria. For some time, transcription-translation coupling was thought to be relevant to specific circumstances, such as transcription attenuation and polarity. Polarity refers to a premature transcrption termination mechanism, which occurs when the translating ribosome slows down, allowing for the Rho termination factor to bind to the mRNA and cause premature termination. Transcription attenuation is an important regulatory mechanism, described extensively in the context of the *trp* operon (Landick, Carey, and Yanofsky 1985). The operon contains an attenuator sequence, which when transcribed can form one of two secondary structures –a terminator, which prevents any downstream genes from being expressed by terminating transcription, and an anti-terminator, which allows for the complex to read through the attenuator sequence and continue transcription. Depending on the availability of amino acids, the pioneering ribosome can promote the formation of one of these two structures.

We now know that transcription-translation coupling in bacteria plays a more significant role in regulation of gene expression. The ribosome translating the mRNA while it is being synthesised can influence transcription through interactions with the RNAP. *In vivo* studies on coupling (Proshkin et al. 2010) showed that the rate of transcription is directly influenced by the rate of translation. They showed that the ribosome decoding the mRNA being transcribed increases the rate of transcription by preventing backtracking. More recently, structural studies have shown direct interactions between the transcription and translation machineries, in a complex named the "expressome". Demo et al. (2017) and Kohler et al. (2017) reported cryo-EM structures showing direct contacts between RNAP and the ribosome in *E. coli.* The former reported a complex between RNAP and the SSU in absence of mRNA. In the latter, a 7.6 Å reconstruction was obtained by colliding the translating ribosome with a stalled RNAP elongation complex (EC). Crosslinking data from Fan et al. (2017) mapped the direct contacts between RNAP and the ribosomal LSU and SSU. They showed direct contacts between RNAP subunits and the 30S ribosome subunit.

1.2 RNA Polymerase

In all organisms, transcription proceeds via a three-stage process. It begins at initiation when RNAP, assisted by initiation factors, locates a promoter on the DNA, binds to it and begins the process of RNA polymerisation. Upon successful initiation, RNAP can then move into the elongation phase, in which the RNA is extended one nucleotide at a time. Finally, at the end of the gene, the process is terminated either intrinsically or with the help of an RNAP-binding protein. This overview, while extremely concise, does not even remotely do justice to the highly complex nature of this extremely fundamental process. In this section, I will describe the structure of RNAP in bacteria as well as its comparisons with eukaryotes and archaea, followed by a detailed look at the various stages of transcription, and how they are regulated.

1.2.1 Evolution

The RNA world hypothesis (Gilbert 1986) describes a primordial landscape made up of single-stranded RNA molecules. The propagation of genetic information in such a scenario would have depended on ribozymes having RNAP-like activity, capable of replicating both itself and other functional RNAs. Ekland and Bartel (1996) demonstrated that a ribozyme generated from random sequences, in the presence of template RNA and NTPs, was capable of extending an RNA primer using the same reactions seen in RNAP enzymes with a remarkable level of fidelity. Recent work confirming the likelihood of this scenario described the *in vitro* evolution of an RNAP ribozyme which was capable of synthesising functional ribozymes (Tjhung et al. 2020).

Over time, these initial RNAs began producing proteins, which eventually led to the differentiation of proteins for enzymatic functions and RNA for storage of genetic information (Eigen 1971; Joyce 2002). It has been hypothesised that all existing DNA-dependant RNAP in all three kingdoms of life co-evolved from the same multi-subunit enzyme, or the last universal common ancestor (LUCA). We see clear evidence for this in the large degree to which both the structural motifs of RNAP and the mechanisms of

	$\boldsymbol{\alpha 1, \alpha 2}$	$\boldsymbol{\beta}$	$\boldsymbol{\beta'}$	$\boldsymbol{\omega}$
E. coli gene	rpoA	rpoB	rpoC	rpoZ
No. of residues	329	1342	1407	91
Molecular weight (kDa)	36.5	150	155	10.2

Table 1.1 – *E. coli* RNAP subunits

transcription are conserved.

Structural comparisons of bacterial RNAP and eukaryotic Pol-II showed very clear similarities between the two – not just in domain architecture, but also in the folding of individual proteins (Ebright 2000) . This not only highlighted the evolutionary ties these two polymerases share but also illustrates the very real value of studying bacterial transcription. Studies on bacterial RNAP and its processes do have very real implications in the understanding of the same processes within Pol-II. Comprehensive structural and sequence analyses of RNAP across various species (bacteria, plant plastids, archaea, eukaryotes) by Lane and Darst (2010a, 2010b) shone more light on the similarities between the different forms of this multi-subunit enzyme across the board. All cellular life forms are thought to have been descended from a common RNA-based life form. The survival of such a life form would have depended on an RNA polymerase ribozyme, capable of transcribing both functional RNA molecules as well as itself in a manner with a high level of fidelity.

1.2.2 Structure of Bacterial RNAP

With a molecular weight of roughly 400kDa, RNAP in bacteria are comprised of five subunits - two α, β and β', and ω. Compared with archaea and eukaryotes, the multi-subunit RNAP in bacteria is relatively much simpler. Table 1.1 lists the five subunits within *E. coli*, with Figure 1.4(f) showing their assembly within the enzyme.

Alpha

The two alpha subunits ($\alpha 1, \alpha 2$) exist as a homodimer (Figure 1.4 (a) & (b)) (Murakami et al. 1997). Structurally, each subunit is divided into an N-terminal and a C-terminal domain (residues 1-235 and 248-329, respectively), connected by a flexible linker. The

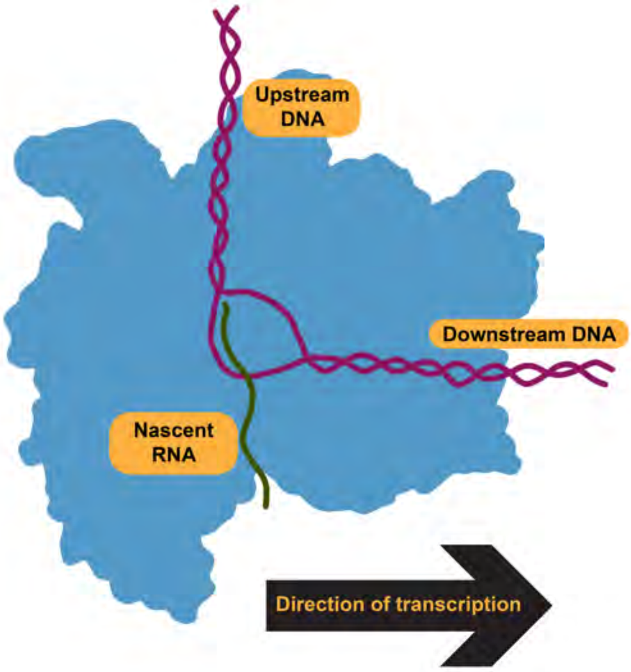


Figure 1.3 – Orientation of the transcribing RNAP with respect to the DNA being transcribed.

NTD can be further divided into two domains - domain 1 and domain 2 - the first being comprised of 2 alpha helices and a four-stranded anti-parallel beta sheet, and the second made of 7 anti-parallel beta sheets and 1 alpha helix. Only αNTD Domain 1 is involved in the dimerization process, with the 2 pairs of alpha helices interlocking to form a hydrophobic core. The formation of the α subunit dimer is the first step in RNAP assembly, with it acting as the scaffold for assembly of the β and β' subunits. $\alpha1$ forms contacts with only the β subunit, while $\alpha2$ contacts both β and β'.

In addition to its role in dimerisation and RNAP assembly, the αNTD also influences gene regulation through its interactions with transcription regulators (Liu et al. 1996). While the αCTD is non-essential for RNAP assembly, its importance lies in its role during transcription activation through interactions with various transcription factors and DNA.

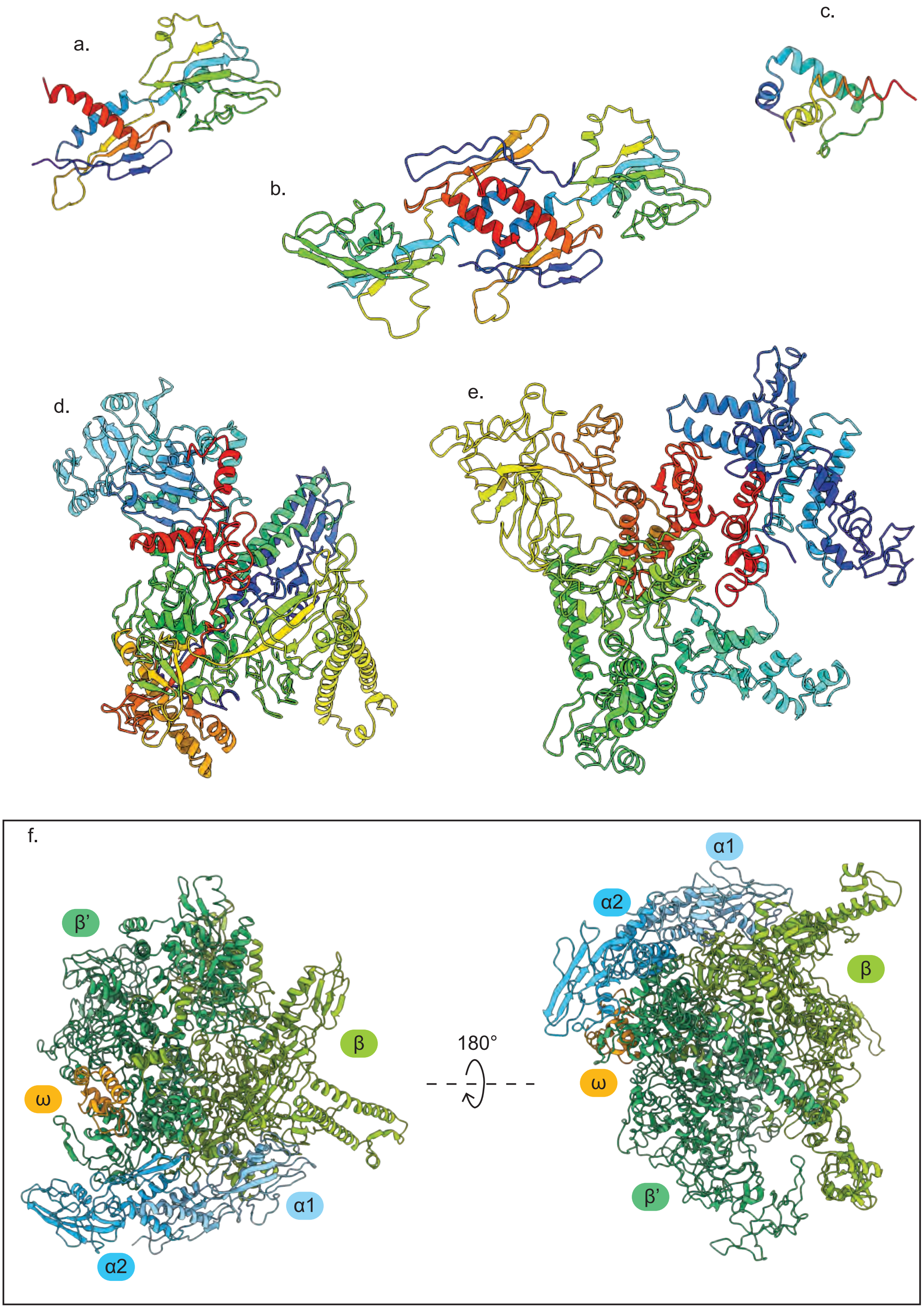


Figure 1.4 – Subunit composition of *E. coli* RNAP, showing (a) a single α subunit and (b) the α subunit dimer, (c) ω, (d) β, and (e) β'

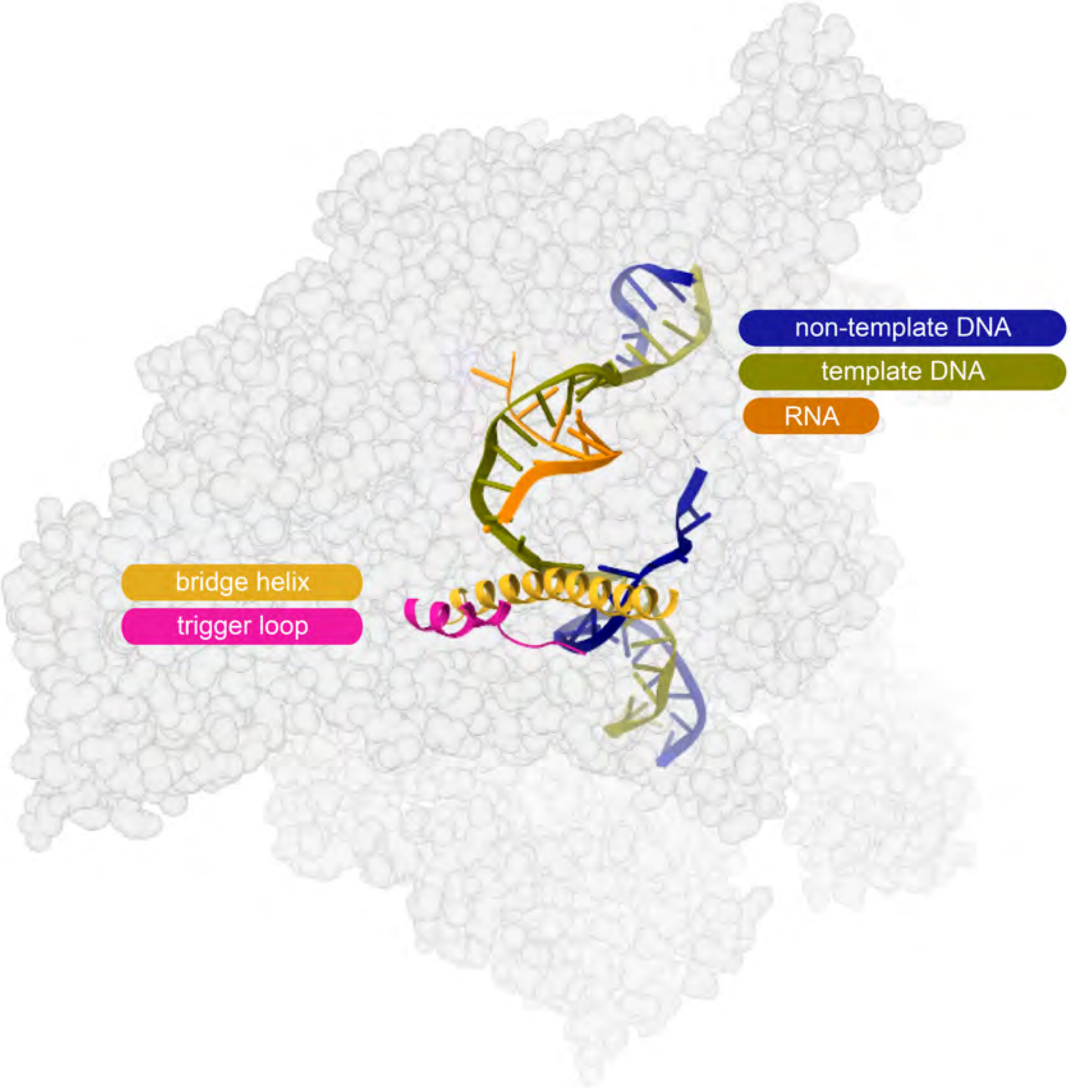


Figure 1.5 – Positions of the bridge helix and trigger loop within the RNAP active site, with respect to the RNA transcript, template and non-template DNA strands.

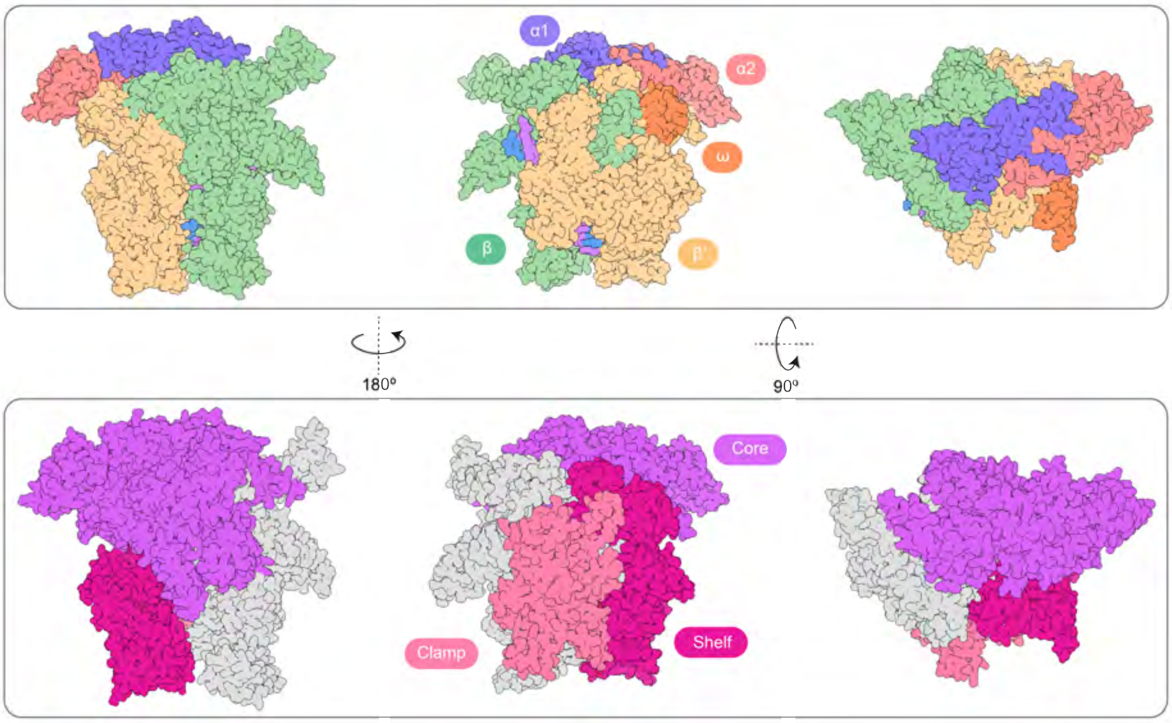


Figure 1.6 – *E. coli* RNAP subunits and relevant domains

Beta, Beta Prime

β and β' are the two largest subunits that make up the RNAP core enzyme (Sutherland and Murakami 2018) (Figure 1.4 (d) & (e)). While they do have comparatively similar sizes in *E. coli*, this is not the case across the board; for instance, the equivalent subunits in the yeast species *Saccharomyces cerevisiae*, Rpb1 and Rpb2, have molecular weights of approximately 191 and 138 kDa respectively.

Encoded by the rpoB and rpoC genes in the β operon (in which rpoB is positioned upstream of rpoC), these two subunits contact each other over a surface area of 7734Å. Their assembly in the RNAP core complex depends on the formation of the α dimer. An interesting point to note about the positions of the two large subunits relative to one another is that the C-terminus of the β subunit lies adjacent to the N-terminus of the β' subunit. This shows that it might well have been the case that these two subunits existed as a single entity in ancient RNAPs. Evidence for this is indicated by the work done by Severinov et al. (1997), in which fusion of the β and β' subunits still resulted in a functional enzyme.

RNAP is known for its characteristic crab claw-like appearance, with the β and β' subunits forming the pincers of the claw. The cleft formed by the claw is what forms the main channel through which the template DNA enters the enzyme core. While the main channel is positively charged to allow for binding of the downstream DNA, the rest of the RNAP surface is predominantly negatively charged to prevent any non-specific DNA interactions. The DNA-binding clamp, which is open when RNAP is not bound to DNA as well as in the initial intermediate states of the RNAP-promoter complex formation, is closed once RNAP forms a stable complex and remains closed during elongation (Chakraborty et al. 2012). Figure 1.6 shows the three major structural motifs in *E. coli* RNAP, along with the subunits that they are a part of. The two large subunits, especially β', make up most of the core, clamp and shelf. The clamp and the shelf, which together form what is called the swivel module, are capable of rotating relative to the core. Rotation of the swivel module, which allows the tips of the RNAP pincers to be positioned closer to one another, has been observed in hairpin-stabilised and in backtrack-stabilised paused ECs (Guo et al. 2018; Abdelkareem et al. 2019).

The β and β' subunits contain most of the structural elements that are conserved across all cellular RNAPs. The first is the active centre or the catalytic core of RNAP, which is contained within two double-psi-beta-barrel (DPBB) domains - one from each of the two largest subunits. The DPBB domain in β' contains a conserved aspartate triad which is responsible for the coordination of the Mg^{2+} ions required for catalysis. The β DPBB domain, on the other hand, contains the basic residues that interact with the NTP during catalysis. The role of these DPBB domains in RNA synthesis is a characteristic feature of all cellular RNAPs. Another important structural motif within the catalytic core is the bridge helix (BH), found in the β' subunit. Structurally, it forms a barrier between the DNA-binding main channel and the secondary channel, which serves as an entry pathway for incoming substrates as well as a binding site for many regulatory transcription factors. The BH is also known to play a key role in the translocation of the DNA/RNA hybrid during transcription elongation (Bar-Nahum et al. 2005).The importance of the BH in the context of the nucleotide addition cycle (NAC) is highlighted in the elemental paused elongation complex (PEC) structures from (Weixlbaumer et al. 2013), in which they showed that kinking of the RNAP BH when it enters a paused state blocks the active site, thereby preventing the formation of new RNA bonds from being catalysed. In addition to the BH, β' also contains the trigger loop (TL), which is capable of participating in catalysis. The positions of the BH and the TL relative to one another and to the RNA and DNA within the complex are shown in Figure 1.5. The functional role of the TL during transcription is explored further in Section 1.3.4.

Other key structural elements which interact with DNA and RNA are the Fork loop 2 and the Switch 3 modules in β, together with the lid and the rudder in β'. Fork loop 2 assists in the separation of the DNA strand downstream of the active site. At the cleft of the main channel, the downstream DNA exists in its duplex form. An arginine from the fork loop 2 stacks on top of the downstream DNA base pair, thus blocking the entry of the DNA duplex into the active site. The Switch 3 module interacts with the separated RNA single strand upstream of the active site and transcription bubble, helping in the separation of the DNA/RNA duplex and promoting reannealing of the two DNA strands. Both the lid and the rudder also interact with the nucleic acids upstream. The lid stacks on the upstream DNA/RNA hybrid base pair, consequently blocking any further growth of the hybrid (Toulokhonov and Landick 2006). The rudder prevents re-association of the separated DNA and RNA strands through interactions with the separated template DNA strand.

Finally, the β flap-domain lines one side of the exit channel for the nascent RNA. This module is implicated in transcriptional pausing and termination: on the formation of an RNA hairpin in the narrow exit channel, the conformation of the flap domain is affected, leading to pausing and termination.

Omega

With a molecular weight of 10.2 kDa, ω is the smallest subunit in RNAP (Figure 1.4 (c)). Structurally, it consists of 5 alpha helices and binds mainly to the β' subunit. In particular, it binds to the DPBB domain, which contains the catalytic active site for

RNA synthesis. By binding to the DPBB domain of RNAP, ω possibly plays a role in maintaining RNAP activity by shielding the DPBB domain

The exact role of the ω subunit hasn't been studied as extensively as the other subunits that make up RNAP, a consequence of early results which showed that it is non-essential for cell growth and for reconstitution of a transcriptionally active enzyme in vitro(Gentry et al. 1991). In the same decade, it was also shown that a ΔrpoZ strain of RNAP co-purified with the protein chaperone GroEL, pointing towards a possible role of the ω subunit in RNAP folding (Mukherjee et al. 1999). This also suggested a role of ω in maintaining the structural stability of RNAP.

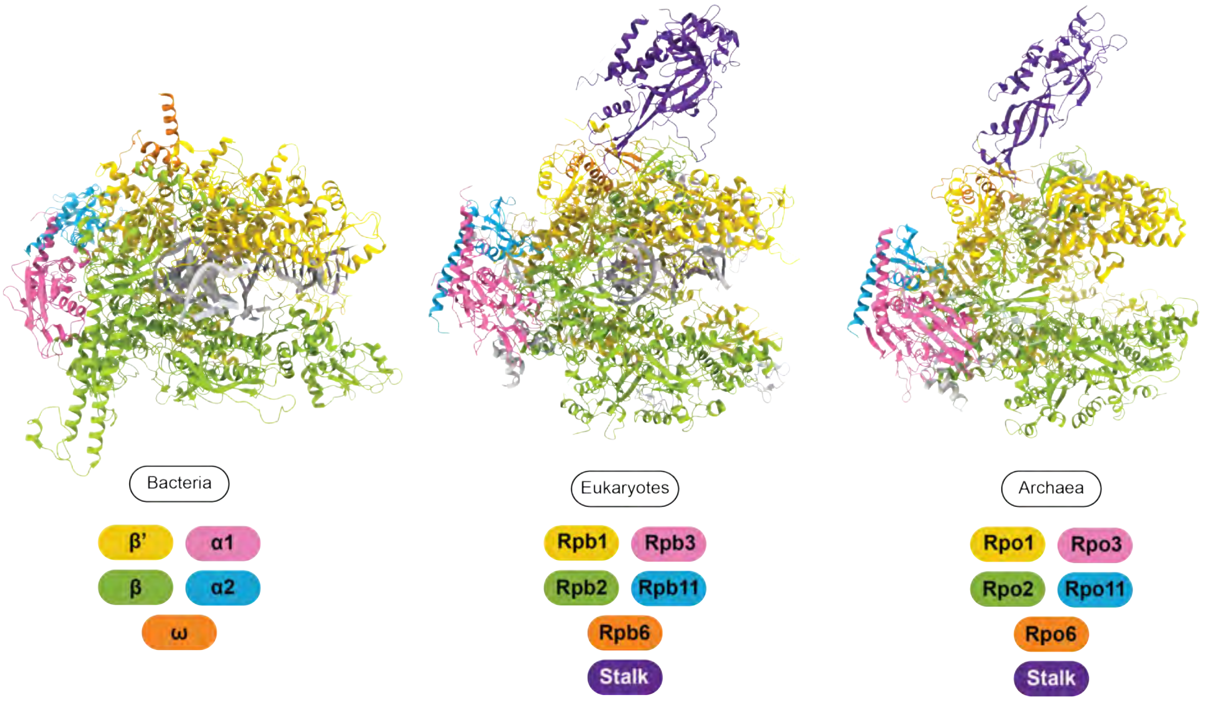


Figure 1.7 – Comparison of RNAP structures of species from all three kingdoms of life, with equivalent subunits coloured similarly. The structures are all oriented with the downstream DNA entering from the right hand side of the page, and the upstream DNA directed outwards from the face of the page. The bacterial structure is from *E. coli* (6ALH) (Kang et al. 2017), Pol-II from *Saccharomyces cerevisiae* (1Y1W) (Kettenberger, Armache, and Cramer 2004), and archaeal RNAP from *Saccharolobus shibatae* (2WAQ) (Korkhin et al. 2009). The archaeal and eukaryotic strutures show the stalk (coloured in purple), which isn't present in bacteria

1.2.3 RNAP in Archaea and Eukaryotes

While Archaea each encode for a single form of RNAP similar to bacteria, eukaryotes contain multiple RNAPs – Pol-I, Pol-II, and Pol-III. In addition to these three, plants also encode two additional RNAPs, Pol-IV and Pol-V, which evolved from Pol-II and transcribe different ncRNAs (Zhou and Law 2015) .The five subunits that make up bacterial RNAPs have homologues in archaea and eukaryotes, but the enzymes in these two kingdoms of life also have additional subunits. The subunit composition in all three domains are listed in Table 1.2. In addition to these conserved subunits, there are also additional RNAP-binding transcription factors which are conserved, some of which are also homologous to subunits in other RNAPs. For example, the Pol-II transcription factor (TF) TFIIS, which is functionally homologous to the bacterial factor GreA, is structurally similar to the subunits A12, Rpb9, and C11 in Pol-I, II and II respectively, and also to the TF TFS in archaea.

Bacteria	***Archaea***	***Eukaryotes***		
		Pol-I	*Pol-II*	*Pol-III*
β'	Rpo1	A190	Rpb1	C160
β	Rpo2	A135	Rpb2	C128
$\alpha 1$	Rpo3	AC40	Rpb3	AC40
$\alpha 2$	Rpo11	AC19	Rpb11	AC19
		A12	Rpb9	C11
	Rpo5	Rpb5	Rpb5	Rpb5
ω	Rpo6	Rpb6	Rpb6	Rpb6
	Rpo8	Rpb8	Rpb8	Rpb8
	Rpo10	Rpb10	Rpb10	Rpb10
	Rpo12	Rpb12	Rpb12	Rpb12
	Rpo4	A14	Rpb4	C17
	Rpo7	A43	Rpb7	C25
	Rpo13			

Table 1.2 – RNAP subunit composition in bacteria, eukaryotes (Pol-I,II,III), and archaea. The bacterial subunits and their homologous subunits in eukaryotes and archaea are highlighted in yellow.

The evolutionary history and similarities within RNAPs from in the three domains has been studied extensively, and was recently discussed by Werner and Grohmann (2011). The primary structural feature conserved throughout is the DPBB motif, which is a part of the two largest subunits: β and β' in bacteria, Rpo1 and Rpo2 in archaea, Rpb1 and Rpb2 in Pol-II. The general structural similarities and differences amongst different RNAPs are shown in Figure 1.7. Apart from the conserved DPBB motif, which contains the active site of RNA polymerisation, other structural similarities are difficult to point out, owing to the large insertions and deletions in the enzyme across all species. Many of the additional subunits in eukaryotes and archaea might not be directly necessary for RNA polymerisation, but are needed for cell viability and serve as contact points between RNAP and TFs, between individual RNAPsubunits, and with nucleic acids.

The main structural difference that eukaryotes and archaea have when compared with bacteria is the stalk. It is formed by the subunits Rpb4 and Rpb7 in Pol-II in eukaryotes, and by Rpo4 and Rpo7 in archaea. While it is stably bound within arachaeal RNAP, it is reversibly bound to Pol-II (Orlicky et al. 2001). The Pol-II stalk, as well as the existance of Pol-II-specific TFs direct us towards an understanding of the differences in transcription regulation amongst the various eukaryotic RNAPs. In their review on transcription in different eukaryotic RNAPs, Barba-Aliaga, Alepuz, and Pérez-Ortín (2021) proposed that the difference seen between these processes, even within the same species, boils down to the transcription products generated by each RNAP. Since Pol-II transcribes mRNA and must interact with an extremely large set of genes, it is therefore the target of a large set of TFs. On the other hand, Pol-I and Pol-III are responsible for transcribing untranslated genes. Pol-I, which transcribes rRNA, needs to carry out transcription at a high level of fidelity, and at a faster rate (Goodfellow and Zomerdijk 2013). To assist in this, rather than having an external cleavage factor like Pol-II, the TFIIS-like subunit in Pol-I, A12, is responsible for promoting cleavage activity and ensuring fidelity. Pol-III also contains a subunit similar to this, C11, which promotes transcription fidelity.

While the comparisons that can be made between different RNAPs is extensive, the one that is most relevant to this work is in the subunits and factors responsible for ensuring transcription fidelity. While TFS, TFIIS and the eukaryotic homologous subunits are structurally and functionally similar, they exhibit the same functional role as the Gre factors in bacteria, described in later sections. Despite the structural dissimilarities, fully understanding the mechanisms by which the Gre factors ensure bacterial transcription fidelity might point towards a deeper understanding of transcription fidelity in eukaryotes and archaea.

1.2.4 The Transcription Cycle

Transcription takes place in three steps: initiation, elongation, and termination. Once initiation has been completed successfully and RNAP begins elongation, the entire complex remains intact without dissociating until it terminates the process at the end of the gene being transcribed. Broadly, the mechanisms and domains involved are conserved. Hence, all of the structures and mechanisms discussed will be mostly in the context of bacterial RNAP. When relevant, the equivalent names and residues in eukaryotes and archaea will be mentioned.

Initiation

For transcription to begin, the RNAP core enzyme must recognise and bind to a promoter site. On its own, the core enzyme cannot initiate transcription and requires the assistance of additional factors. These initiation factors in bacteria are known as the sigma (σ) factors, and together with the RNAP core complex form the RNAP holoenzyme.

The σ factors work by recognising specific sequence motifs and guide the RNAP core to the promoter site. There are multiple σ factors, each of which is required under different physiological circumstances. The number of σ factors encoded within the genome of each species of bacteria varies to a fairly large extent from a single factor in *Mycoplasma genitalium* to 7 in *E. coli* and 65 in *Streptomyces coelicolor* (Gruber and Gross 2003). The primary housekeeping σ factor in *E. coli* is σ^{70} which is the dominant σ factor in cells during exponential growth. The other known factors in bacteria are σ^{38}(chemotaxis and flagella formation), σ^{32} (heat shock response), σ^{24} (heat shock response), σ^{54} (nitrogen fixation), σ^{28} (flagellar gene expression) and σ^{19} (ferric nitrate transport) (Tripathi, Zhang, and Lin 2014). The numbers used in the naming of each σ correspond to its approximate molecular weight in kDa.

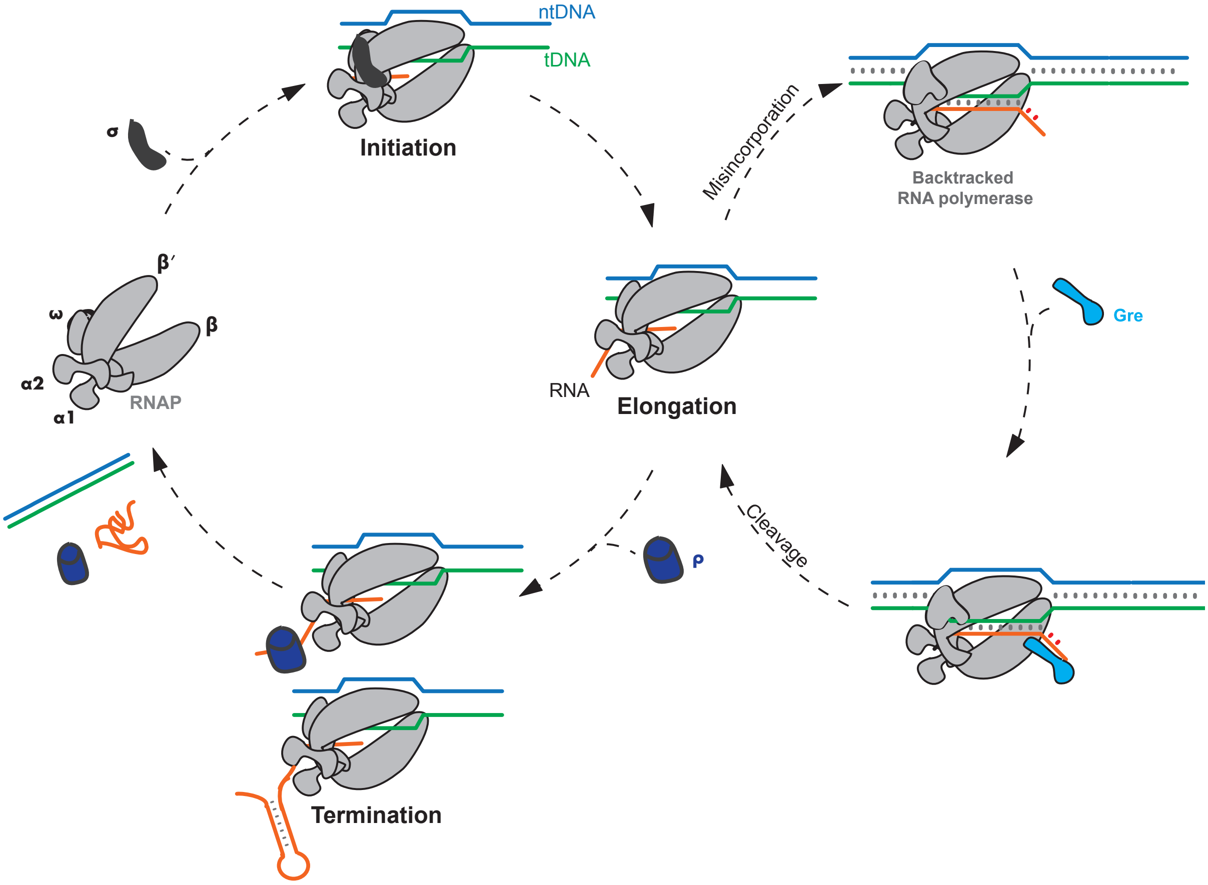


Figure 1.8 – Schematic of the transcription cycle in *E. coli*, showing the three main phases - initiation, elongation, and termination

At the very beginning, RNAP with a σ subunit bound recognises and binds to a specific promoter, thus forming the initial closed RNAP-promoter complex. What follows this is a series of changes that brings the complex into an open RNAP-promoter complex at the promoter. The intermediate steps involved in formation of the open initiation complex are described in (Ruff, Thomas Record, and Artsimovitch 2015). The intermediate conformations between the closed and open complexes correspond to bending of the upstream DNA, bending of the downstream DNA into the RNAP cleft and opening of the DNA double helix, and stabilising the open complex. Looking at the kinetics involved in the transition between each of these intermediate states, we see that the main rate-limiting step is the point at which DNA starts to be unwound in the cleft in order to allow for a single strand to be guided into the acceptor site (A-site), a part of the active site that binds an incoming rNTP substrate.

Once in the open complex conformation, the DNA strand downstream is melted past the start site, thus forming the transcription bubble. NTP substrates entering through the secondary channel (SC) are polymerised one by one to form an initial RNA transcript that is about 12 nucleotides long. Situated close to the A-site is a portion of the σ_3 domain of the σ factor called the $\sigma_{3.2}$ loop, proposed to play a significant role in abortive initiation (Murakami, Masuda, and Darst 2002). During abortive initiation, the transcript can either "push past" the $\sigma_{3.2}$ loop and extend up to the required length, or is unable to do so and is likely released through the SC. Abortive initiation ends once the RNA reaches a length of about 12 nucleotides, which at that point is long enough to form the complete DNA-RNA hybrid and enter the RNA exit channel, displacing the $\sigma_{3.2}$ loop.

In the final stage of initiation, the interactions of the σ factor with both RNAP and the promoter are destabilised, allowing RNAP to transition into the elongation phase by escaping the promoter and starting the process of forward translocation along the DNA.

Elongation

Once RNAP has escaped from the promoter site, it begins the process of extending the nascent RNA at the 3'-end one nucleotide at a time, while simultaneously translocating forward along the DNA.

Extension of the nascent RNA happens through the NAC. During the NAC, the catalytic core of RNAP undergoes a series of highly dynamic conformational changes. Two binding positions within the active site are relevant here: the A-site, and the product site (P-site) which lies adjacent to it. At the start of the cycle, the 3'-hydroxyl of the RNA is positioned at the P-site, with the A-site available for substrate binding. NTPs enter the RNAP core through the secondary channel. An NTP complementary to the DNA base in the A-site needs to be positioned correctly, for which two important screening mechanisms need to come into play: differentiation between deoxyribonucleotides (dNTP) and ribonucleotides (NTP), as well as detection of the correct ribonucleotide. The RNAP TL plays a key role in substrate selection, which is described later in Section 1.3.4. The positioning of a complementary NTP in the A-site triggers the process of formation of a new phosphodiester bond. This takes place via a two metal ion-mediated nucleophilic substitution (S_N2) reaction, the schematic for which is shown in Figure 1.11. In this reaction, the terminal 3' oxygen of the nascent RNA, positioned at the P-site, functions as the nucleophile. It attacks the α-phosphate of the NTP, resulting in the formation of a covalent bond. The bond between the α- and β-phosphates is broken, and the pyrophosphate is released through the SC. The elements within the active centre, which were specifically positioned to allow for phosphodiester bond formation, are de-stabilised by the release of the pyrophospate. This allows for the EC to be driven forward along the DNA by one base pair. The active centre element mainly responsible for translocation is the BH, which does so through mechanism which, when simplified, resembles a ratchet and pawl device (Bar-Nahum et al. 2005). Forward translocation of the transcription elongation complex (TEC) simultaneously results in the shifting of the newly added RNA base into the P-site, unwinding of one DNA base downstream of the transcription bubble, and re-annealing of one DNA base pair upstream of the transcription bubble.

Termination

Termination of transcription is essential and plays a vital role in controlling the efficiency of gene expression. Transcription termination at the end of each operon allows for RNAP to be recycled and avoids unnecessary expression of downstream genes.

Two forms of termination may occur at the end of the transcription cycle: intrinsic termination or Rho-dependent termination (Figure 1.8). Intrinsic termination relies on formation of secondary structures within the transcript in order to dissociate the DNA, RNA, and RNAP. In Rho-dependent termination, dissociation of the EC relies on the RNA translocase activity of the Rho factor. Regardless of the pathway used, the main goal within termination is to destabilise the EC enough to allow for RNAP to dissociate from the DNA template and the newly-synthesised RNA. Pausing is a key mechanism which promotes termination. By stalling the EC for long enough, it allows for more time for the complex to enter a termination pathway faster than NTP addition to take place

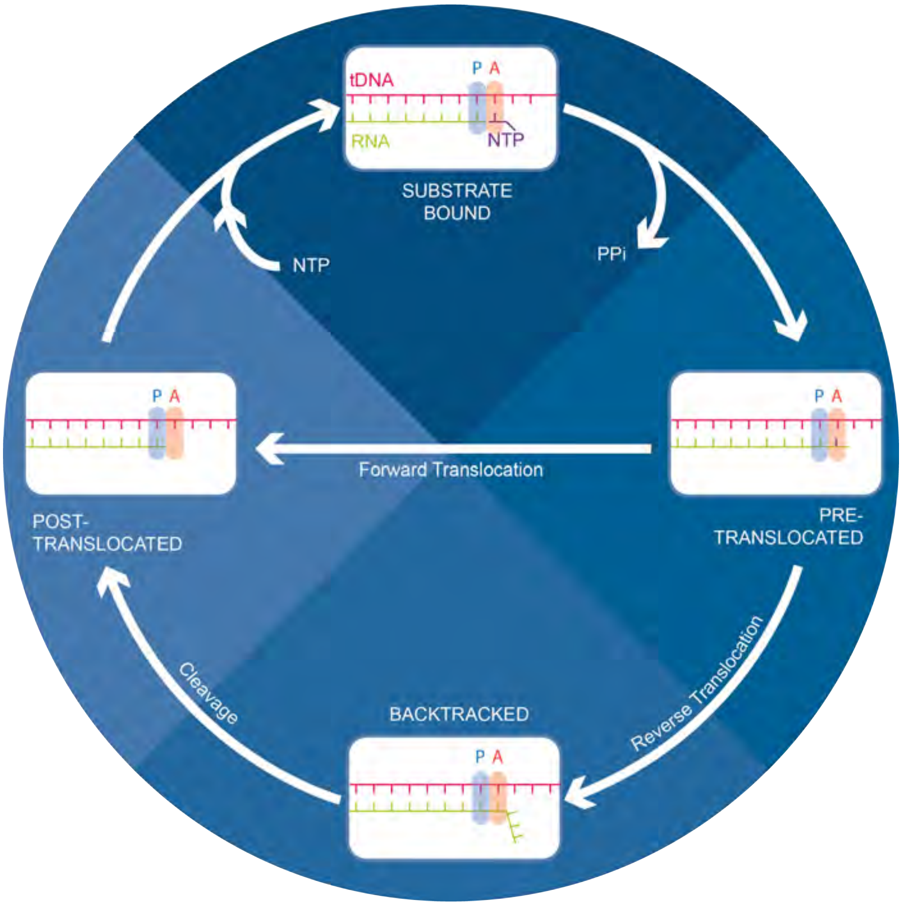


Figure 1.9 – Nucleotide addition cycle during transcription elongation, showing the A- and P-sites

and read through the termination site by shifting the transcription bubble forward.

Termination may also take place via a third mechanism, mediated by the transcription-coupled repair (TCR) factor Mfd (Selby and Sancar 1993). Mfd recognises ECs stalled due to DNA damage and binds to both the stalled RNAP and the DNA, removing the stalled RNAP. Recent structures from the Darst lab (Kang et al. 2021) describe the mechanism by which Mfd recognises stalled ECs and dissociates them to allow TCR to take place. The transcription-coupled repair mechanisms that follow ultimately then dissociate the EC, thus terminating transcription.

1.2.5 Regulatory Mechanisms

From initiation up until termination, transcription is a highly regulated process. Each step is controlled by a number of regulatory mechanisms carried out by the core enzyme and by external RNAP and nucleic acid-binding protein factors and ncRNAs. The various aspects of transcription regulation are well-documented and extensive, and this section will only illustrate a few examples of regulation at each step of the transcription cycle, with the particular focus on the role of different SC-binding factors and their effects on transcription elongation.

Regulation of Initiation

Different σ factors, responsible for recognising different promoter sequences, affect the transcription of genes in response to different environmental factors. Regulation of transcription of various operons by the RNAP holoenzyme containing one of the many σ factors is carried out by binding of different repressors and activators to the DNA upstream of the RNAP binding site. Regulation may also occur through modifications to the promoter DNA which affect the binding of RNAP and regulatory proteins to the promoter region. In the system of regulation of the expression of certain genes, we often see a two-component system that comes into play: one protein acts as a 'sensor' to a particular environmental change, while the second 'receiver' protein acts towards responding to the change. An example of such a system is the well-documented PhoB-PhoR regulatory system, which controls the response in bacteria to a reduction in the concentration of free phosphate in its environment (Santos-Beneit 2015). In this system PhoR, a transmembrane protein, functions as the sensor while PhoB, a cytosolic protein, functions as the receiver. A drop in the phosphate concentration in the outer environment causes phosphate to diffuse out of the cell, in turn causing a change within the cytoplasmic domain of PhoR. This results in the transfer of an ATP γ-phosphate to a specific side chain in PhoB. This switches this trancription activator to an active state, and the phosphorylated PhoB induces transcription initiation at the promoter sites of genes to help the cell survive.

In their review on regulation of transcription initiation, Browning and Busby (2016) discuss the evolution of the various regulatory processes we see during transcription initiation. Early transcription was likely completely unregulated, with relics of it possibly being the "pervasive transcription" seen in bacteria, in which some transcripts formed are simply noise and unrelated to specific genes. The evolution of the σ factors which allowed bacterial RNAP to bind to specific start sites would have been followed by the evolution of various repressors and activators.

Regulation During Elongation

Once in the elongation phase, RNAP moves along the DNA template one base pair at a time, extending the RNA as it does so. This isn't a smooth process, and is frequently stalled in paused states. During pausing, RNAP translocation stalls at a point on the DNA and halts RNA synthesis without the TEC dissociating. Pausing occurs more frequently and across more organisms than what was initially expected, and takes place at an average rate of 20-100 bp (Kang et al. 2019). Pausing is often accompanied by reverse translocation of RNAP along the DNA in what has been termed as backtracking, first described by Nudler et al. (1997) and Komissarova and Kashlev (1997). Backtracking does not always take place at regulatory pause sites, as has been described by Kireeva and Kashlev (2009). Backtracking and its significance in transcription will be explored in more detail in Section 1.3.1.

Initially thought to only take place at promoter and terminator regions through the formation of RNA secondary structure hairpins, the development and implementation of different techniques to map pausing *in vivo* and genome wide have revealed a much broader landscape of pause sites (Larson et al. 2014; Vvedenskaya et al. 2014). Pausing is a vital regulatory process in transcription. It assists in the coordination of transcription and translation (see section 1.1.3), allows for formation of RNA secondary structures and binding of elongation factors to the EC, and is required for transcription termination. Although transcription initiation was thought to be the rate-limiting step of the entire cycle, promoter-proximal pausing has been shown to be an important factor influencing the rate of transcription. Pausing can either be stabilised through the formation of RNA hairpins or by backtracking, which forms the basis of their classification into Class I and Class II pauses (Artsimovitch and Landick 2000). In both cases, elongation is prevented through disruption of the elements within the catalytic centre (Class I) or by blocking the A-site by nascent RNA (Class II). Through x-ray crystallographic studies and, later, single-particle cryo-EM, structures of different paused elongation complexes have provided insights into the various steps involved in different types of pausing. The first crystal structures of a PEC was published by Wang et al. (2009), which was of a backtrack-stabilised yeast Pol-II PEC. More structures were solved by x-ray crystallography soon after, with the *T. thermophilus* elemental PECs (ePEC) structures from (Weixlbaumer et al. 2013) and the *T. thermophilus* backtracked PEC from Sekine et al. (2015). 2018 saw the publication of a number of PECs solved through cryo-EM, which included elemental, hairpin-stabilised, and TF-bound PECs (Kang, Mishanina, Bellecourt, et al. 2018; Kang, Mooney, et al. 2018; Guo et al. 2018). The transcription factor NusA stabilises Class I pauses through interactions with the paused EC (Guo et al. 2018). NusG, a universally conserved transcription factor (Spt5 in eukaryotes and archaea), interacts with RNAP and has been shown to promote elongation (J. Li et al. 1992; Burova et al. 1995). Both NusA and NusG interact with RNAP at positions which also serve as binding sites for other transcription factors. NusG contacts the RNAP β' clamp helices, which is also where a non-essential protein RfaH binds. Both NusG and RfaH promote transcription elongation, but differ in their effects on ρ-dependent termination –NusG promotes while RfaH lowers it, depending on the termination site (Belogurov et al. 2009). NusA and the σ^{70} initiation fcator bind the same region on the RNAP core, but are required for different stages of the transcription cycle.

Regulation of Termination

Both ρ-dependent and intrinsic transcription termination are mediated by external factors which either promote or inhibit termination. Two transcription factors which play a significant role in termination and antitermination are NusG and NusA. NusA can both enhance and inhibit termination, depending on the terminator in question. For instance, it enhances termination at the λtR2 and *E. coli* rrnB T1 terminators (Schmidt and Chamberlin 1987). On the flipside, it reduces the efficiency of termination at other

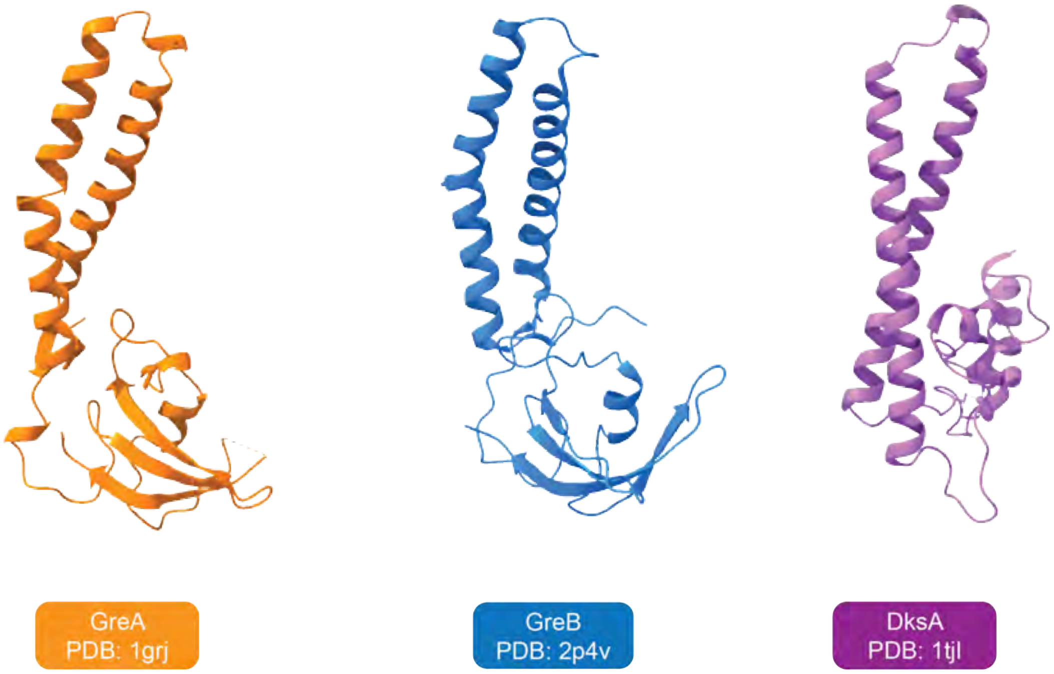


Figure 1.10 – Structures of the secondary channel-binding factors (SCBF)'s from *E. coli* - GreA, GreB, and DksA

sites such as the intrinsic terminator preceding the rpoB gene (which encodes for the RNAP β subunit) (Linn and Greenblatt 1992). Read-through of this terminator is also increased in the presence of NusG. NusA further promotes anti-termination by stabilising the interactions of the antiterminator λN protein with RNAP (Mason, Li, and Greenblatt 1992). Antitermination by λN is also assisted by interactions with NusG (J. Li et al. 1992).

Secondary Channel Binding Factors

A family of transcription factors relevant to this work is the group of SCBF. The bacterial Gre factors belong to the group of SCBFs along with *E. coli* DksA and Rnk, and *Thermus thermophilus* Gfh1. Bacterial SCBFs are structurally similar to one another, as shown in Figure 1.10: they possess a coiled coil amino-terminal domain (NTD) and a globular carboxyl-terminal domain (CTD). The NTD is the portion of the protein which actually enters the SC, allowing it to contact the catalytic core, while the CTD likely stabilises it by binding to the RNAP surface.

Since all the SCBFs within a particular species bind to the same narrow channel, it is all but natural to expect that there will be some degree of competition for binding amongst the different factors. However, each of these proteins binds to RNAP at specific points in the transcription cycle, potentially reducing any adverse effects on the functionality of one by competitive binding of another protein. For instance, DksA (DnaK suppressor A) primarily mediates transcription initiation at rRNA promoters, while GreA and GreB aid in the cleavage of RNA in backtracked complexes during transcription elongation. DksA was shown to be an essential transcription regulatory factor for initiation at rRNA promoters (Paul et al. 2004), alongside guanosine tetraphosphate (ppGpp) and the initiating NTP.

Similar to GreA and GreB, DksA also contains two highly conserved acidic residues –two aspartates in this case –capable of coordinating a Mg^{2+} ion bound to ppGpp, which would in turn stabilise the ppGpp-RNAP complex (Perederina et al. 2004). The ppGpp response network is essential for the survival of bacterial cells under conditions of amino acid deficiency. When bound to RNAP, ppGpp inhibits rRNA and tRNA transcription, and promotes transcription of genes encoding for proteins involved in amino acid synthesis. The effects of DksA on transcription might not be limited to rgulation of initiation. Perederina et al. (2004) compared the effects of DksA and GreA on elongation *in vitro.* While DksA has no effect on the ability of a TEC to read through a pause site (pausing was eliminated in the presence of GreA), it prevented transcription arrest in the presence of an excess of NTPs (as did GreA) but did not induce faster cleavage in an arrested complex. *In vivo* analysis of GreA/GreB and DksA on various *E. coli* strains showed that these proteins can display some degree of redundancy in certain cases and might also work antagonistically to one another (Vinella et al. 2012).

While DksA displays some degree of diversity with respect to its effects on transcription, another SCBF, Gfh1, can be clearly distinguished from the Gre factors owing to its clear inhibitory role in elongation. A crystal structure of *T. thermophilus* Gfh1 (Symersky et al. 2006) revealed that unlike the previously discussed SCBFs, this factor contained a more flexible coiled-coil domain with four acidic residues at the tip, enabling it to position RNAP core elements in an inactive state. It is interesting to note that in the case of bacterial SCBFs, these proteins all display surprising homology in the sequences, especially within the coiled-coil domains. A more detailed understanding of the mechanisms through which each of them works to enhance or inhibit transcription is still needed.

1.3 Proofreading

During RNA synthesis, multiple mechanisms ensure that the correct NTP is added to the nascent RNA. However, this does not negate the possibility of an incorrect substrate being incorporated. In the event of a mismatched NTP being incorporated into the RNA transcript, a series of proofreading mechanisms can kick in. The process of proofreading is two-fold: the EC first undergoes backtracking, which is then followed by removal of the unwanted portion of the RNA. The cleavage reaction responsible for removal of the backtracked portion of the RNA is often catalysed in the presence of RNAP-binding transcription factors which specifically bind to its SC. In addition to the interactions involving proofreading factors, specific elements within the enzyme core also participate in catalytic mechanisms. However, a general consensus of the role of these elements hasn't been reached yet. All of these points are explored in this section.

1.3.1 Backtracking

In 1992, DNase footprinting of *E. coli* TECs along single transcription units (Krummel and Chamberlin 1992) showed large template-dependent irregularities in the size and position of the DNase footprints, which did not agree with the long-standing idea of a single unchanging polymerase moving along the DNA during elongation. While acknowledging the limitations within that particular study, the results did lead the authors to postulate an "inchworming" model of elongation, in which RNAP would expand and contract in order to translocate forward along the DNA. Through mapping of the DNA-binding, catalytic, and RNA-binding sites of a TEC, it was subsequently shown that for the most part, RNAP demonstrated a fairly monotonous movement along the DNA, with the proposed inchworming only taking place at specific DNA sites (Nudler, Goldfarb, and Kashlev 1994). This idea of RNAP inchowrming along the DNA at any sites was finally put to rest a few years later when RNA-DNA cross-linking data (Nudler et al. 1997) showed that the TEC maintains an 8-9 base pair register throughout elongation, and that the phenomenon taking place at what had been identified as inchowrming sites was, in fact, simply the complex reversibly sliding backwards along the DNA –backtracking.

Transcription is not a continuously smooth process. Elongation is frequently interrupted by pausing, in which the transcribing complex is stalled. The role of pausing in transcription regulation has been discussed in the previous section (section 1.2.5). Here, the focus will be more on the specific roles that backtracking plays, and its effects on elongation. Backtrack-stabilised pausing was initially thought to be predominant for promoter-proximal pausing, in which contacts with initiation factors, lack of nascent RNA secondary structures and absence of trailing ECs and ribosomes would make the EC more prone to backtracking. Later, results published by Churchman and Weissman (2011) showed that backtracking also stabilised pausing away from the promoter site.

Backtracking is characterised by the translocation of RNAP backwards relative to the DNA. In 1997 Komissarova and Kashlev (1997) proved that this is accompanied by the backward translocation of RNAP. Through DNA and RNA footprinting, they showed that when RNAP backtracks, it threads the 3'-end of the nascent RNA out of the A-site. X-ray structures of backtracked elongation complexes from *S. cerevisiae* (Cheung and Cramer 2011; Wang et al. 2009) and *T. thermophilus* (Sekine et al. 2015) showed that inactivation of elongation takes place through extrusion of the 3'-end of the nascent RNA out of the ative site and into the RNAP SC (pore and funnel in eukaryotic Pol-II). However, these structural studies relied on complex formation with mutant proteins or chimeric transcription fcators. Recent work from (Abdelkareem et al. 2019) structurally described the entire process of backtracking in *E. coli* RNAP, from the first step in which RNAP enters the backtracked state, to the cleavage of the extruded RNA, to finally restarting transcription. Backtracking plays an extremely significant role in transcription regulation,

When backtracking takes place, the extruded portion of the RNA gets pushed into the SC of the enzyme. For RNA polymerisation to take place, substrate loading must be allowed to happen first. The A-site needs to be free and available for NTP-binding, and the SC must be available for NTPs and metal ions to enter the core. During backtracking, both of these sites are occupied, and RNA extension is stalled. To restart the process, the extruded RNA occupying the A-site and blocking the SC must be cleaved, or cut off, in a hydrolysis reaction that is intrinsic to RNAP and which can also be accelerated in the presence of TFs.

Backtracking doesn't only play a role in regulating transcription elongation. It has been shown that backtracking could be linked to genome instability. Backtracked ECs are more likely to collide with replisomes, resulting in double strand break (DSB)s (Dutta et al. 2011). Mechanisms in place which prevent backtracking or rescue ECs from a backtracked state, such as coupled translation (Proshkin et al. 2010) and the Gre elongation factors contribute towards genome stability. The same paper from Dutta et al. showed that the frequency of mutations seen in GreB-deficient *E. coli* cells was higher than in the wild-type strain, the former being more prone to backtracking induced DSBs. They suggested that the occurence of increased DSBs due to backtracking could explain the mutagenic response to stresses like nutrient depletion and antibiotic exposure.

1.3.2 Hydrolysis

Cleavage of the extruded RNA must take place in order for RNAP to escape its backtracked state and resume transcription. This is achieved through hydrolysis of the phosphodiester bond between the residues in the A-site and the P-site. The process of cleavage proceeds in a similar fashion to the sequence of reactions that take place during the formation of new phosphodiester bonds (Figure 1.11). In this case, cleavage of the phosphodiester bond between the bases in the A-site and P-site takes place through nucleophilic attack of the bond by an activated water molecule. A hydroxyl ion, generated through donation of a proton by a water molecule in the active site, attacks the RNA phosphodiester bond between the P-site and A-site, with the leaving group being the newly generated 3'-end of the shortened RNA.

Hydrolysis requires the coordination of two metal ions(Beese and Steitz 1991), similar to what we know is required for the exonuclease activity in DNA polymerase (Sosunov et al. 2003). Of these two metal ions, the first, Mg-I, is stably bound while the binding of the second ion, Mg-II, is more transient in nature. Hence, we see that in elongation complexes in which catalytic activities are meant to be observed, the resolution for Mg-II is either missing or noticeably poorer than that of the first. There has also been evidence of there being a third metal ion-binding site (Wang et al. 2006). In that study, Pol-II substrate-bound elongation complexes in low-Mg^{2+} showed two distinct metal-binding sites, in which the Mg^{2+} ions were coordinated by specific residues in the enzyme core as well as by the α, β, and γ phosphate of the NTP occupying the A-site. These two sites are consistent with the two distinct sites for coordination of Mg-I and Mg-II during catalysis. At high concentrations of Mg^{2+} , they observed that Mg-II was missing and Mg^{2+} was instead positioned at a third site. Unlike the lower occupancy typically observed for Mg-II, the occupancy at this third distinct site was comparable with that of Mg-I. While it is important to note that there is evidence of a third site for Mg^{2+}-coordination, occupancy of any ions at this site during catalysis has only been observed in the presence of high Mg^{2+} concentrations. There hasn't been conclusive evidence for this third site being occupied at lower Mg^{2+} concentrations.

Not only are the two Mg^{2+} ions important for catalysis, but also additional residues within the enzyme itself. Relevant to the two ions is an aspartate triad (Asp460, Asp462, Asp464 in *E. coli*) in the β' subunit. When RNAP is in an inactive state, these three residues together coordinate Mg-I. For either of the catalytic mechanisms at the A-site to take place, a second ion is needed. To accomodate for the binding of Mg-II, the three residues reorient themselves in what has been described as 'active centre tuning' (Sosunova et al. 2013). In this transient state, Mg-I remains bound by the three aspartates, albeit with fewer coordination bonds, while the coordination of Mg-II is assissted by Asp462 and Asp460.

The backtracked RNA itself assists in hydrolysis, as demonstrated by Zenkin, Yuzenkova, and Severinov (2006). In it, they tested the cleavage of RNA in different correct and misincorporated elongation complexes under different conditions such as increasing concentrations of Mg^{2+} and introducing non-hydrolysable NTPs. In their results, they showed that dinucleotide cleavage of misincorporated RNA is efficient, with the extruded RNA likely positioning itself in the E-site of the active centre. The E-site is the position at which an NTP substrate first binds to the complex before being screened and positioned at the A-site. Most importantly, they observed that the cleavage rates for different complexes containing different misincorporated nucleotides was differet, leading to their model of "product-assisted catalysis".

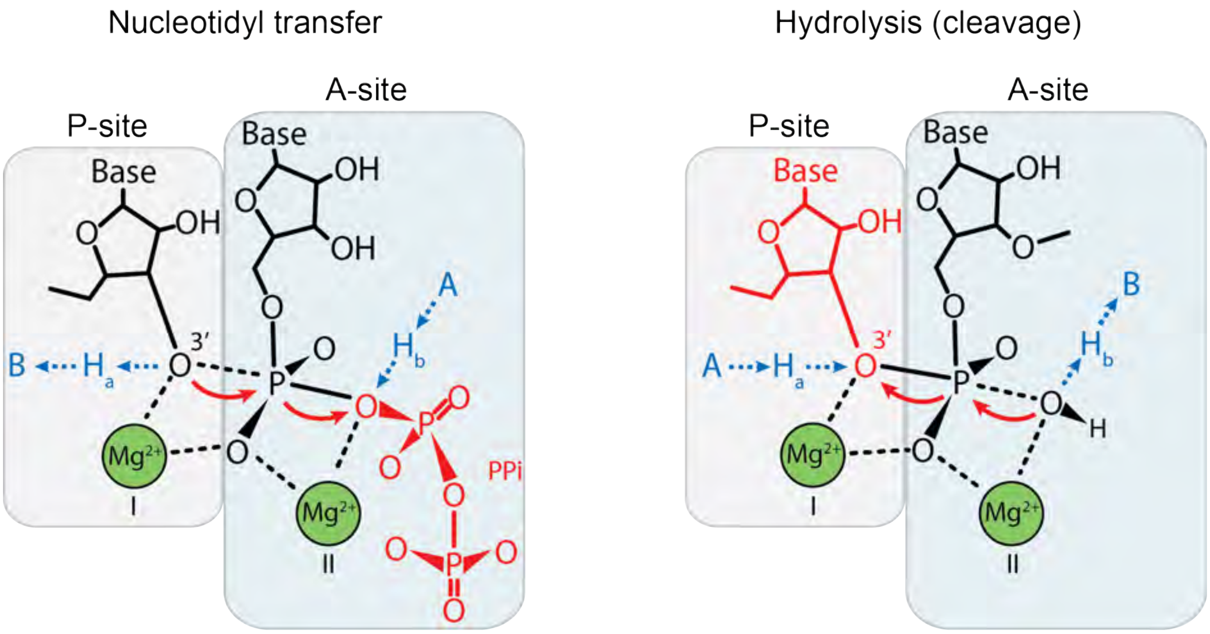
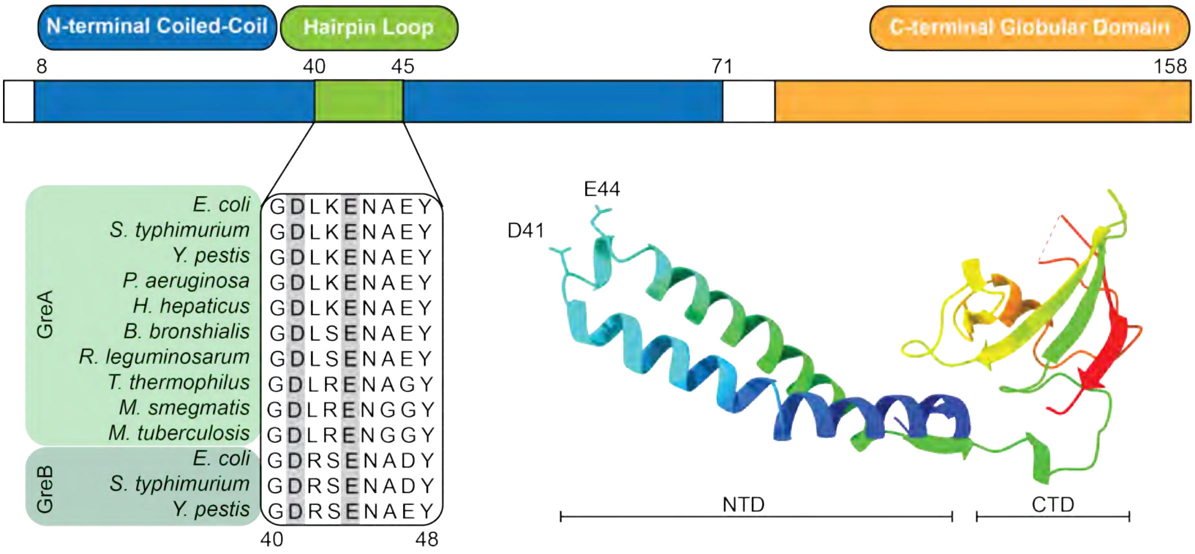


Figure 1.11 – Schematic showing the reactions responsible for formation and hydrolysis of phosphodiester bonds. The red arrows denote the direction of transfer of electrons. Both are nucleophilic substitution reactions (S_N2), with the 3'-OH in the P-site and the 2'-OH in the A-site acting as the nucleophiles in each case

1.3.3 GreA

The intrinsic cleavage activity of RNAP through which it hydrolyses the extruded portion of the RNA is accelerated in the presence of certain TFs. In bacteria, the TFs which stimulate cleavage are known as the Gre factors, named as such after they were discovered for their growth regulatory effects on transcription elongation in the 1990s (Borukhov et al. 1992). Both GreA and GreB have been identified for their strong influence on transcription cleavage as well as for inhibiting transcriptional arrest (Orlova et al. 1995). Structurally, they are analogous to other bacterial SCBF. Functionally, they are analogous



GreA	*E. coli*	GDLKENAEY
	S. typhimurium	GDLKENAEY
	Y. pestis	GDLKENAEY
	P. aeruginosa	GDLKENAEY
	H. hepaticus	GDLKENAEY
	B. bronshialis	GDLSENAEY
	R. leguminosarum	GDLSENAEY
	T. thermophilus	GDLRENAGY
	M. smegmatis	GDLRENGGY
	M. tuberculosis	GDLRENGGY
GreB	*E. coli*	GDRSENADY
	S. typhimurium	GDRSENADY
	Y. pestis	GDRSENAEY

Figure 1.12 – Domain organisation and structure of *E. coli* GreA. The sequence alignment shown is that of the conserved hairpin region in GreA and GreB across different bacterial species. Structure of GreA (PDB 1grj) includes the two conserved residues, D41 and E44.

to the SC-binding cleavage factors TFIIS (or SII) in eukaryotes (Izban and Luse 1993) and TFS in archaea (Lange and Hausner 2004). Whilst structurally dissimilar, GreA/GreB, TFS and TFIIS influence cleavage by contacting the respective RNAP active site with two conserved acidic residues. The similarities between the bacterial and archaeal/eukaryotic cleavage factors essentially ends at this point. The major differences between these TFs lies in their structures. Compared to their eukaryotic and archaeal counterparts, GreA and GreB are relatively simpler, and are comprised of a coiled-coil NTD and a globular CTD. On the other hand, the Pol-II TF TFIIS is made up of three domains (Morin et al. 1996). Of these, the N-terminal domain I is non-essential for cleavage activity, the central domain II is needed for binding to Pol-II, and the C-terminal domain III forms a zinc ribbon that contacts the active site and promotes cleavage. The two conserved acidic residues which are positioned at the tip of the coiled-coil domains in GreA/B are positioned in a β-hairpin in the domain III Zn ribbon in TFIIS (Kettenberger, Armache, and Cramer 2003). In the case of Pol-I and Pol-III, there are not any external RNAP-binding factors that promote cleavage. Rather, this role is fulfilled by specific subunits within the enzymes themselves (Alic et al. 2007). Pol-I and Pol-III all contain TFIIS-like subunits – A12 and C11 respectively. The reason why these polymerases rely on their proofreading subunits rather than an external factor likely lies in fact that they are responsible for transcribing RNA that does not get translated, making fidelity at this step extremely vital. Archaea also contain a TFIIS-like factor, called TFS. The TFS gene was first identified by Langer and Zillig (1993) and later functionally analysed by Hausner, Lange, and Musfeldt (2000).

As with other bacterial SCBFs, *E. coli* GreA contains a coiled-coil NTD and a globular CTD, shown in Figures 1.10 and 1.12. The NTD contains a hairpin loop, the residues for which are conserved in GreA and GreB across various bacterial species. The two most important residues within the hairpin are the two acidic residues, an aspratate and a glutamate (D41 and E44), which interact with the active centre of RNAP and are known to influence hydrolysis. *E. coli* GreA shares many other conserved residues with the GreA proteins in other species, as seen in Figure 1.12. The relevance of GreA for maintaining transcription fidelity was described by Bubunenko et al. (2017), whose results showed that deletion of GreA in *E. coli* produced a hundred-fold increase in transcription errors compared with the wild-type *E. coli.* They also showed that although ΔGreA cells showed an increase in misincorporation errors, over-expression of GreB in the same cells reduced the errors to wild-type levels.

Following the identification of GreA as a cleavage factor, efforts were made in understanding the roles of its two domains as well as to identify its differences with the other cleavage factor GreB. The role of the Gre NTDs in distinguishing between their specific roles was explored through comparisons of their crystal structures and testing the functionality of GreA/GreB hybrids (Koulich et al. 1997). By exchanging the NTD residues of GreA with the equivalent residues of GreB, the hybrid protein switched its preference for di-nucleotide cleavage with tetra-nucleotide cleavage. These experiments showed while the NTD was important for distinguishing between the types of cleavage, the CTD had no effect. Along the NTD, GreB also has a much more prominent patch of basic residues, which would be important for stabilising longer lengths of backtracked RNA in the SC for cleavage. Shortly after this, Koulich, Nikiforov, and Borukhov (1998) expressed the

GreA NTD and CTD separately to more accurately define their roles. *In vitro* assays in the presence of the NTD, CTD, and the whole protein revealed that the NTD is essential for cleavage activity. The CTD on its own did not induce cleavage of the RNA transcript, but showed strong cleavage along with the NTD. Competition binding assays in the same paper showed that CTD is responsible for specific binding of GreA to RNAP. Manual docking of the *E. coli* GreA and *T. thermophilus* RNAP holoenzyme resulted in a model in which the GreA CTD binds to the outer portion of the SC where it would also be able to contact the sequence insertion 3 domain of *E. coli* RNAP, with the tip of the NTD contacting active centre elements (Laptenko et al. 2003). In the same work, the importance of the two acidic residues at the NTD tip (D41 and E44) for coordination of Mg^{2+} required for catalytic activity was demonstrated through mutational analysis of the tip, in which the acidic residues were mutated accordingly to either disrupt interaction forces or steric fit within the active centre. A paper from Sosunova et al. (2003) in the same year also showed that D41 and E44 are essential for coordination of magnesium in the A-site, with additional data showing that the increased cleavage activity triggered by the Gre factors is extremely likely due to increased retention of the second magnesium ion (Mg-II) needed for catalysis in the two metal-ion model.

While GreA is important for elongation, it is not essential for cell growth. The main reason for this is certainly that RNAP is capable of intrinsically cleaving the backtracked portion of the RNA. However, another explanation might be that rate of elongation is tied to the cooperation of different ECs along a single transcription unit. Epshtein et al. (2003) demonstrated that, *in vitro*, ECs were capable of reading through site-specific DNA roadblocks when assisted by additional ECs transcribing the same DNA. The read-through was attributed to cooperation of ECs and not cleavage activity by comparing the same reaction with the addition of equimolar quantitties of GreB.

1.3.4 Role of the Trigger Loop

Part of the β' subunit, the trigger loop is an element of RNAP that is present in the catalytic core of RNAP. Situated in close proximity to the BH, it is a key element in

RNAP core, which participates in substrate selection and NTP addition. It exists in two states – folded and unfolded, described structurally both in bacterial RNAP (Vassylyev et al. 2007) and in Pol-II (Wang et al. 2006). When folded, it forms a helical bundle referred to as the trigger helices (TH). Many studies have been published over the years which have focused on assessing the role that the TL plays in the catalytic mechanisms within the RNAP core across different species. Of these, a large portion of them have centred around one key residue within the TL: a universally conserved histidine (β' His 936 in *E. coli*, β' His 1242 in *T. aquaticus*, Rpb1 His 1085 in *S. cerevisiae*) which contacts the A-site when the TL is in its folded state.

The TL also influences the translocation of RNAP along the DNA by interacting with the BH. Owing to its proximity to the BH, it makes sense that the conformation of the TL would then affect the BH. A model for such a mechanism was reported by Bar-Nahum et al. (2005), which describes the movement of RNAP along the DNA being governed by the binding of the incoming substrate and bending of the BH. The bending and straightening of the BH drives the EC forward, with the TL in turn controlling the rate of oscillation between the two BH states and the equilibrium between them. The influence that the TL has on the BH explains its role in pausing. Pausing of ECs, which takes place frequently throughout the elongation phase of the transcription cycle, is stabilised either by RNA hairpin formation or by backtracking. Hairpin-stabilised pausing of an EC is accompanied by rearrangement of elements within the A-site. Prior to entering a stabilised paused state, the TEC enters an intermediate 'elemental' paused state. By studying RNAPs containing TL deletions and mutation, Toulokhonov et al. (2007) proposed a model for TL-mediated rearrangement of the A-site in an elemental paused state. Structural studies on elemental paused complexes (Weixlbaumer et al. 2013) later revealed that in all of their structures, the TL appeared disordered, suggesting that its role in rearangemet during pausing might be indirect through its influence on BH conformations

Substrate selection and Nucleotide Addition

Analysis of different Pol-II elongation complexes, described in (Wang et al. 2006) paints a clear picture of the role that the TL plays during substrate selection and during catalysis. Of the 14 structures they looked at, only the two containing the correct NTP bound to the A-site had the TL positioned in close proximity to the A-site. They showed that the TL influences substrate selectivity through direct and indirect interactions with the NTP in the A-site: it interacts directly with the β-phosphate and the base, and indirectly with the 2'- and 3'-OH groups of the ribose. The exact positioning of the TL was seen to be influenced by contacts with different residues within the enzyme itself, indicating a system in which the TL positioning would allow for discrimination against binding of deoxynucleoside triphosphate (dNTP)s as well as pyrimidine-pyrimidine and purine-purine pairings. In addition to its role in substrate selection, the TL would also participate in formation of the phosphodiester bond. Through its contact with the β phosphate of the NTP in the A-site, the conserved histidine would assist in the SN2 attck on the RNA 3'-OH, leading to the formation of a new phosphodiester bond. These models for TL-mediated substrate selection and bond formation also points towards the TL coupling the two processes. The release of a pyrophosphate upon successful incorporation of a complementary base would then displace the histidine and other elements near the A-site, unfolding the TL and subsequently leaving room for the BH to facilitate translocation of the EC.

Hydrolysis

Of the two catalytic processes that take place within the RNAP core –nucleotide addition and hydrolysis –the role that the TL plays in the latter remains somewhat unclear.

Two key papers on RNA hydrolysis in bacteria which highlighted the importance of the TL came from Yuzenkova and Zenkin (2010) and Roghanian, Yuzenkova, and Zenkin (2011). In the first one, which assessed the role of the TL during intrinsic hydrolysis (in the absence of cleavage factors) and the mechanism by which it might participate in cleavage, the authors compared cleavage between wild-type RNAP and RNAP containing point mutations of the TL histidine (H1242A) and deletions of the entire domain (ΔTL). *In vitro* assays showed that intrinsic cleavage was significantly reduced for both the mutants. In addition to this, their data supported a transcript-assisted cleavage model previously proposed by the same group (Zenkin, Yuzenkova, and Severinov 2006), by showing that the TL is essential for preferentially hydrolysing the second phosphodiester bond in 1 nucleotide backtracked complexes, possibly by orienting the 3' nucleoside monophosphate (NMP). Importantly, the same experiments with *E. coli* RNAP yielded the same results. Building on these results, Roghanian, Yuzenkova, and Zenkin (2011) tested the effects of GreA on cleavage in WT RNAP and ΔTL RNAP in *T. aquaticus*. Based on the results of their *in vitro* kinetic assays, they suggested that the enzyme active centre might switch between a "GreA active centre" and a "TL active centre", depending on whether the TEC is in a backtracked state or a productive elongation state. This model of the active centre switching between GreA and TL-mediated catalysis was based on very thorough examination of their effects on hydrolysis. Intrinsic cleavage activity that was lost through deletion of the TL was mostly restored on addition of GreA. However, comparisons of these kinetics with those involving the H1242A mutation in the TL showed that not only was the GreA-assisted cleavage comparable with that of GreA + WT RNAP, it was also quicker than cleavage in the GreA + ΔTL RNAP.

In their *in vitro* assays performed for studying the catalytic role of the TL, Zhang, Palangat, and Landick (2010) constructed *E. coli* RNAP mutants in which two adjacent residues within the TL (β' 930 and 931) were substituted for proline, preventing it from folding into the TH. These mutations did not affect tri-nucleotide transcript cleavage in the presence and absence of the cleavage factor GreB. Deletion of the entire TL domain, which includes the *E. coli*-specific sequence insertion 3 (SI3) domain, drastically reduced the rate of GreB-assisted cleavage, suggesting that Gre factor-mediated hydrolysis requires the SI3 domain and not the folded TH. The SI3 domain, also known as β'i6, has been proposed to be part of the gating system that mediates binding of different transcription factors to the SC (Furman et al. 2013). Structural data on backtracked Pol-II (Wang et al. 2009) also showed the TL in a partially open conformation, suggesting that its folding into the TH might not be vital for hydrolysis to take place.

The TL is likely necessary for cleavage to take place, but was initially thought to participate through the invariant histidine, which could function as a general base in catalysis. However, for the histidine to be able to participate in cleavage, the TL needs to fold into the TH to allow the residue to contact the active site. Biochemical and structural data of backtracking and backtrack rescue RNAP from different species have shown that folding of the TL is not important for hydrolysis, leading to the question of how the domain might play a role in cleavage. Recently, Mishanina et al. (2017) proposed a model for catalysis within the RNAP active site in which, rather than participating as a general acid-base, the TL could instead function as a positional catalyst.

1.4 Cryo-Electron Microscopy

The structures for all the complexes described in the results (Chapters 4, 5, and 6) were obtained through single-particle cryo-EM. The history of electron microscopy being used in biology dates back to the mid-1900's, culminating with the so-called 'resolution revolution' in the last decade.

1.4.1 Architecture of a TEM

All transmission electron microscopy (TEM)s have the same general layout (Figure 1.13). At the very top is an electron gun, which generates a highly coherent electron beam generally in the range of 100-300 keV. Electron guns in modern-day high-resolution electron microscopes used for cryo-EM are of the field emission type, which means that electron beams are produced as a result of a strong electric field generated by gun tip and a positively charged anode. Field emission guns (FEG) were preceded by thermionic emission guns which generated less coherent beams with larger diameters. There are two types of FEGs –a Schottky-type FEG (used in the Titan Krios, Glacios, Polara, among others) and a cold-FEG (JEOL Cryo-ARM 300 microscope).

Once generated, the electron beam passes into the illumination system of the microscope. In principle, the different sets of components work analogously to the lens system in an optical microscope. Electromagnetic lenses are used to manipulate the electron beam through the formation of a strong magnetic field generated by passing a strong current through a coil. The first set of lenses are the condenser lenses, which focus and control the size of the beam. Defects in the condenser lens are responsible for generating spherical and chromatic aberrations as well as astigmatism in the beam. This is followed by the objective lens, which is the largest and the strongest of the lenses within the microscope. The specimen to be imaged is placed within the objective lens. After passing through the specimen, the objective lens focuses them to form the beam image in the image focal plane of the lens. The intermediate lens allows the user to switch between the imaging mode and diffraction mode, in which the diffraction pattern formed in a different plane of the objective lens is projected onto the detector.

The detector used in an electron microscope is one of the biggest resolution-limiting factors within the imaging system. The development of direct electron detectors like the Gatan K2 summit and the Falcon detectors contributed to the boom in high resolution cryo-EM structures.

1.4.2 Cryo-Electron Microscopy & SPA

In the initial years of transmission electron microscopy being used for imaging biological specimens, samples were stained with heavy metals to achieve high contrast (see Figure 1.1). The limiting factor at this stage was specimen damage within the electron beam. Specimen damage could be reduced by maintaining the samples at cryogenic temperatures, which was shown to be feasible by Taylor and Glaeser (1974). Jacques Dubochet and Alasdair McDowall changed the game when they showed that rapid freezing could prevent water from forming ice crystals (Dubochet and McDowall 1981). A few years later, they published electron electron micrographs of unstained viruses embedded in vitreous, or amorphous, ice (Adrian et al. 1984).

Around this time, progress was also being made in the determination of 3D models of specimens through electron microscopy. An early example was the determination of a 3D model of the purple membrane from *Halobacterium halobium* through a combination of electron imaging and diffraction. Simultaneously, techniques were being developed to generate 3D maps from 2D projections of particles in electron micrographs (Frank et al. 1978).

A combination of the advances in sample preparation and computational analysis led to the development of cryo-EM as an alternative to x-ray crystallography for biomolecular structure determination, especially for samples that were challenging to crystallise. Embedding the sample particles in vitreous ice and imaging them at lower electron doses does reduce the extent of radiation damage produced in the sample, but also results in very noisy data from which the signal needs to be extracted. Projection averaging algorithms form Penczek, Grassucci, and Frank (1994) solved this, with computational advancements in the following decades drastically improving the quality of data. Of note was the development of beam-induced motion correction, which corrected for the blurring produced by particle movements within the electron beam (Brilot et al. 2012). The giant strides made in direct electron detector technology further pushed cryo-EM to the forefront of structure determination by allowing for high resolution structures to be determined from smaller datasets, as demonstrated by Bai et al. (2013).

Cryo-electron microscopes are used for a variety of techniques:

- single-particle analysis
- cryo-electron tomography
- electron crystallography
- micro-ED (electron diffraction)

In single-particle analysis, a dataset comprising of individual movies is collected. Each movie comprises of a fixed number of frames, each recorded at a low electron dose (typically around 1 e$^-$/Å^2/frame) so as to not bombard the imaged area with a high dose all at once. The frames within each movie are dose-weighted and motion corrected, following which estimation of the contrast transfer function (CTF) allows for determining defocus and astigmatism (Erickson and Lug 1971). Individual particles are then picked from the micrographs. 2D class averaging of the picked particles is the first step in cleaning up the dataset by removing classes representing ice, contaminants, and dissociated complexes. The particles are then used to produce a map *de novo* (also called an *ab initio* map in some cases). This map is then used to generate a high-resolution map.

Single-particle analysis has proven to be a powerful tool for solving structures of transcription complexes. These complexes are often flexible, adopting short-lived but functionally relevant states. Through SPA, we can visualise these complexes in their native states, often with multiple conformational states of a complex solved within a single dataset (Hanske, Sadian, and Müller 2018).

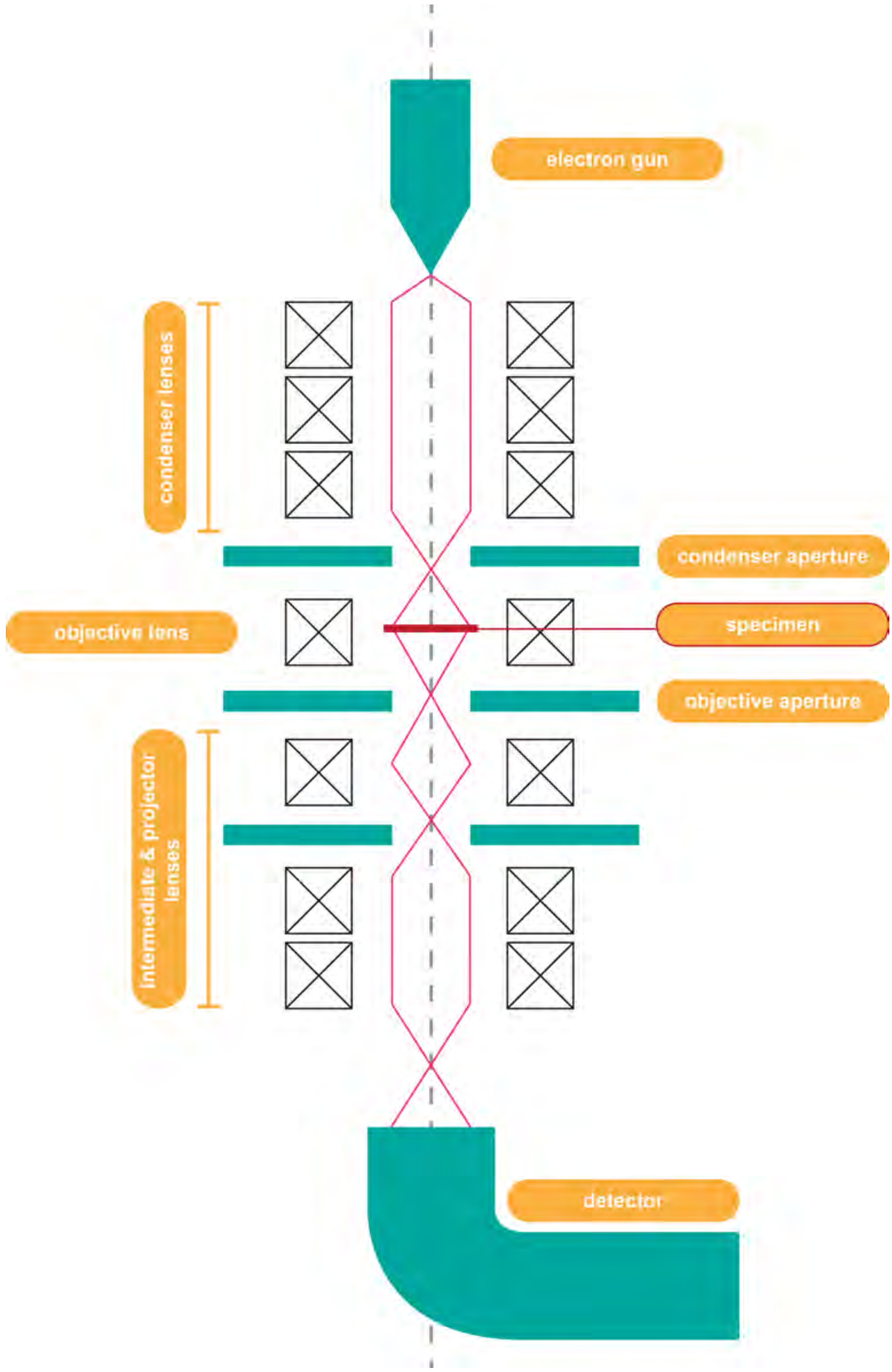


Figure 1.13 – Layout of a transmission electron microscope

Chapter 2

Aims and Objectives

The first step of gene expression – transcription – is carried out by RNA polymerases. Many of the core architecture, catalytic mechanisms and regulatory processes are largely conserved across the three kingdoms of life. As a direct result of this, insights that we gain from studying transcription mechanisms in a more straightforward model like *Escherichia coli* has implications in understanding the same mechanisms in more complex transcription systems like the eukaryotic RNAPs.

Here, the focus of the project was to address some questions related to proofreading in transcription. Elongation complexes backtracked by a short length as a result of misincorporations in the nascent mRNA are rescued through cleavage of the backtracked portion of the RNA. The active site of RNAP, which catalyses the formation of new bonds, also catalyses the hydrolysis of RNA bonds. Hydrolysis is known to be more efficient in the presence of specific proofreading factors (GreA in bacteria, TFIIS in eukaryotes and TFS in archaea), and has been studied extensively over the past two of decades. However structurally, we are yet to have a more complete understanding of proofreading for a few reasons. A number of backtracked Pol-II complexes had been solved by x-ray crystallography (Kettenberger, Armache, and Cramer 2003; Wang et al. 2009; Cheung and Cramer 2011). However, these structures didn't provide us with the in-depth understanding of proofreading, each for a different reason. While the Pol-II + TFIIS structure from Kettenberger, Armache, and Cramer (2003) did not contain RNA, the structure published by Cheung and Cramer (2011) was backtracked by 8 nucleotides which is much longer than the backtracked length involved in transcriptional proofreading. The complexes published by Wang et al. (2009) were backtracked by short lengths of RNA, however they used a mutant form of TFIIS, with one of the acidic tip residues (Glu 291, equivalent to Glu 44 in *E. coli* GreA) was mutated to a histidine. Two structures of backtracked complexes in bacteria were also solved by x-ray crystallography (Sekine et al. 2015). These structures of RNAP from *Thermus thermophilus* were backtracked by 1 nucleotide, however the transcription factor (TF)-bound structure used a GreA-Gfh1 chimeric protein. The authors were also unable to model the nucleic acids in the RNAP + TF structure. The cryo-EM structures of *E. coli* backtracked complexes from Abdelkareem et al. (2019)

were backtracked by lengths of RNA implicated in backtrack-stabilised pauses rather than proofreading, and also contained GreB. As of this point, no published structures of wild-type GreA bound to a short backtracked RNA Polymerase (RNAP) complex exist. This is due to the fact that GreA had a low affinity for RNAP, making the complex difficult to capture without modifying the TF. Single-particle cryo-EM allows the user to capture a complex in a short-lived state. This, combined with the recent advances that have made it possible to obtain higher resolution reconstructions, meant that it would be possible to capture a complex of a wild-type GreA-bound *E. coli* RNAP complex backtracked by a short length, in its pre-catalytic state.

Through *in vitro* cleavage assays and single-particle cryo-EM, I set out to structurally and functionally characterise a complex in its pre-catalytic state. The complex would need to be backtracked by up to 2 nucleotides with a misincorporated NMP at the 3'-end of the RNA, and with the wild-type form of GreA bound to the secondary channel, without any modifications to the protein that would affect its interactions with the backtracked complex.

This complex would allow us to answer certain questions pertaining to proofreading in transcription in *E. coli*:

- What is the exact role of GreA in hydrolysis?
- Despite the extensive structural and sequential homology between GreA and GreB, how does the complex differentiate between these two proteins when backtracked by different lengths?
- Does the trigger loop actively participate in hydrolysis, and if so, how?
- Are there structural changes that would explain how a backtracked base might assist in its own hydrolysis?

Chapter 3

Materials and Methods

3.1 Protein Purification

Purification of *E. coli* RNA Polymerase

Lysis buffer	
Tris, pH 8.0 at 4°C	50 mM
Glycerol	5%
EDTA	1 mM
$ZnCl_2$	10 µM
DTT	10 mM
PMSF	0.1 mM
Benzamidine	1 mM

10x TGE	
Tris, pH 8.0 at 4°C	100 mM
Glycerol	50%
EDTA	1 mM

PEI wash buffer	
TGE	1x
NaCl	0.5 M
$ZnCl_2$	10 µM
DTT	1 mM
PMSF	0.1 mM
Benzamidine	1 mM

PEI elution buffer	
TGE	1x
NaCl	1 M
$ZnCl_2$	10 µM
β-mercaptoethanol	5 mM
PMSF	0.1 mM
Benzamidine	1 mM

All proteins were expressed in and purified from an *Escherichia coli* LOBSTR strain (Andersen, Leksa, and Schwartz 2013) with a knock-out for the RNase I and II genes (*E. coli* LACR II, to be published). *E. coli* RNA Polymerase (RNAP) containing a β' subunit with a C-terminal His_{10}-tag was expressed in *E. coli* LACR II. An LB culture containing 100 µg/ml Ampicilin and 34 µg/ml Chloramphenicol was induced with 0.5 mM IPTG at an OD_{60} of 0.8, for 3 hours at 37°C. The cells were subsequently harvested, pelleted, frozen and stored.

On the first day of the purification, the cells were thawed and resuspended in the lysis buffer. Protease inhibitor cocktail (PIC) tablets (1 tablet per 50 ml) and DNase I (20 mg/ml) were then added before sonication of the resuspended cells. PEI precipitation was then carried out by adding 10% polyethyleneimine (PEI) solution at pH 8.0 to the clarified lysate drop-wise. The precipitate was collected by centrifugation and then washed with the PEI wash buffer by resuspending it in the buffer and centrifuging until the supernatant turned colourless. The precipitate pellet was then resuspended in the PEI elution buffer and centrifuged, keeping the pellet, which was resuspended in a small volume of the PEI elution buffer.

The next step was ammonium sulphate precipitation, in which ammonium sulphate powder was added slowly to the pooled elution fractions from the previous step, to reach a final concentration of 35 g per 100 ml. This solution was left stirring on ice overnight in a cold room.

IMAC buffer A	
Tris, pH 8.0	20 mM
NaCl	1 M
Glycerol	5%
β-mercaptoethanol	5 mM
PMSF	0.1 mM
Benzamidine	1 mM
$ZnCl_2$	10 µM

IMAC buffer B	
Tris, pH 8.0	20 mM
NaCl	1 M
Glycerol	5%
Imidazole pH 8.0	250 mM
β-mercaptoethanol	5 mM
PMSF	0.1 mM
Benzamidine	1 mM
$ZnCl_2$	10 µM

The precipitate was pelleted the following day, and resuspended in IMAC buffer A for the affinity chromatography step. The sample was loaded onto a 20 ml Ni Sephadex Fast Flow column (collecting the flow through (FT)), washed first with IMAC buffer A, followed by 2% buffer B, a 2-16% gradient of buffer B, and 16% buffer B. Elution was carried out with 100% buffer B and fractions were collected in 96-well plates. The column was then washed with buffer A.

Dialysis buffer 1	
Tris pH 8.0	20 mM
NaCl	1 M
Glycerol	5%
β-mercaptoethanol	5 mM
$ZnCl_2$	10 µM

The pooled elution fractions were dialysed in a dialysis membrane with His-ppx (1 mg per 5-7 mg of protein) in dialysis buffer 1 overnight in a cold room. The dialysed sample was loaded onto a 20 ml IMAC column equilibrated with IMAC buffer A, collecting the FT. Following a wash with buffer A, the column was eluted with 100% buffer B. SDS-PAGE gels confirmed that the subtractive IMAC successfully removed the protease in the elution step while keeping the protein in the FT.

Biorex dialysis buffer	
TGE	1x
DTT	1 mM
PMSF	0.1 mM
Benzamidine	1 mM
$ZnCl_2$	10 µM

Biorex buffer C	
TGE	1x
NaCl	100 mM
DTT	1 mM
PMSF	0.1 mM
Benzamidine	1 mM
$ZnCl_2$	10 µM

Biorex buffer D	
TGE	1x
NaCl	1 M
Dtt	1 mM
PMSF	0.1 mM
Benzamidine	1 mM
$ZnCl_2$	10 µM

The FT from the subtractive IMAC step was then dialysed to BioRex dialysis buffer for four hours before carrying out the ion exchange chromatography step using a 50 ml BioRex70 column. The sample was loaded onto the column equilibrated with BioRex buffer C. It was then washed with BioRex buffer C and eluted with a 0-100% gradient of BioRex buffer D. The peak fractions were pooled and concentrated.

SD buffer	
Hepes pH 8.0	10 mM
KCl	0.5 M
Glycerol	1%
DTT	2 mM
PMSF	0.1 mM
Benzamidine	1 mM
$ZnCl_2$	10 µM
$MgCl_2$	1 M

The concentrated pooled fractions were filtered to remove aggregates, and injected onto a 300 ml Superdex 200 Increase 26/60 column equilibrated with SD buffer. The column was washed with the SD buffer until the sample eluted.

Cryo-EM dialysis buffer	
Hepes pH 8.0	10 mM
KOAc	150 mM
MgOAc	5 mM
DTT	2 mM
$ZnCl_2$	10 µM

Peak fractions were pooled and dialysed into the Cryo-EM buffer overnight. The purified protein was concentrated to 70-80 mg/ml, aliquoted, frozen and stored at -80°C. The final yield of *E. coli* RNAP from 12 litres of culture was 26.4 mg, at a concentration of 71.5 mg/ml.

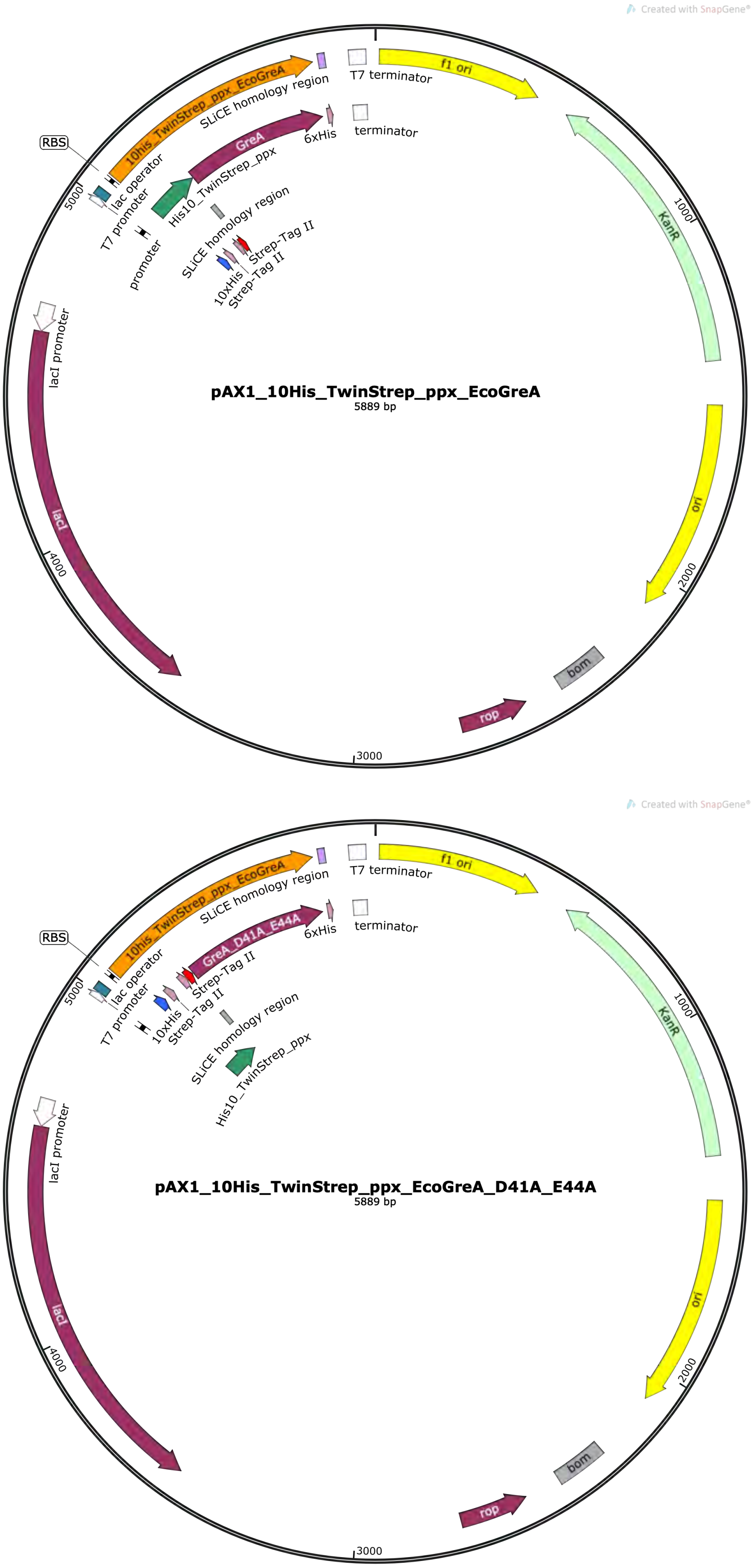


Figure 3.1 – Plasmid maps for GreA wild type (top) and mutant (bottom) expression vectors

Purification of *E. coli* GreA

Buffer A	
Tris-HCl, pH 7.5	40 mM
NaCl	600 mM
β-mercaptoethanol	2 mM

Buffer B	
Tris-HCl, pH 7.5	40 mM
NaCl	600 mM
Imidazole	250 mM
β-mercaptoethanol	2 mM

Buffer C
Buffer A + 2.5 mM desthiobiotin

Cryo-EM dialysis buffer	
Hepes-KOH, pH 8.0	10 mM
KOAc	150 mM
MgOAc	5 mM
DTT	2 mM
$ZnCl_2$	10 µM

E. coli GreA containing an N-terminal His_{10}-TwinStrep tag (Figure 3.1) was expressed in the *E. coli* LACR II strain. 6 litres of culture were grown at 37°Cin LB containing 50 µg/ml of the antibiotic Kanamycin (expression construct contains a Kanamycin resistance marker). 1mM IPTG added at an OD_{600} of 0.7 was used to induce overexpression of GreA for 3 hours at 37°C. The Cells were harvested and resuspended in 5 volumes of the lysis buffer along with an EDTA-free protease inhibitor cocktail (1 tablet/50ml, Sigma Aldrich). Cells were lysed by sonication. The lysate was then cleared by centrifugation at 30000g, 4°C for 30 minutes.

The cleared lysate was loaded onto a 5ml IMAC column (HiTrap HP) and washed with Buffer A. The sample was then eluted with Buffer B and loaded onto two 5ml StrepTrap HP columns. This was followed by a wash with Buffer A and elution with Buffer C. The peak fractions, containing the tagged protein, were pooled and incubated with HRV3C protease (1 mg/20 mg of protein) at 4°Cfor 15 minutes on a shaker. The sample, now containing GreA and the cleavaed tag as well as some amount of the uncleaved protein, was injected onto a subtractive IMAC column (HiTrap HP). This was washed with 0%

Buffer B before a gradient elution to 100% Buffer B.

The peak fractions of the cleaved wild-type GreA, collected in the wash step, were pooled and dialysed into the Cryo-EM buffer overnight. The protein was finally concentrated in a 4 ml Amicon concentrator (3 MW cut-off), spun several times for 10 minutes at 3500 rpm until an optimum concentration was reached. After reaching a concentration of 12 mg/ml (680µM), the protein was aiquoted and snap frozen in liquid nitrogen and stored at -80°C.

The purification of the *E. coli* GreA mutant (E41A D44A) followed the same procedure as the purification of the wild-type, with the exception that the mutant eluted during the gradient elution step at 38% buffer B. The final concentration of the *E. coli* GreA E41A D44A mutant was 16 mg/ml (7.2 mg from a 6 litre culture).

3.2 Transcription Assays

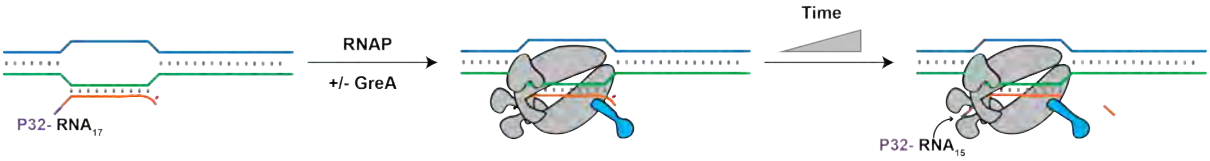


Figure 3.2 – Experimental schematic of the *in vitro* cleavage assays

All assays were carried out in the same cryo-EM buffer used for protein storage. The general experimental setup followed the schematic outlined in Figure 3.2. Chemically synthesised DNA and RNA oligonucleotides were annealed to form a nucleic acid scaffold on which RNAP could bind. Each pair of DNA oligonucleotides used in the scaffolds – the template DNA strand (tDNA) and the non-template DNA strand (ntDNA) – consisted of approximately 14 base pairs upstream and 14 base pairs downstream of the transcription bubble. 8 to 10 mismatched bases mimicking the transcription bubble within a transcription elongation complex served as the binding site for RNAP.

RNA Endlabelling

In order to visualise and therefore quantify the extension and cleavage of ribonucleic acid (RNA) in a scaffold for each *in vitro* cleavage assay, it needed to be labelled with a radioactive isotope, Phosphorus-32 (^{32}P, half life = 14.3 days). Labelling was done using an adenosine triphosphate (ATP) substrate consisting of ^{32}P at the γ-phosphate position (^{32}P-γ ATP).

For each scaffold, 5µM of the RNA was incubated with ^{32}P-γ ATP and T4 polynucleotide kinase (T4 PNK, New England Biolabs) at 37°C for 1 hour. For a 20µl reaction mixture, 2µl ATP was used. The labelling reactions were all carried out in the T4 PNK buffer supplied by the manufacturer (Buffer composition: 70 mM Tris-HCl, 10 mM $MgCl_2$, 5 mM DTT, pH 7.6 at 25°C). Following inactivation of the enzyme at 95°C for 5 minutes, the labelling mixture was spun down through a 1ml column loaded with Sephadex G50 resin (Mini bio-spin columns from Biorad, G50 resin from Sigma Aldrich. The supernatant, comprising of ^{32}P-endlabelled RNA, was used for annealing of scaffolds.

Cleavage Assays

The first step was the annealing of a scaffold made up of the tDNA and the labelled RNA. A mixture containing 4.5 µM of unlabelled RNA, 0.5 µM of the labelled RNA, and 10 µM of the tDNA in the reconstitution buffer (10 mM Tris-HCl pH 8.0, 40 mM KCl, 5 mM $MgCl_2$) was brought up to approximately 95°C and then allowed to cool down to room temperature in a water bath. The labelled scaffold was either used on the same day or stored at -20°C and used as soon as possible.

Per reaction:

EM buffer	1x (5 mM Mg^{2+})
Chapso	8 mM
DTT	1 mM
BSA	0.2 mg/ml
Scaffold	0.5 µM
Non-template DNA	0.5 µM
E. coli RNAP	1 µM
E. coli GreA (when needed)	5 µM

For each reaction, the buffer, detergent, DTT, Chapso, and scaffold were first mixed together. RNAP and the ntDNA were mixed separately and incubated for a few seconds over 37°C. To begin the reaction (time point 0), the reaction mixture was added to RNAP and, when applicable, GreA. At each pre-determined time point, an aliquot of the reaction mixture was added to an equal volume of denaturing RNA loading buffer (8 M urea, 20 mM EDTA pH 8.0, 5 mM Tris-HCl pH7.5, 0.5% bromophenol blue, 0.5% xylene cyanol). 1 mM NTPs were added following the final time point and left for an additional 5 minutes to allow for RNA extension to continue, before also taking a sample and mixing it in the RNA loading buffer. The tubes containing the individual time points were boiled at 95°C for 5 minutes, spun down, and stored. Samples were loaded onto denaturing polyacrylamide gels (15% or 20% acrylamide, 7M urea). Gels were exposed using phosphor imager screens at -80°C overnight, which were then scanned using a Typhoon PhosphorImager (GE Healthcare).

3.3 Cryo-Electron Microscopy

Sample Preparation

For single-particle cryo-EM, the best sample is often the most homogeneous. To ensure homogeneity for assembled complexes such as the ones described in this work, the best

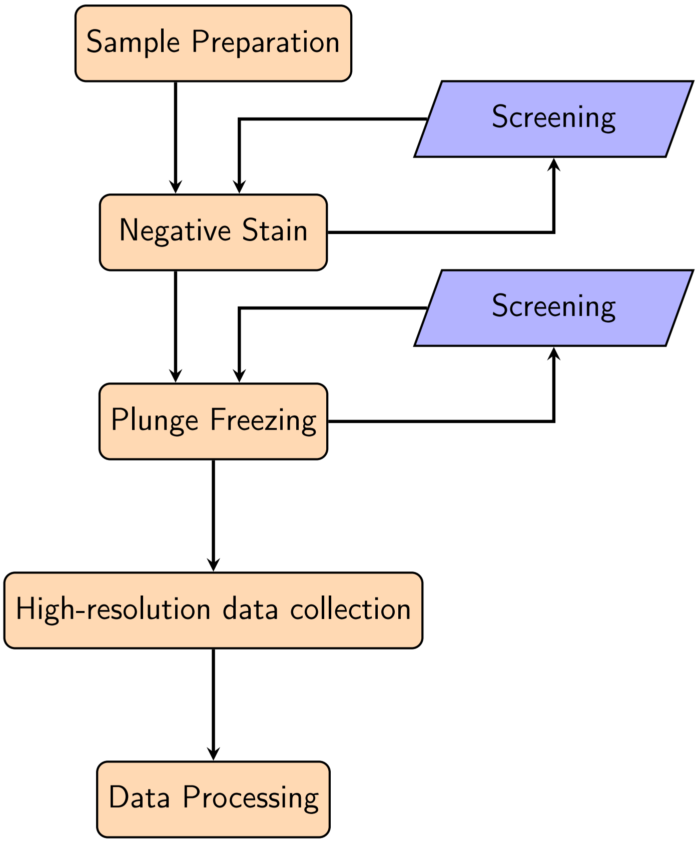


Figure 3.3 – Cryo-EM workflow

approach is to run the complex over a gel filtration column, which would then ensure that the final sample being applied to the grid would be the entire complex with the enzyme, transcription factor and nucleic acids bound. A 24ml Superose6 Increase 10/300 GL column was used for size exclusion chromatography. For each run, 100µl of sample was injected onto the column. A sample containing 1 mM DTT, 100 µM scaffold, 50 µM RNAP and 250 µM GreA in Cryo-EM buffer was injected onto the column equilibrated in cryo-EM buffer. The column was then washed until the sample eluted. Peak fractions were pooled and concentrated, and then loaded onto SDS-PAGE and denaturing urea gels to check for the presence of individual complex components.

To be able to visualise biological samples in their native states using cryo-EM, they need to be rapidly frozen in order to prevent the formation of ice crystals and suspend the particles in vitreous ice (an amorphous solid) (Dubochet and McDowall 1981). To achieve this, samples applied to cryo-EM grids are blotted and then plunge-frozen in liquid ethane (melting point = 85 K).

In order to prevent preferential orientation of *E. coli* RNAP particles, all frozen samples contained the detergent CHAPSO[1] at its CMC of 8mM (Chen et al. 2019). Grids were made hydrophillic through plasma treatment using a Fischione plasma cleaner (model 1070). The grids were then used immediately for plunge freezing. Samples were frozen using a Vitrobot Mark IV (Thermo Fisher Scientific). All samples collected on were frozen with the same parameters:

Plasma Cleaning (Fischione 1070)	
Gas mix	9:1 Ar:O_2
Time	40 s
Power	70%

Plunge Freezing (Vitrobot Mark IV)	
Sample volume	3 µl
Blot time	2 s
Blot force	8

All samples were either frozen on UltrAuFoil 1.2/1.3 300 Au mesh grids, or on QuantiFoil 2/2 300 Cu/Rh mesh grids. Grids were immediately stored in dedicated cryo-EM grid boxes in liquid nitrogen. Samples applied to all grids contained the same 1:2:5 ratio of RNAP:scaffold:GreA in Cryo-EM buffer containing 1 mM DTT and 8 mM Chapso. The samples applied to the gold foil (UltrAuFoil) and carbon foil (QuantiFoil) grids contained 4 mg/ml RNAP. Different sample concentrations were only used when applied to graphene oxide (GO)-coated grids, since the higher affinity of biomolecules for graphene meant that a lower sample concentration would have to be used.

Graphene Oxide Grids

GO grids were prepared as per the protocol in the paper from Bokori-Brown et al. (2016). A 2 mg/ml stock solution of GO flakes was diluted to 0.2 mg/ml in water. Holey carbon foil grids were plasma cleaned using the same parameters as before, and a 3 µl drop was applied to the C-side of each grid. The solution was incubated on each grid for 1 minute before washing. The grids were washed three times by lifting 20 µl drops of water and blotting them edge-on on Whatman blotting paper, twice on the C-side and once on the

[1] 3-([3-Cholamidopropyl]dimethylammonio)-2-hydroxy-1-propanesulfonate

mesh side. The grids were air dried and used within 30 minutes.

Single Particle Cryo-EM

Prior to collecting any dataset, the grids were first screened for particle density, ice thickness and contamination. Screening of grids was initially carried out on a Polara (FEI) within the IGBMC, and later on the Glacios (Thermo Fischer Scientific) at Novalix, situated on the Illkirch Campus. Grids screened on the Glacios could be directly loaded onto a Titan Krios (Thermo Fisher Scientific), owing to the two microscopes using the same clipping mechanisms and docking cassettes. Screened grids were first used to collect preliminary datasets in order to confirm whether or not the entire complex was intact, and without preferential orientation. Once a preliminary reconstruction confirmed that the sample contained an intact complex containing all the components, a high resolution dataset could be collected. For the data collection sessions at the IGBMC and at the EMBL, on-the-fly pre-processing and monitoring was carried out using WARP (Tegunov and Cramer 2019). The Warp-extracted particles and filtered good movies were used for further processing. Frame alignments and dose weighting of the movies was done using either Motioncor2 (Zheng et al. 2017; X. Li et al. 2013) or the patch motion correction algorithm in CryoSPARC (Punjani et al. 2017). This was followed by CTF estimation by Ctffind4 (Rohou and Grigorieff 2015). Different particle extraction methods were used for different datasets, and have been indicated accordingly in Tables 5.1 and 6.1 listing the data collection parameters. For particles picked with Topaz (Bepler et al. 2019), the Warp-extracted particles were used to train the model. Once the particles had been cleaned up by 2D class averaging, *ab initio* models were generated, which were then used as the input volumes for generating the reconstructions. The fourier shell correlation (FSC) plots were used to determine the resolutions of maps. 3D classification was finally used to check for different classes of particles and to clean up the dataset.

Published *E. coli* RNAP structures were used as references when building the atomic models in Coot (Emsley et al., n.d.), namely the *E. coli* RNAP elongation complex (6ALH) from Kang et al. (2017) and the GreB-bound precleavage complex (6RIN) from

Abdelkareem et al. (2019). The *E. coli* GreA crystal structure from Stebbins et al. (1995) (1GRJ) was used in the GreA pre-cleavage models. Refinements of the models was carried out in Phenix (Adams et al. 2010; Liebschner et al. 2019), with model-based map sharpening done using LocScale from Jakobi, Wilmanns, and Sachse (2017). Post-processing of maps was also carried out with DeepEMhancer (Sanchez-Garcia et al. 2020). Analysis of the data was carried out in Coot and ChimeraX (Pettersen et al. 2021; Goddard et al. 2018).

Results

Chapter 4

Complex Formation and Preliminary Results

I first set out to obtain a functional, homogeneous complex of a backtracked RNA Polymerase (RNAP) with GreA bound to the secondary channel (SC). The aim was to be able and capture it in its pre-catalytic stage - i.e. before RNA cleavage. This started off with carrying out *in vitro* transcription assays to assess the functionality of complexes assembled with purified proteins. The next step was to identify and optimize the best DNA/RNA scaffold for assembling a homogeneous pre-cleavage complex. Finally, it was important to collect a preliminary dataset of the assembled complex before moving onto a high-resolution dataset.

4.1 Functional Characterisation

The activity of RNAP and GreA purified as per the protocol in 3.1 had to first be confirmed to make sure the complexes assembled thereafter would be functional. Initial transcription assays used to test (a) the functionality of the assembled complex, and (b) the first set of DNA-RNA scaffolds are shown in Figure 4.1. The RNA used in both scaffolds was the same, with the template and non-template DNA strands designed accordingly. As seen in the scaffold schematics in the figure, both complexes were backtracked by the same length of RNA (3 nucleotides). The difference between the two was the size of the

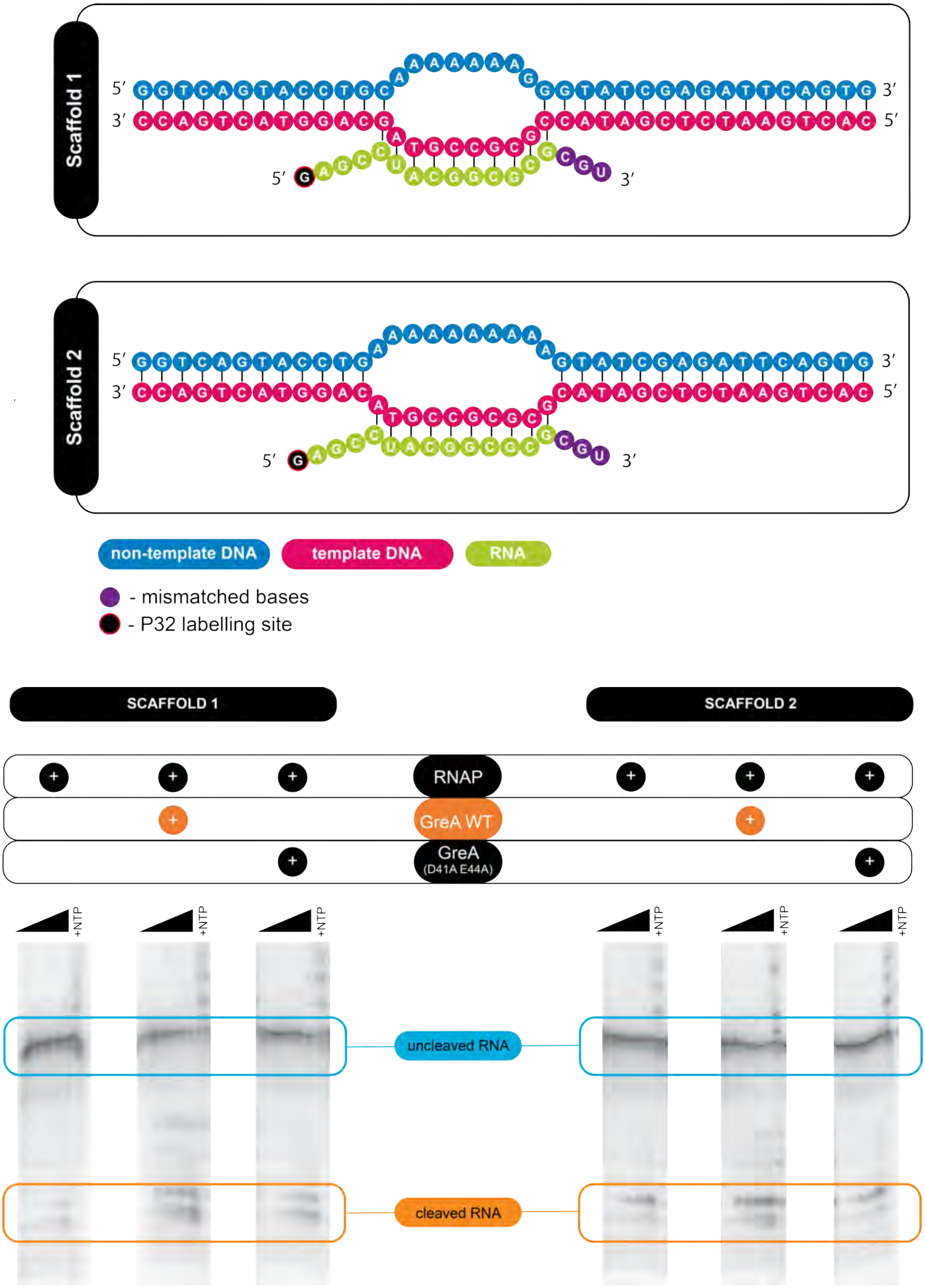


Figure 4.1 – Results of the first set of transcription assays, which were used to assess the functionality of the assembled complex as well as to start testing out scaffolds that would yield a single homogeneous product. The time course used in these was 0/1’/5’/10’/30’, followed by a chase of NTP (A+C+G+U).

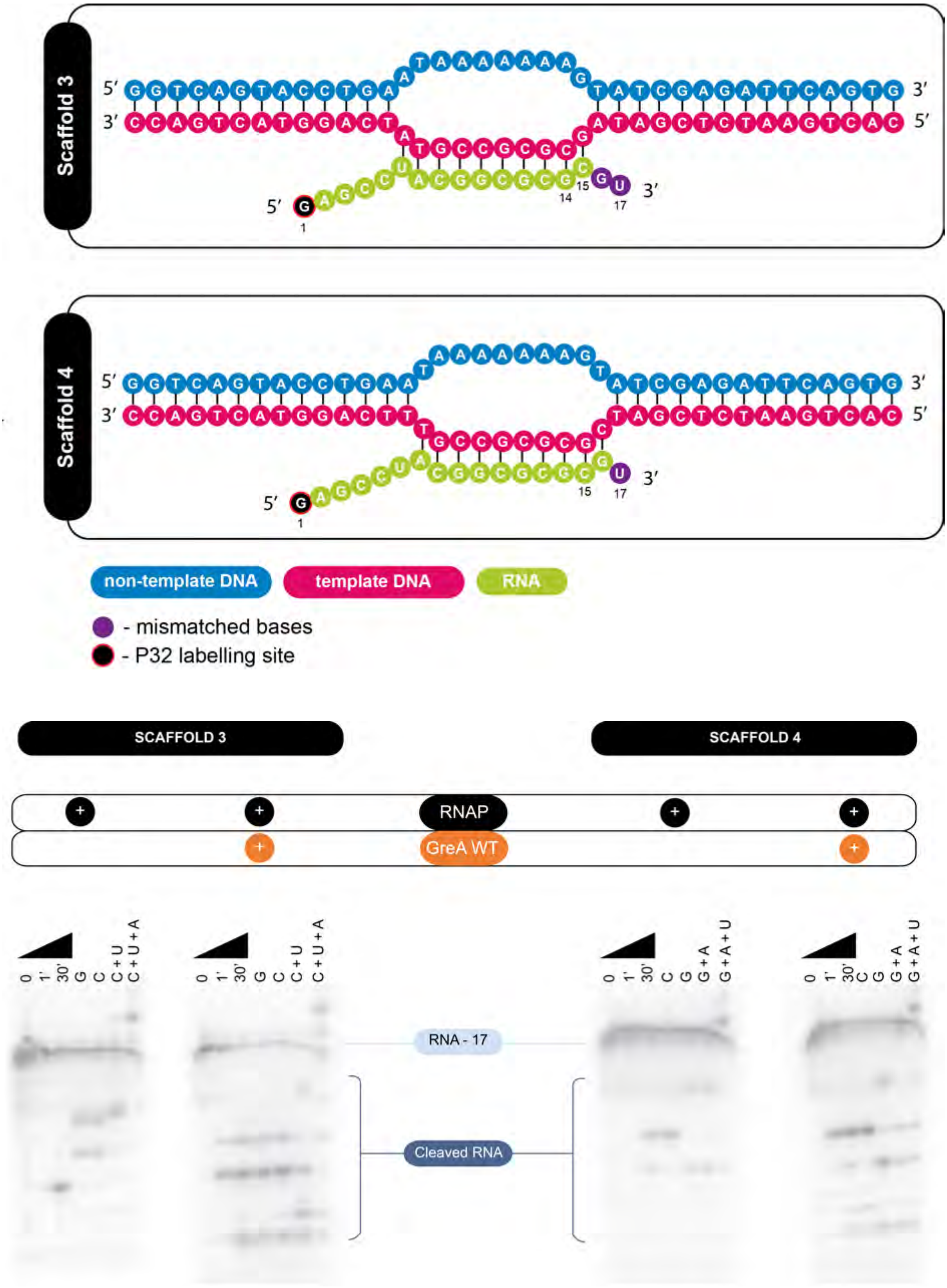


Figure 4.2 – Assays to test homogeneity of tri- versus di-nucleotide cleavage, with RNA positions 1, 15, and 17 labelled in the scaffold diagrams. Uncleaved RNA is 17 nt-long.

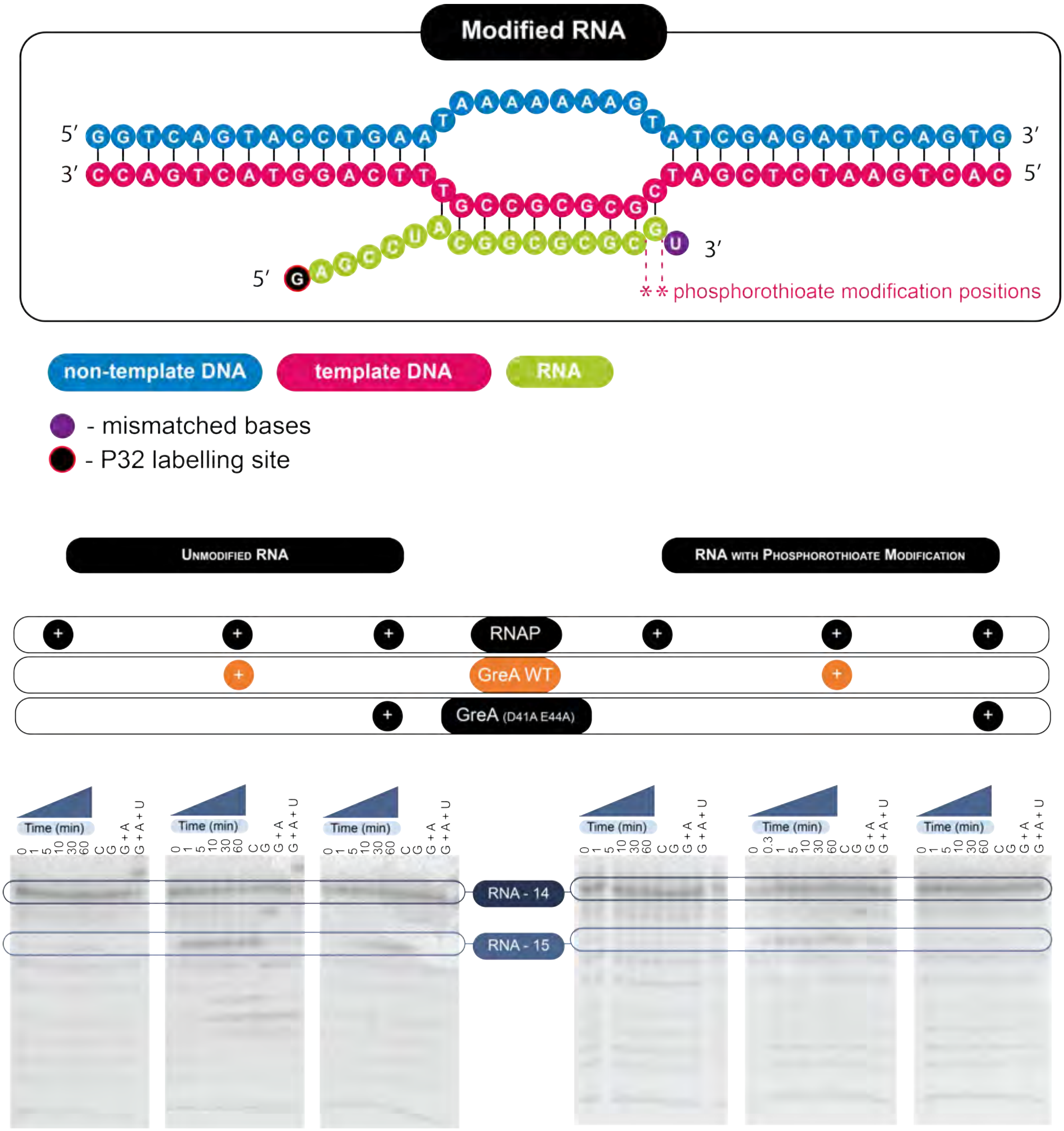

Modified RNA
5′ GGTCAGTACCTGAA T AAAAAAAG T ATCGAGATTCAGTG 3′
3′ CCAGTCATGGACTT T GCCGCGCG C TAGCTCTAAGTCAC 5′
5′ G AGCCUA CGGCGCGC G U 3′
* * phosphorothioate modification positions
non-template DNA
template DNA
RNA
- mismatched bases
- P32 labelling site
Unmodified RNA
RNA with Phosphorothioate Modification
RNAP
GreA WT
GreA (D41A E44A)
+
Time (min)
0 1 5 10 30 60 C G G + A G + A + U
0 0.3 1 5 10 30 60 C G G + A G + A + U
RNA - 14
RNA - 15

Figure 4.3 – Assays to compare scaffolds with and without RNA modifications

transcription bubble –an 8 base pair central mismatch in the first, and a 10 base pair mismatch in the second. Cleavage reactions were tested under three conditions:

- RNAP alone
- RNAP + wild-type GreA (WT GreA)
- RNAP + mutant GreA (D41A E44A)

The gels displayed in Figures 4.1 and 4.2 were 15% polyacrylamide (AA) - 7M urea. They resulted in blurrier bands when compared with the sharper bands seen in 20% AA gels run with the same samples. However, the separation between the bands was a lot more distinct. Nevertheless, beginning with Figure 4.3, the denaturing gels used to run all samples on were 20% AA, in which preparation and running conditions were adjusted to obtain more uniform bands with clear separation of the RNA in the samples.

For both scaffolds 1 and 2, we see that the RNA does not get completely cleaved over the course of the reaction (30 minutes). For the fraction that does get cleaved, the reaction seems to be more efficient in the presence of WT GreA. In the case of the reaction containing the GreA mutant, cleavage is not completely identical to what is observed for RNAP alone, but is less efficient than when assisted by WT GreA. For both, two cleavage products were obtained. However, a higher proportion of the longer cleavage product (the upper band, resulting from cleavage of fewer nucleotides) was present in the second reaction set. Although a larger bubble would likely allow for more flexibility in binding, the size of the bubble might have led to a preference in cleavage of one bond over the other, which was not seen in the case of the shorter transcription bubble in Scaffold 1.

After this, two more scaffolds were designed, both with a 10 nucleotide central mismatch and a 17 nucleotide-long RNA (scaffold schematics in Figure 4.2). The reactions were carried out with and without WT GreA, and a time course of 0/1'/30' was used. The main aim of these reactions was to test the cleavage positions for scaffolds with shorter backtracked RNA. As with the first two sets of reactions, the scaffolds here were also designed with the same RNA, but with different template and non-template DNA

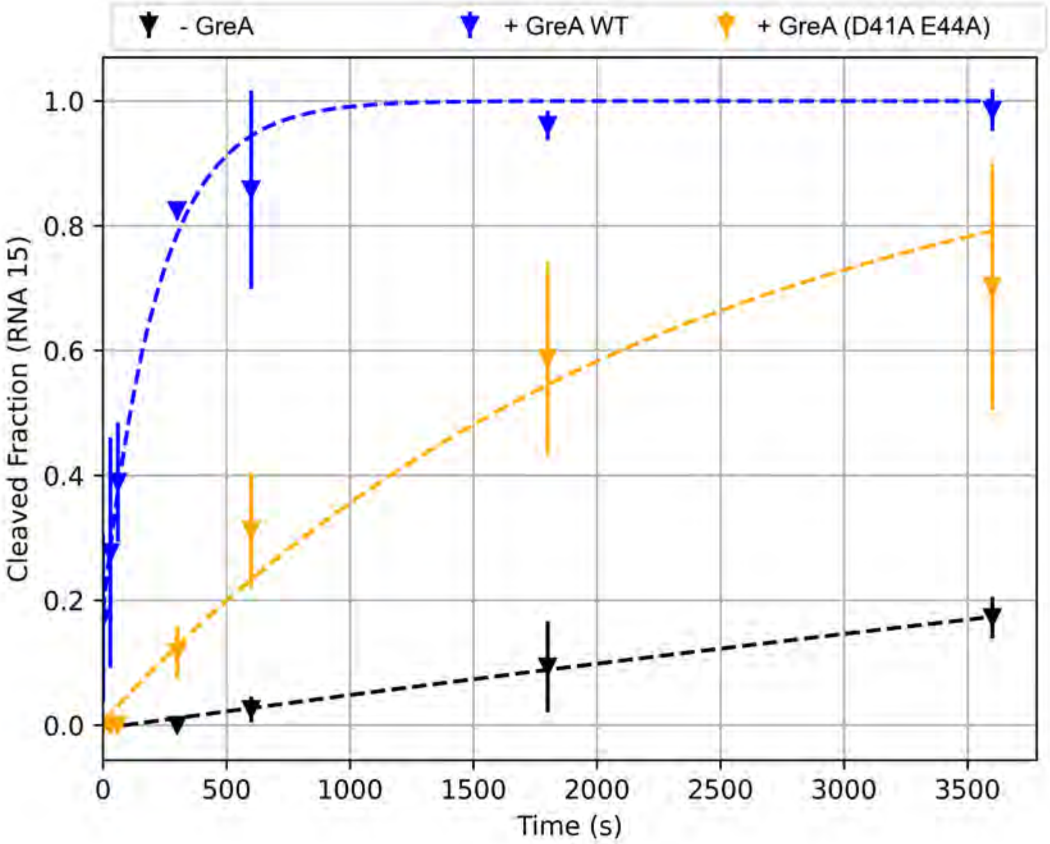


Figure 4.4 – Graph of cleavage fraction versus time for different conditions: RNAP alone, RNAP + WT GreA, and RNAP + mutant GreA. The half life for each fitted curve are (top to bottom) 2.6, 26.6, and 212.8 minutes.

oligonucleotides, which would form a 2 nucleotide (scaffold 3) and a 1 nucleotide (scaffold 4) backtrack. Instead of chasing the reaction with a G+C+A+U NTP mix, the chase was carried out with different combinations of NTPs, indicated in Figure 4.2.

For scaffolds 3 and 4, I again saw that the amount of cleavage was significantly increased in the presence of WT GreA. For these two sets of reactions, the complex with the longer backtracked (scaffold 3) was expected to result in cleaved RNA that was 14 nt long, with the shorter backtracked scaffold (scaffold 4) forming a 15 nt cleavage product. For scaffold 3, addition of only GTP would not provide the complex with a substrate complementary to the DNA in the acceptor site (A-site), thereby not favouring RNA extension. Addition of CTP alone would extend the RNA by 1 nt, CTP+UTP would extend it by 2 nt, and addition of CTP+ATP+UTP would extend it by upto 4 nt. Likewise with scaffold 4, CTP alone was expected to produce no extension, with the subsequent chases (GTP,

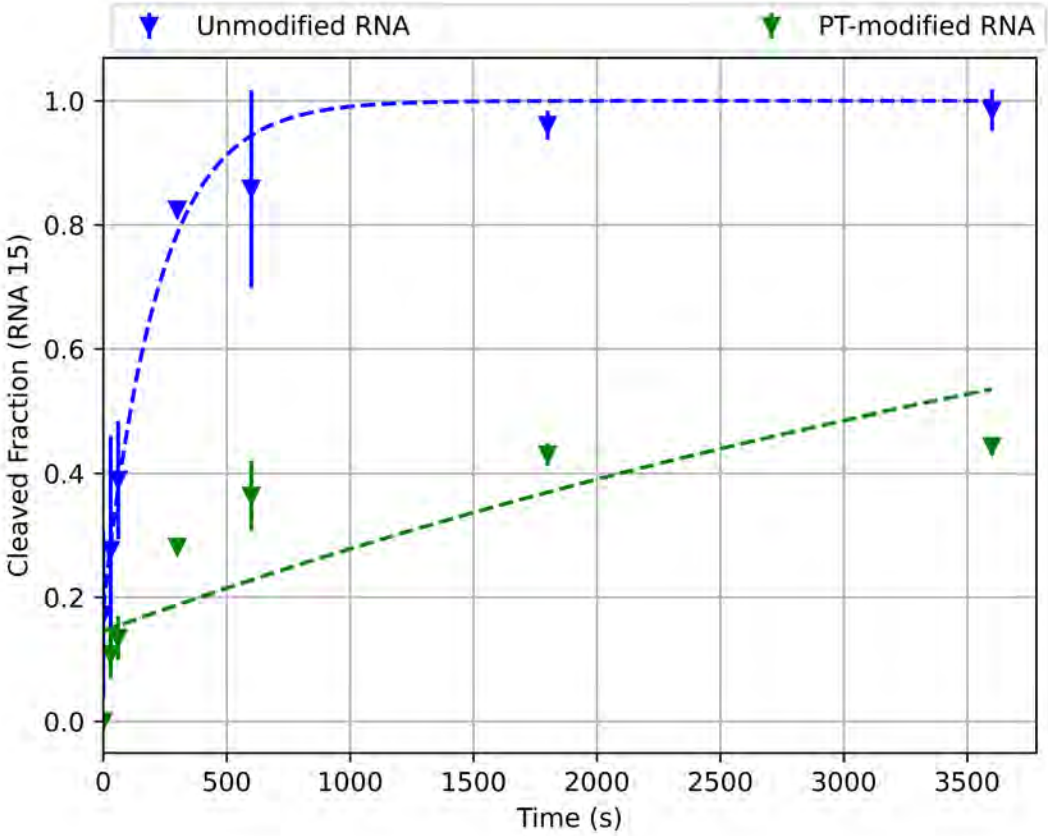


Figure 4.5 – Graph comparing the di-nucleotide cleavage of 1 nucleotide backtracked RNA with and without phosphorothioate modifications in the presence of GreA

GTP+ATP, GTP+ATP+UTP) forming gradually longer extension products. However, the results showed that both the cleavage products and extension products in the chase were very heterogeneous for scaffold 3, with more homogeneity observed in scaffold 4. Scaffold 4 (with a 1nt backtrack) was chosen between the two, and the cleavage assay repeated, as shown in the gels on the left in Figure 4.2.

In all four scaffolds tested in the beginning, it was clear that cleavage in the presence of WT GreA happened rapidly, with a large quantity of cleaved product formed within the first minute. Cleavage in the presence of the GreA mutant was comparatively slower. However, since the aim was to be able to use the WT form of GreA for the cryo-EM reconstructions, using the mutant form of the protein was not the first course of action. Instead, I tested RNA containing phosphorothioate modifications. An oligonucleotide with a phosphorothioate modification has a non-bridging oxygen in its phosphate backbone substituted for a sulphur atom. This wouldn't alter the actual structure of the oligonucleotide. However, sulphur has a lower affinity for magnesium (Mg^{2+}) (Pecoraro, Hermes, and Cleland 1984). Therefore RNA containing this modification at its 3'-end would undergo hydrolytic cleavage at a much slower rate.

An RNA with two phosphorothioate modifications at the 3'-end (indicated by the two red asterisks in the scaffold schematic in Figure 4.3) was used within a 1 nt backtracked complex and its affect on cleavage tested in comparison with the same complex with an unmodified RNA. The two scaffolds were tested in three complexes: RNAP alone, with WT GreA, and with the mutant GreA. As was previously seen, cleavage of the unmodified RNA was observed within the first minute in the presence of WT GreA. In the case of the unmodified RNA, almost no cleavage was observed for the complexes of RNAP alone and RNAP with the GreA mutant. In the presence of the WT, cleavage did occur as sulphur only reduces the affinity for the Mg^{2+} required for cleavage, but doesn't inhibit it. The cleavage rate was sufficiently reduced to be confident that I would be able to capture the complex in its pre-catalytic state for cryo-EM.

4.2 Cryo-EM

Once it had been established that a functional complex backtracked by 1 nt could be formed in which the rate of cleavage could be slowed down enough to capture it in its pre-catalytic state, the next stage was to move on to cryo-EM.

The very first step at this stage was to establish whether the complex could be purified by size exclusion chromatography prior to freezing it on an EM grid. Within single-particle cryo-EM, the best strategy is to use a sample that is as homogeneous as possible. While the more recent advances in processing algorithms has made it easier to deal with more heterogeneous samples, the preference is to always start from a point that is as close to homogeneity as possible. Size exclusion chromatography was performed using a sample containing RNAP, DNA-RNA scaffold and WT GreA in a ratio of 1:2:5, as per the protocol in Section 3.3.

Injecting a sample with an RNAP concentration of ∼4 mg/ml (10 µM) resulted in a peak that didn't contain any clear band for GreA. Increasing the RNAP concentration in the sample five times to ∼20 mg/ml (50 µM) resulted in a faint band for GreA in the main peak (Figure 4.6). Although present in the complex after injection, the amount of GreA was sub-stoichiometric, evident when comparing the intensity of the bands for GreA and the ω subunit of RNAP. This showed that although some of the GreA stayed bound after passing the sample over a gel filtration (GF) column, not enough of the entire complex would be intact in order to get a reliably large fraction of the GreA-bound complex for single-particle analysis (SPA). Looking at the results from the GF runs and the cleavage assays, I decided that the best strategy for SPA would be to mix the purified comlex components on the bench prior to freezing, with WT GreA added at the very end just before application to the grids.

From the beginning, there were two complexes that I aimed to collect on:

- 1 nt backtracked complex without GreA (backtracked complex)
- 1nt backtracked complex with GreA, its pre-cleavage state (pre-cleavage complex)

In addition to this, I also wanted to try alternative supports for cryo-EM, since up until that point all of the datasets collected within the lab had been done on samples frozen on either C-flat grids or Quantifoil grids (both with holey carbon foils). I wanted to test graphene oxide (GO)-coated grids since they are useful supports for lower concentration samples (Palovcak et al. 2018), as well as gold foil grids as they are meant to be more stale within the electron beam (Christopher J Russo and Lori A Passmore 2016a, 2016b). For I decided to test the backtracked complex on GO-coated grids, and the pre-cleavage complex on gold foil grids (UltrAuFoil).

For the GO grids, the optimum sample concentration had to first be determined. Graphene has a very high affinity for bio-molecules, and therefore the sample concentrations used for these grids would have to be significantly lower than the normal 4 mg/ml used for standard holey foil grids. Examples of some of the screening of the GO grids are shown in Figure 4.7. It can be seen that even at 0.1 mg/ml, the sample forms a dense carpet on the entire sheet. Screening for GO conditions was done on an FEI Polara, which

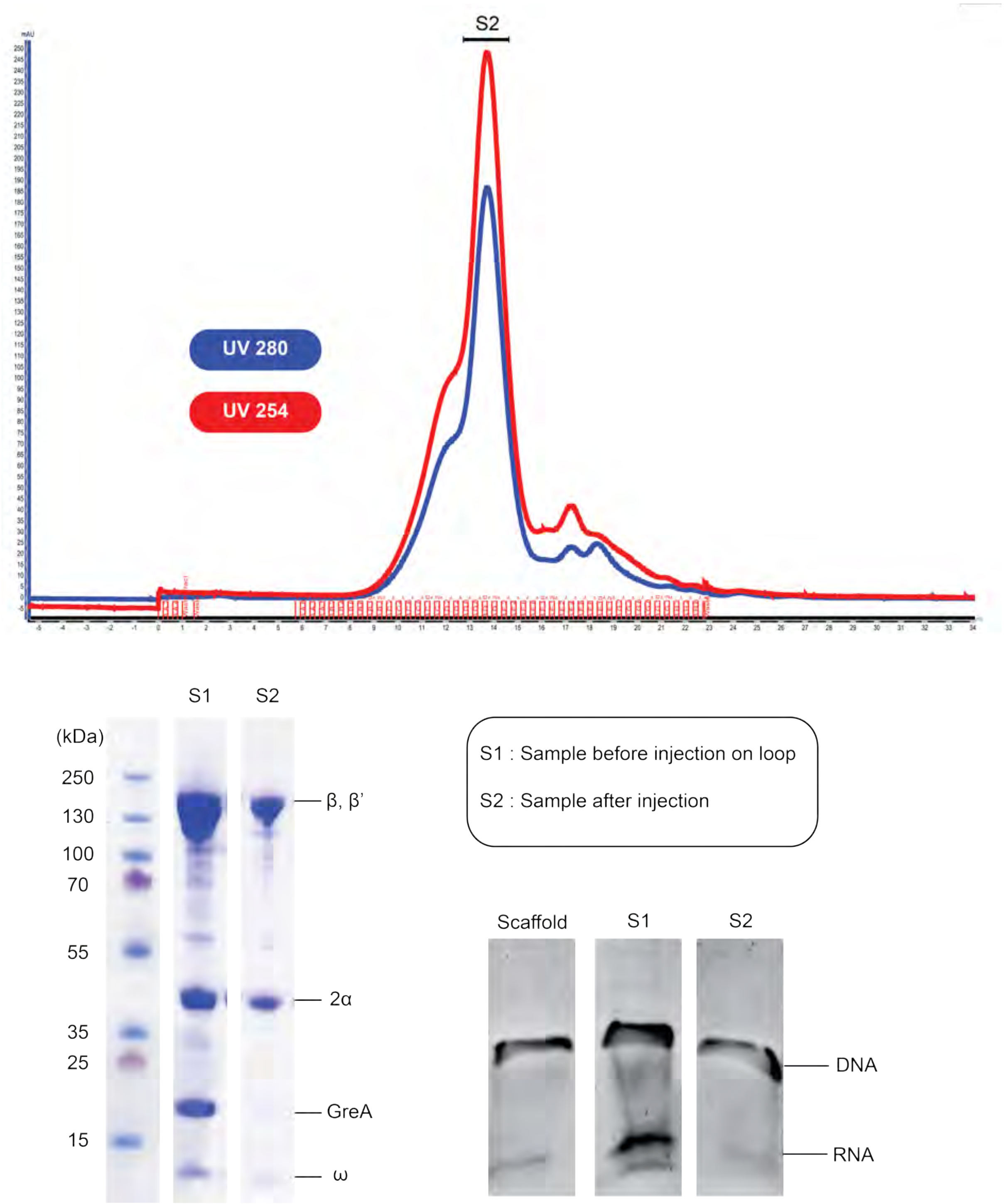


Figure 4.6 – *Top*: Size exclusion chromatogram of the entire assembled complex (x-axis = mAU, y-axis = collection fractions). *Bottom left*: SDS-PAGE columns of the protein ladder, the injected sample and the sample contained in the central peak on the chromatogram. *Bottom right*: Urea gel showing nucleic acid components within the injected sample and the chromatogram main peak, with the complex scaffold used as a reference.

operates at 100 keV. This resulted in better contrast during the screening, allowing us to easily spot the squares and holes that were covered with GO. This turned out to be a challenge during the acquisition of a larger dataset on the Titan Krois, as the lower contrast meant that identifying squares that were covered with GO flakes was extremely difficult. In datasets for which the results are shown in Figure 4.8, the total number of micrographs collected was approximately 3000. On inspection of the micrographs, most of them appeared to have been collected in holes that were empty because they lacked GO sheets. Therefore, the entire dataset had to be manually inspected to pick good micrographs, which in the end was a little over 300. After loading and processing the selected micrographs in CryoSPARC2, the final reconstruction showed very clear and strong preferential orientation.

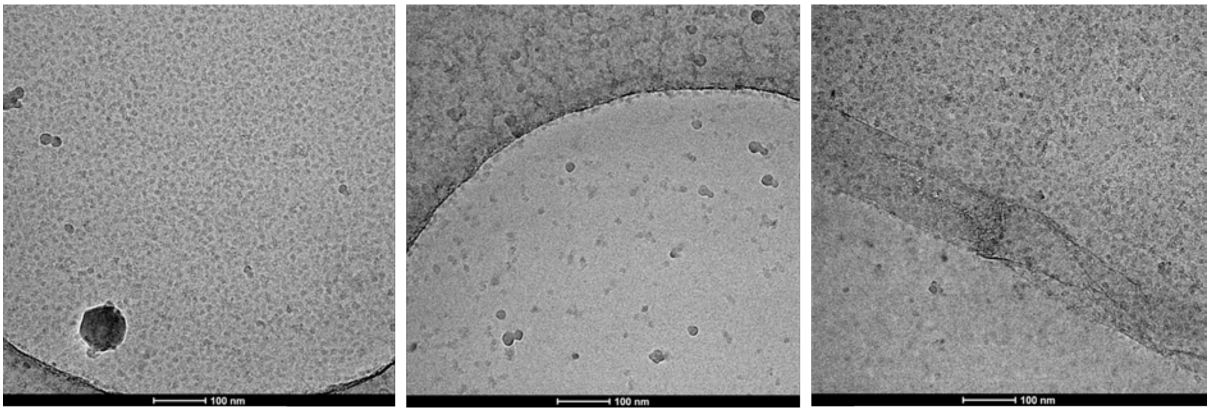


Figure 4.7 – Images of holes within graphene oxide-coated grids. The images on the left shows a hole covered with GO, with a sample concentration of 1 mg/ml RNAP. The hole in the centre is from the same grid, but without GO and hence no particles. The image on the right is from a partially covered hole on a different grid with a sample concentration of 0.1 mg/ml.

Obtaining a structure of the pre-cleavage complex remained the main objective. Optimizing particle orientation on the GO grids for the backtracked complex would likely include extensive screening for detergents and I decided to use the regular holey carbon/gold grids, for which the freezing conditions had already been optimized. Screening of the pre-cleavage complex on gold grids was also carried out on the Polara, in which the particle distribution for an RNAP sample concentration of 4 mg/ml was ideal, with almost no contamination seen inside the holes (atlas and particle distribution in Figure 4.9). Two test datasets were collected for this complex on gold grids, one in November

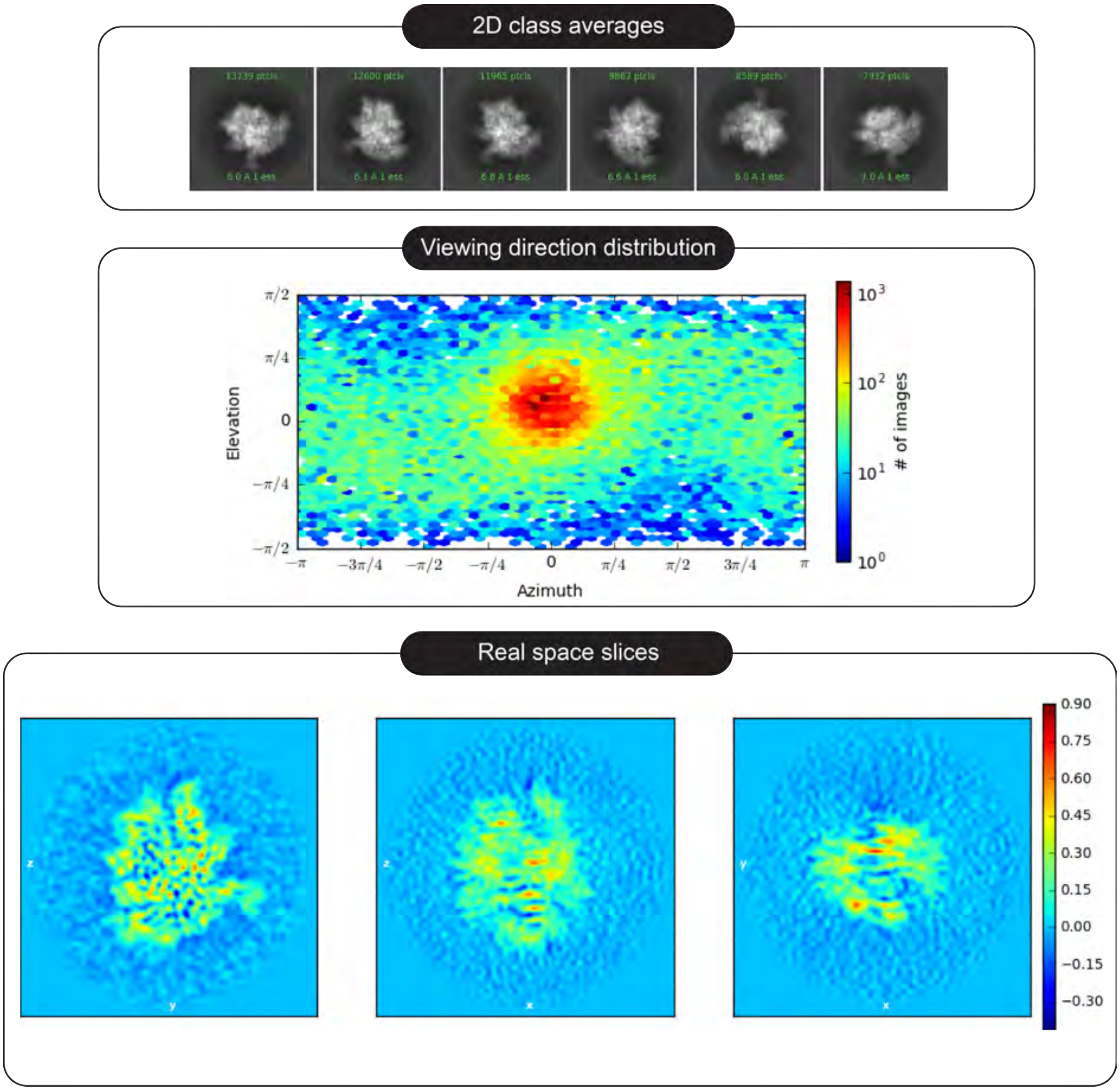


Figure 4.8 – Results for first data collection in September 2018: The 2D class averages after cleanup showed a preference for one orientation of the RNAP complex, which was reflected in the viewing direction distribution plot and real space slices (streaking artefacts) generated from a refinement job containing all of the "clean particles".

2018 and the other in December 2018. The results from November are shown in Figure 4.9. This dataset was collected on the Titan Krios at the IGBMC, with the Falcon 2 detector. The obtained resolution was poor, and not high enough to determine any actual features of the complex. The dataset from December 2018 could not be used due to problems with the objective aperture of the microscope.

After these initial issues collecting on gold grids, I decided that for the next data collection slot, the safest strategy would be to go back and use carbon foil grids, at least so that a preliminary reconstruction could be obtained. Samples were frozen on Quantifoil 2/2 Cu/Rh holey carbon foil grids, and a preliminary dataset was collected in January 2019. Results for this dataset are shown in Figure 4.10, which reached an FSC resultion of 4.2 Å. More important than the resolution was that I could clearly see GreA in the SC, coloured in orange in the figure. The resolution of the refined map was good enough to also see the first magnesium ion bound in the A-site, Mg-I.

With these results, I saw that even with a phosphorothioate modification in the RNA, cleavage in the presence of WT GreA took place too rapidly for size exclusion chromatography to be carried out prior to grid freezing without the complex dissociating. Still, due to the presence of the modifications in the RNA, the time window in which the complex could be captured in a pre-catalytic state was sufficient enough to mix the purified components on the bench before adding GreA and applying the sample to a cryo-EM grid. GreA stays bound to the complex during freezing, as was seen with the preliminary reconstruction, and it would be possible to move on and collect a higher resolution dataset for the same complex, as well as for other pre-cleavage complexes with different backtracked bases.

For tests of the alternative EM grids, the use of GO-coated grids, while theoretically would be extremely advantageous for samples with low concentrations, was not a feasible strategy for high-resolution data collection. Although the preliminary datasets with the gold grids were unsuccessful, they were largely due to problems not concerning the grid itself. For many the high resolution datasets that followed, I typically prepared and screened both carbon and gold foil grids, choosing the best one for data collection based on the particle distribution and overall quality of the grid.

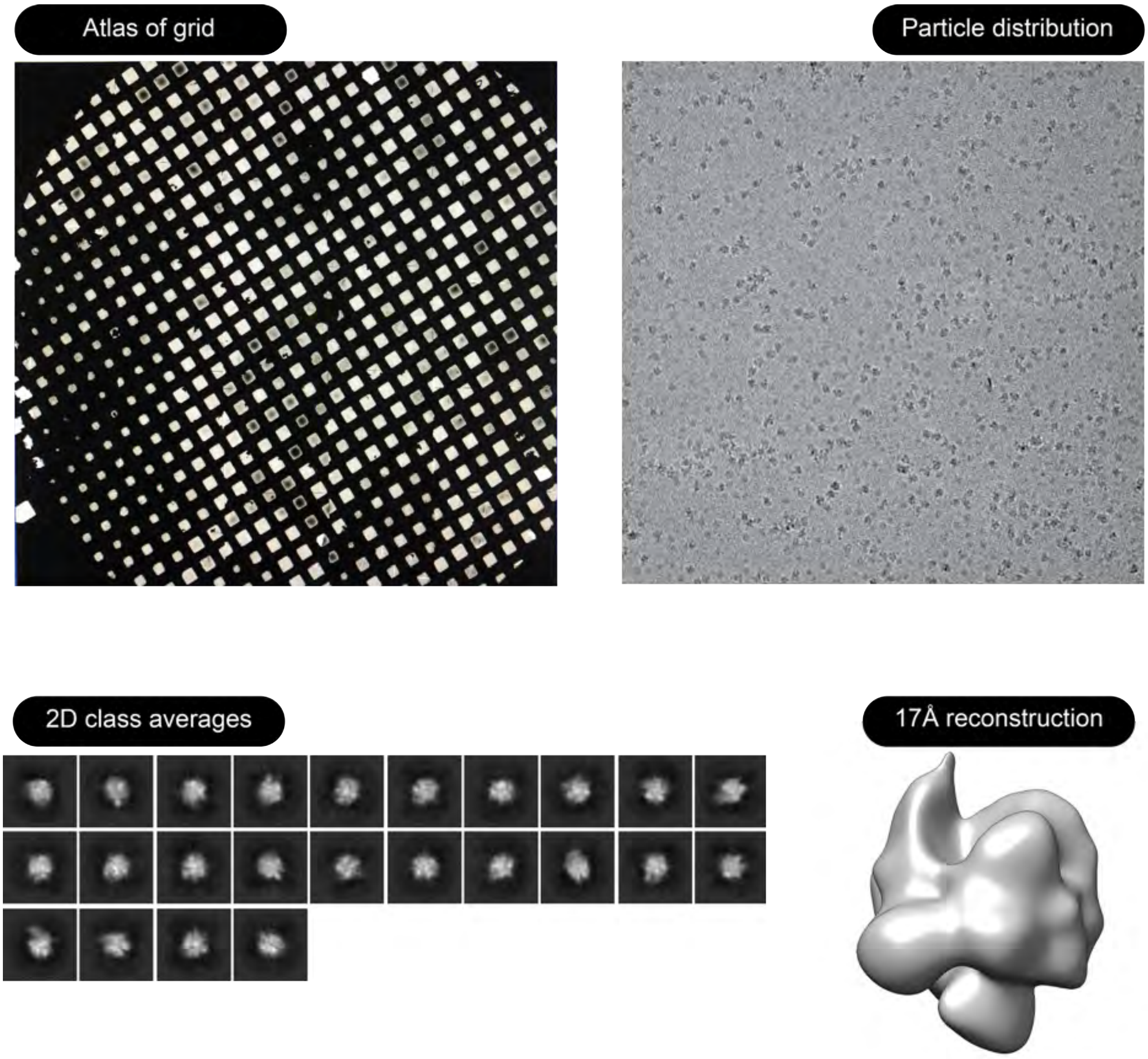


Figure 4.9 – November 2018 Titan data: The image of the EM grid atlas on the top left shows good ice thickness overall, with a representative image of the particle distribution within the grid holes on the top right showing good particle distribution and no contamination. However, analysis of the data yielded a reconstruction with poor resolution which could not be used for structural analysis.

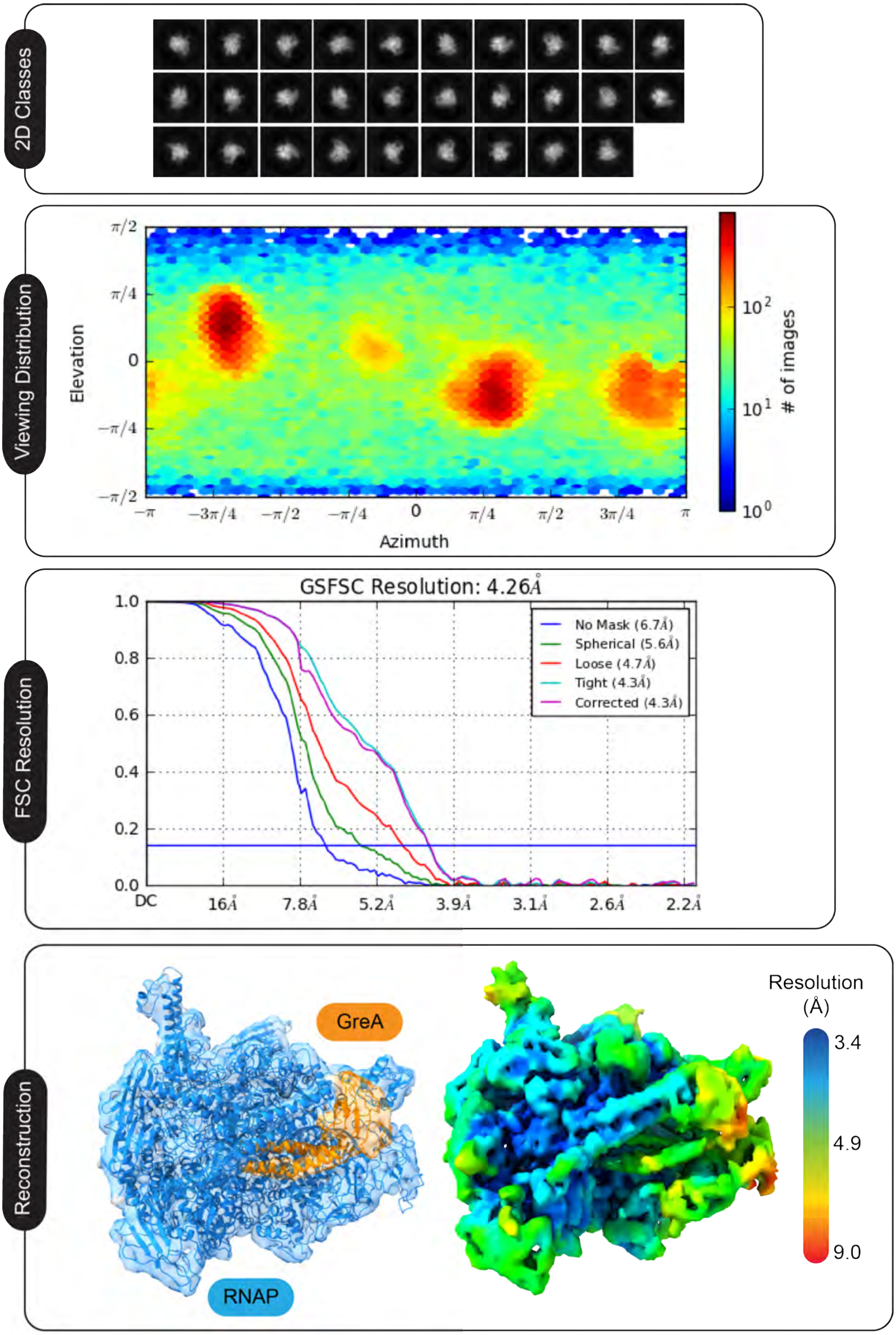


Figure 4.10 – Preliminary reconstruction of the GreA-bound pre-cleavage complex

Chapter 5

Pre-Cleavage Complexes

The preliminary results described in the previous chapter showed that despite the challenge of trying to capture a complex of a backtracked transcription elongation complex (TEC) with GreA bound to the secondary channel (SC), it was possible to obtain the structure of the entire complex, and that with a larger dataset I would be able to reach higher resolution. A higher resolution reconstruction would allow us to get a better understanding of the internal mechanisms within the catalytic core during cleavage. To understand the possible effects of the nature of different backtracked bases on the same catalytic process, a total of four datasets were to be be collected on pre-cleavage complexes which would be near-identical in all regards except for the nature of the backtracked base.

5.1 Data Collection

The scaffolds used for each of these complexes are shown in Figure 5.1, which shows that for scaffolds containing the U, G, and C backtracks, the only differences were in the backtracked base. The scaffold with backtracked base A had an additional difference of one base pair of the DNA immediately downstream of the transcription bubble. Before freezing these complexes on cryo-EM grids, the complexes containing the new scaffolds (G, C, A) were first tested to make sure that they were functional and that there would be enough of a time window to freeze grids. Representative examples of these assays, which were run in triplicate, are shown in Figure 5.1. The gels shown in the figure are

	Pre-Cleavage			
	U	***A***	***G***	***C***
Grid	UltrAuFoil 1.2/1.3	Quantifoil 2/2	UltrAuFoil 1.2/1.3	UltrAuFoil 1.2/1.3
Particles Picked	425805 (CryoSPARC)	449502 (CryoSPARC)	195346 (Topaz)	334366 (WARP)
Particles Used in Reconstruction	212762	67208	124433	173700
Raw Micrographs	11976	4426	2478	1966
Pixel Size (Å)	0.8	1.0525	0.862	0.862
Defocus Range	-0.7 to -2	-0.8 to -2	-0.8 to -2	-1 to -2
Voltage (kV)	300	300	300	300
Electron Dose (e-/Å2)	42.6	56.1	48.6	49.76
Microscope	Titan Krios	Titan Krios	Titan Krios	Titan Krios
	EMBL (Heidelberg)	ESRF (Grenoble)	IGBMC	IGBMC
Detector	K2	K2	K2	K2

Table 5.1 – Cryo-EM data collection parameters used for the four high-resolution pre-cleavage complex datasets

for scaffolds containing phosphorothioate modifications in the same positions as shown in Figure 4.3. As with what was seen for the previous transcription assays, here I also saw that although the amount of cleavage in the presence of GreA was higher, the modifications to the phosphate backbone of the RNA also slowed down the reaction enough to capture each of these complexes in their pre-catalytic states while plunge freezing. For the sake of being concise, the different pre-cleavage complexes will be referred to as PC-U, PC-A, PC-G and PC-C for each of the different backtracked bases.

As per the standard practice expected for high resolution data collection, all grids used for data collection were screened prior to loading them on the microscope. An example of the screening process is shown in Figure 5.3. At this point, even though a carbon foil grid was used to collect the preliminary reconstruction in Section 4.2, the goal of wanting to use gold foil grids was still in place. The main reason for sticking with it was that consistently throughout screening complexes for data collection, the gold foil grids were consistently cleaner (less contamination within the holes, more uniform ice thickness

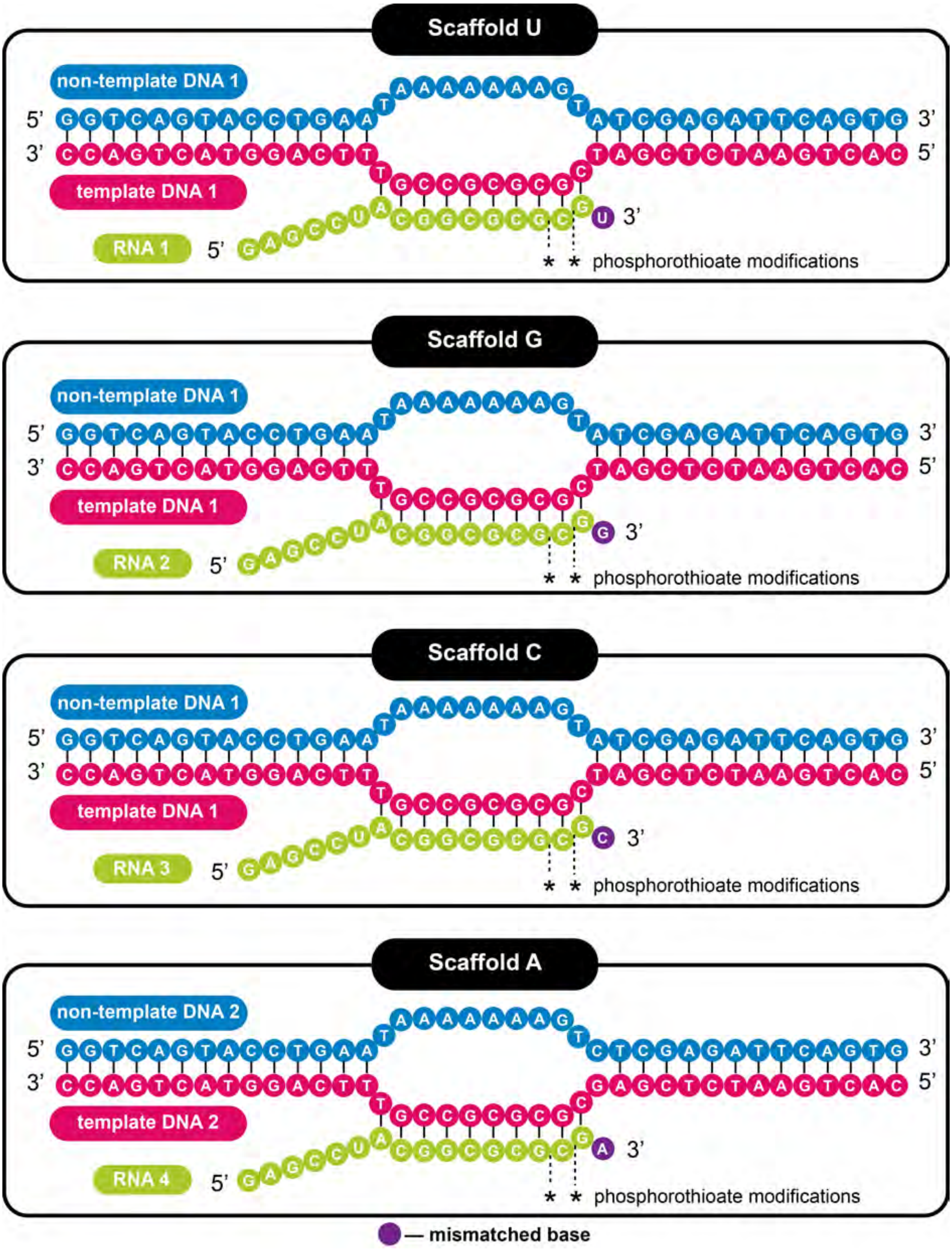

Scaffold U
non-template DNA 1
5' GGTCAGTACCTGAA T AAAAAAAG T ATCGAGATTCAGTG 3'
3' CCAGTCATGGACTT T GCCGCGCG C TAGCTCTAAGTCAC 5'
template DNA 1
RNA 1
5' GAGCCUA CGGCGCGC G U 3'
* * phosphorothioate modifications
Scaffold G
non-template DNA 1
5' GGTCAGTACCTGAA T AAAAAAAG T ATCGAGATTCAGTG 3'
3' CCAGTCATGGACTT T GCCGCGCG C TAGCTCTAAGTCAC 5'
template DNA 1
RNA 2
5' GAGCCUA CGGCGCGC G G 3'
* * phosphorothioate modifications
Scaffold C
non-template DNA 1
5' GGTCAGTACCTGAA T AAAAAAAG T ATCGAGATTCAGTG 3'
3' CCAGTCATGGACTT T GCCGCGCG C TAGCTCTAAGTCAC 5'
template DNA 1
RNA 3
5' GAGCCUA CGGCGCGC G C 3'
* * phosphorothioate modifications
Scaffold A
non-template DNA 2
5' GGTCAGTACCTGAA T AAAAAAAG T CTCGAGATTCAGTG 3'
3' CCAGTCATGGACTT T GCCGCGCG C GAGCTCTAAGTCAC 5'
template DNA 2
RNA 4
5' GAGCCUA CGGCGCGC G A 3'
* * phosphorothioate modifications
● — mismatched base

Figure 5.1 – Nucleic acid scaffold designs used for cleavage assays

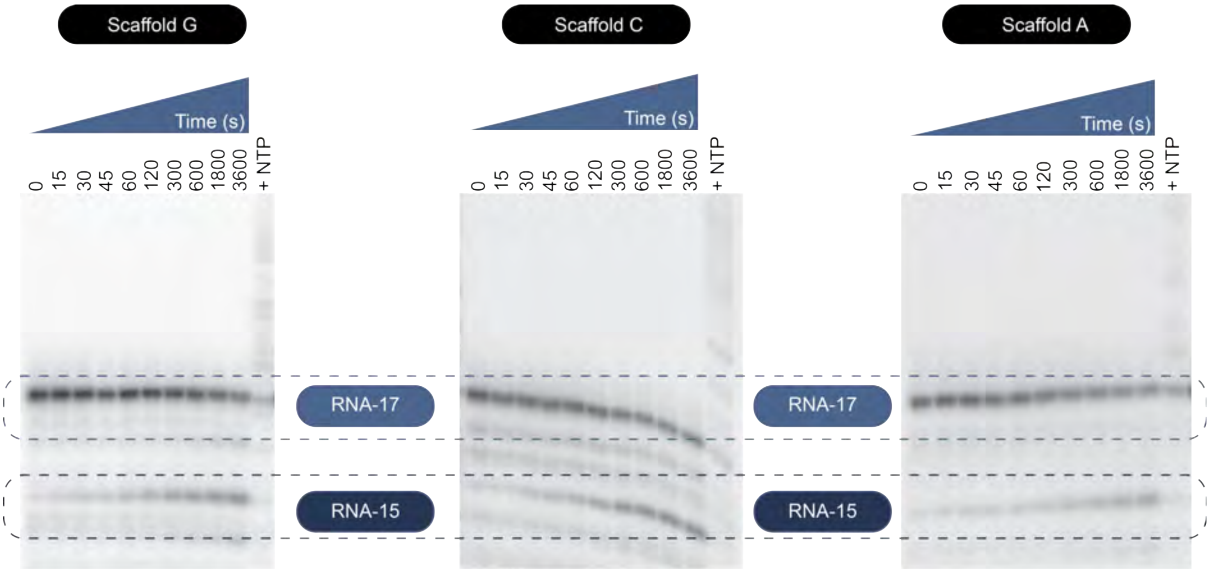


Figure 5.2 – Transcription assays to assess the functionality and GreA-assisted cleavage of remaining 1 nt backtracked complexes

throughout) than the carbon foil grids. The first two high resolution datasets of the four pre-cleavage complexes were for PC-A and PC-U, which were collected at the ESRF in Grenoble and at the EMBL in Heidelberg, respectively. Data collection parameters are listed in Table 5.1. Figures 5.4 and 5.5 show the results for these two reconstructions, which reached resolutions of 2.9 and 3.8 Å.

The final two datasets, for the PC-G and PC-C complexes, were collected in May 2020 over two separate overnight sessions. Despite having only about half of the number of micrographs collected for the PC-A dataset, these two reconstructions still reached reasonable resolutions of 3.9 and 4.2 Å. Atomic models were fit and refined for all four maps. On comparing the maps and models for the four pre-cleavage structures, it was evident that the overall structure and conformations of the core elements within the enzyme were the same. Therefore, the GreA pre-cleavage structure discussed in the rest of this chapter and in following chapters will refer to the PC-U complex, which reached the highest resolution of the four.

The first note for the pre-cleavage structures is that they all formed a single conformational class. This was similar to the GreB sructures in Abdelkareem et al. (2019), in which the GreB pre-cleavage complex was also present as a single class, while the 3 nt backtracked complex adopted two different conformational states termed 'swivelled'

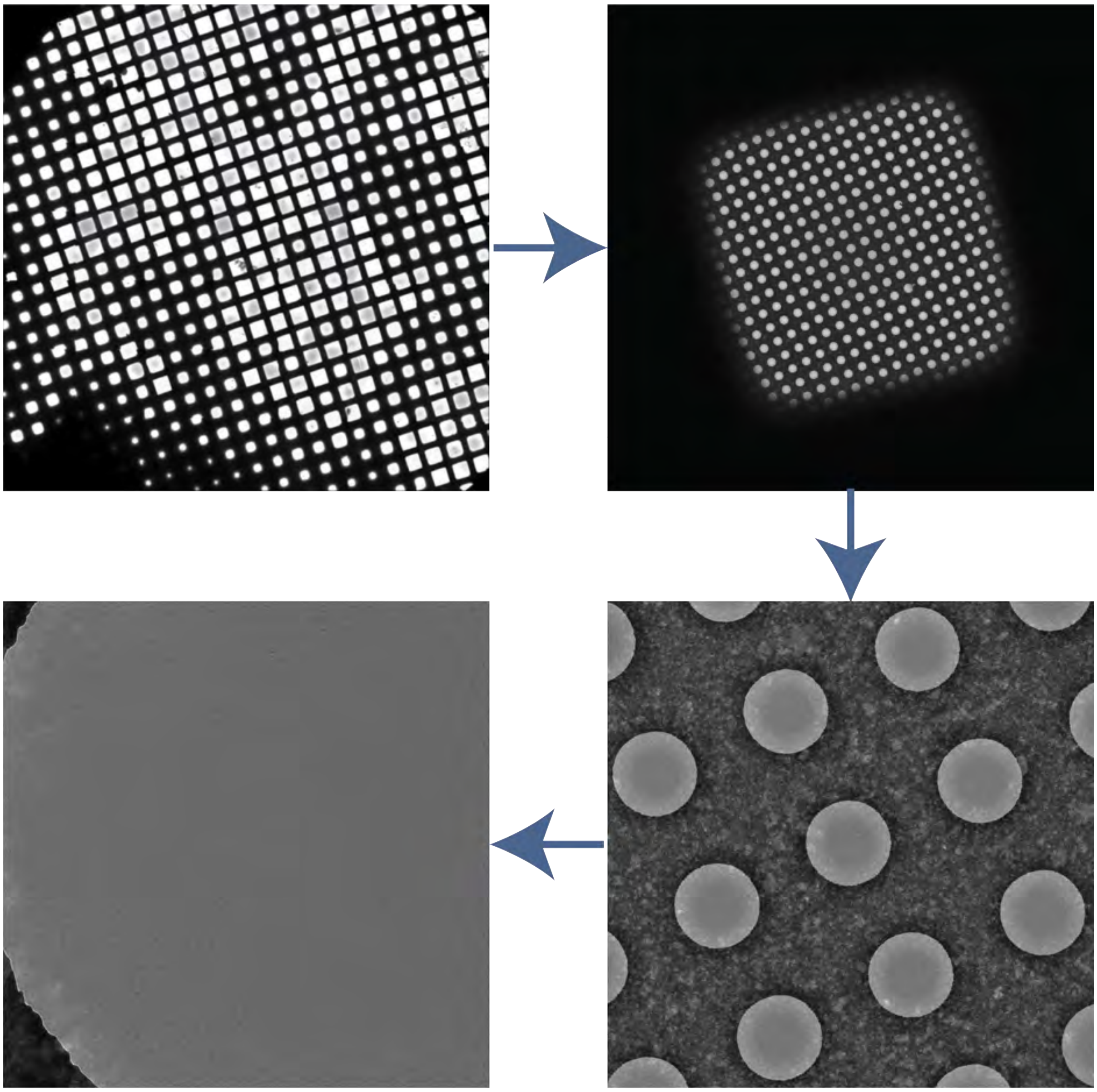


Figure 5.3 – Figures of the grid used for collection of the data shown in Figure 5.4. Clockwise starting from the top left: grid atlas, square map, shot over a hole, and an example of the particle distribution within a hole

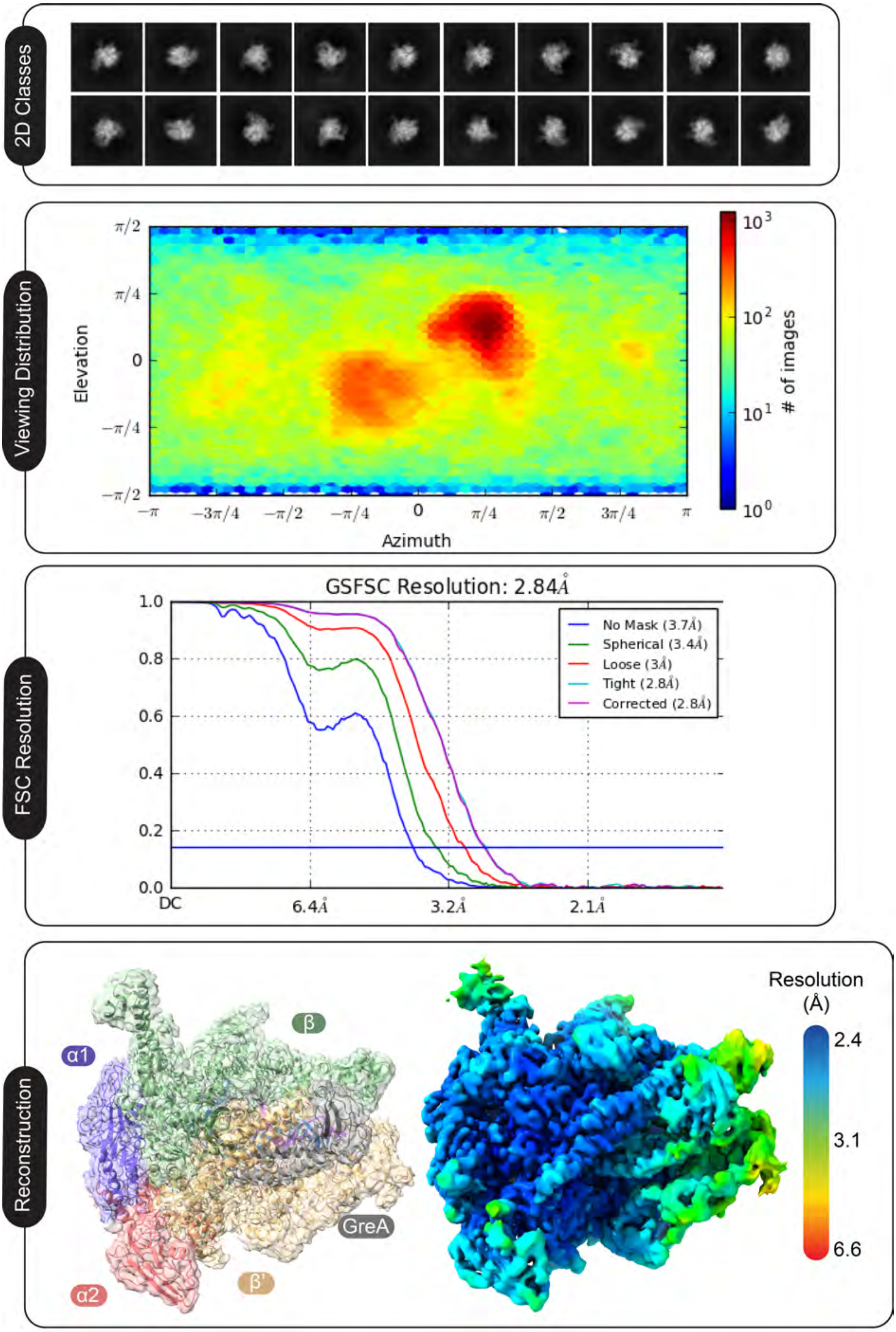


Figure 5.4 – Results of GreA-bound pre-cleavage complex with backtracked base ’U’

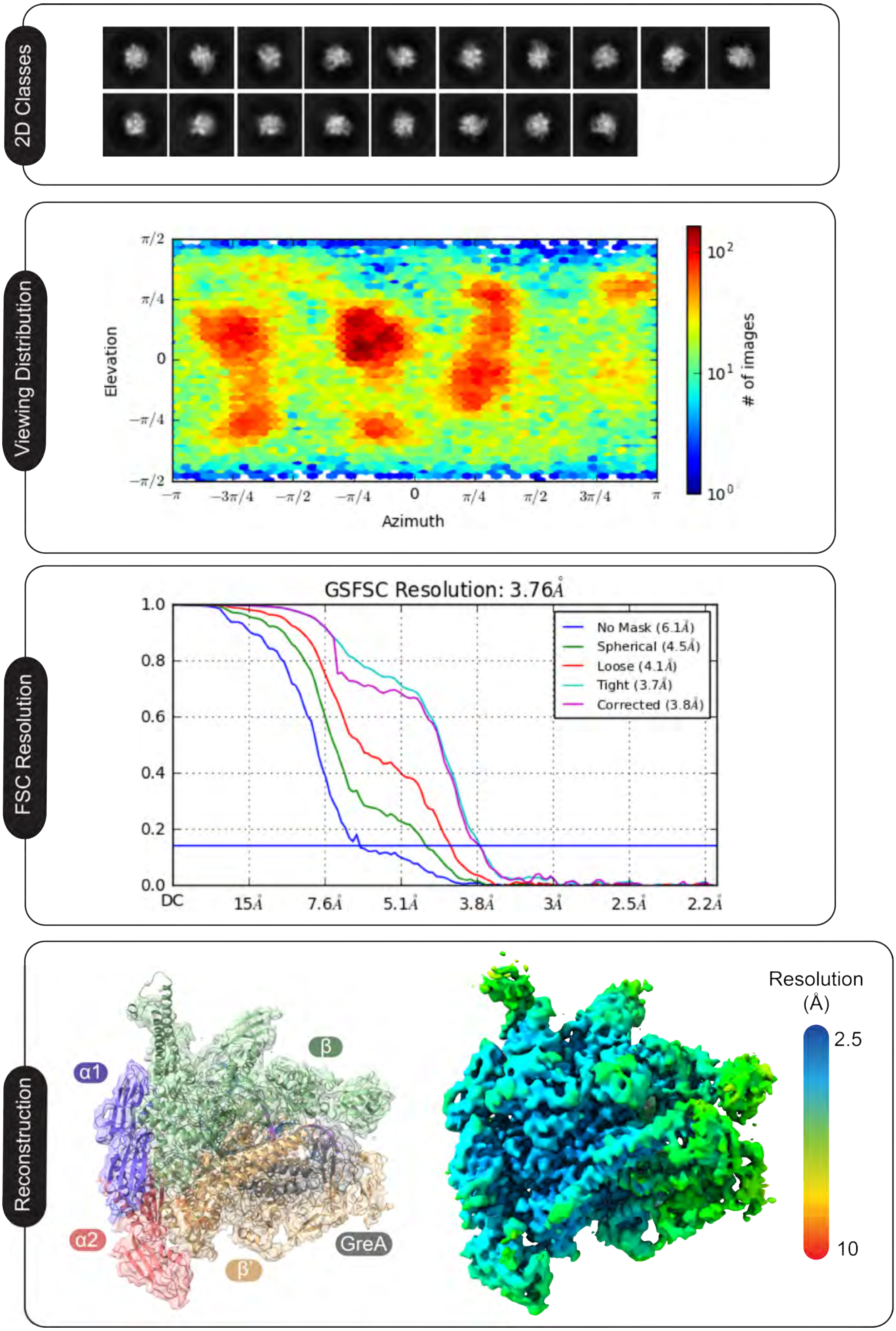


Figure 5.5 – Results of GreA-bound pre-cleavage complex with backtracked base 'A'

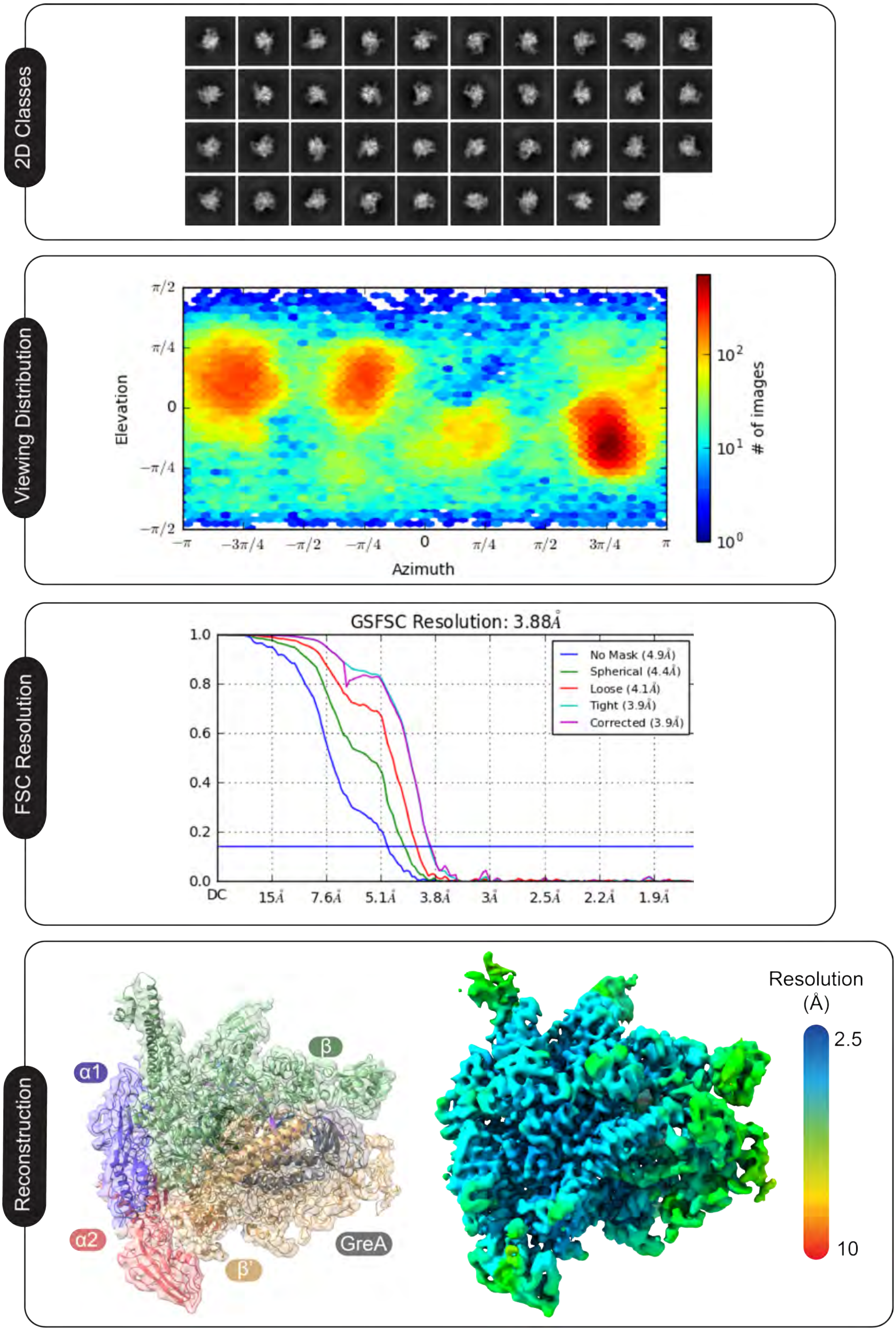


Figure 5.6 – Results of GreA-bound pre-cleavage complex with backtracked base 'G'

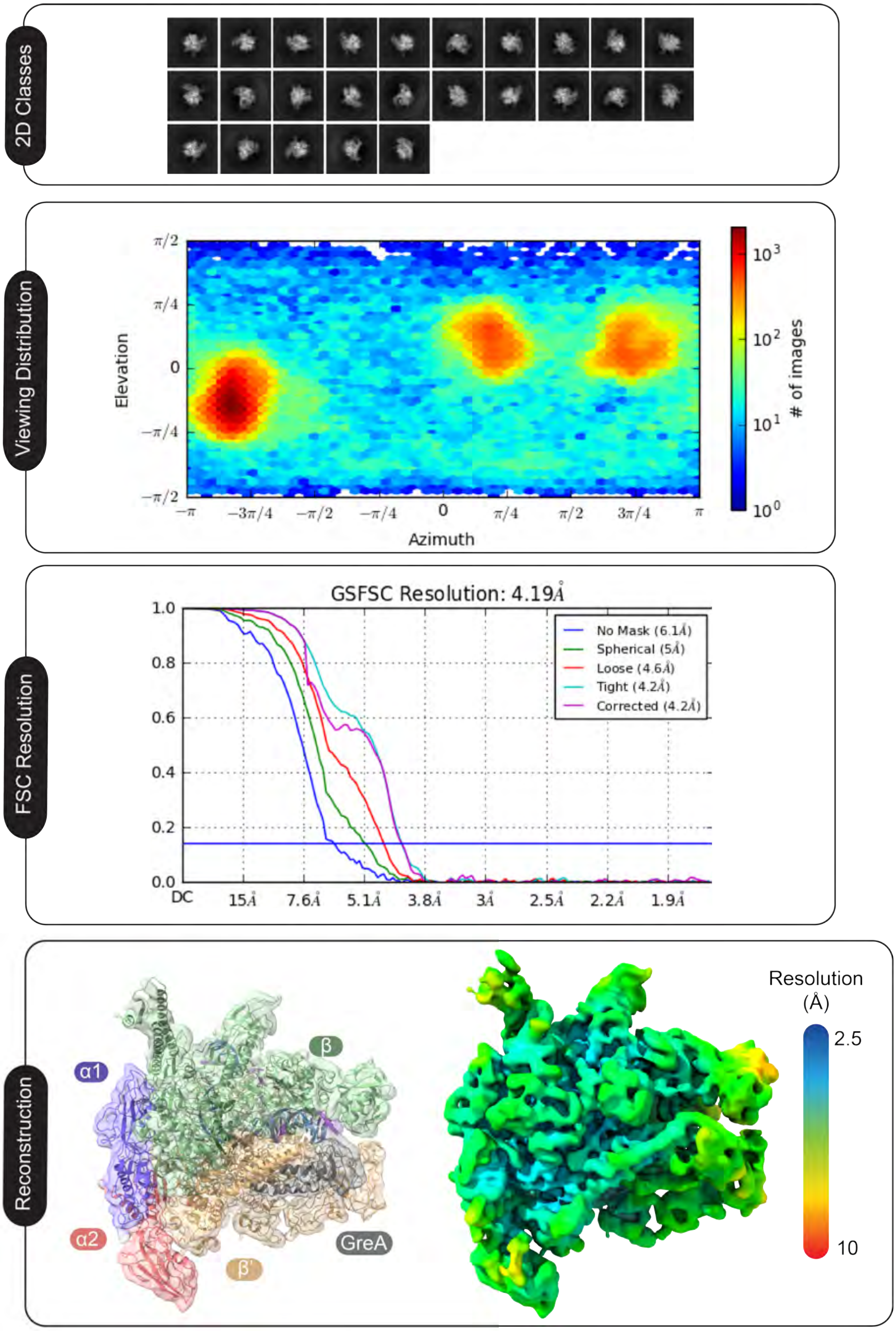


Figure 5.7 – Results of GreA-bound pre-cleavage complex with backtracked base ’C’

and 'non-swivelled'. Figure 5.8 shows a comparison of the swivel modules of the 3 nt backtracked structures and the GreB-bound pre-cleavage complex from Abdelkareem et al. (2019) with the GreA-bound pre-cleavage complex. In it, all the models wre aligned to the core, with only the swivel modules shown to highlight the degrees of swivelling. The GreB structure can clearly be seen in the non-swivelled state, with the GreA structure closer to the swivelled state.

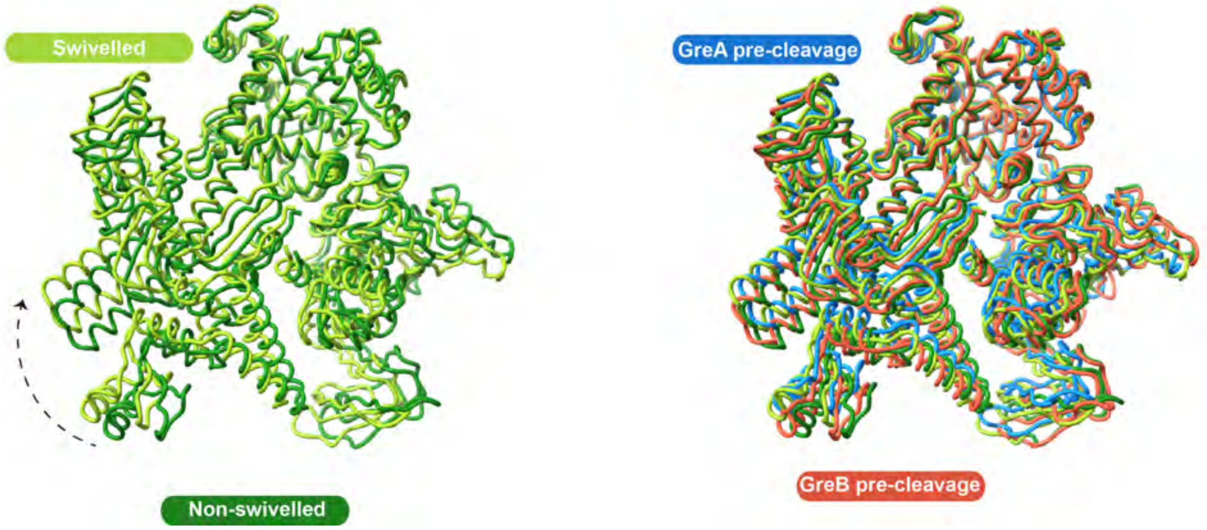


Figure 5.8 – Comparison of swivelling in GreA versus GreB pre-cleavage complexes. Only the swivel modules are shown, and models were aligned to the enzyme core. The two swivelled conformations in the 3 nt backtracked complex (PDB 6RI9 and 6RIP) are shown on the left, with the two Gre-bound pre-cleavage structures on the right (GreB pre-cleavage complex PDB 6RIN)

Most relevant to understanding the catalytic mechanisms which drive hydrolysis are the elements within the catalytic core around the acceptor site (A-site). Along with specific residues within the β and β' subunits, this includes:

1. the bridge helix (BH)

2. the two magnesium ione (Mg-I and Mg-II)

3. the trigger loop (TL), and

4. the NTD tip of GreA

The positioning of these elements relative to one another and to the template DNA and nascent RNA are shown in Figure 5.10. In it, RNA positions ’i’ and ’i+1’ denote RNA NMPs in the A-site and in the backtracked position, respectively. The definitions of all structural motifs discussed here and in later chapters are listed in Table 5.2.

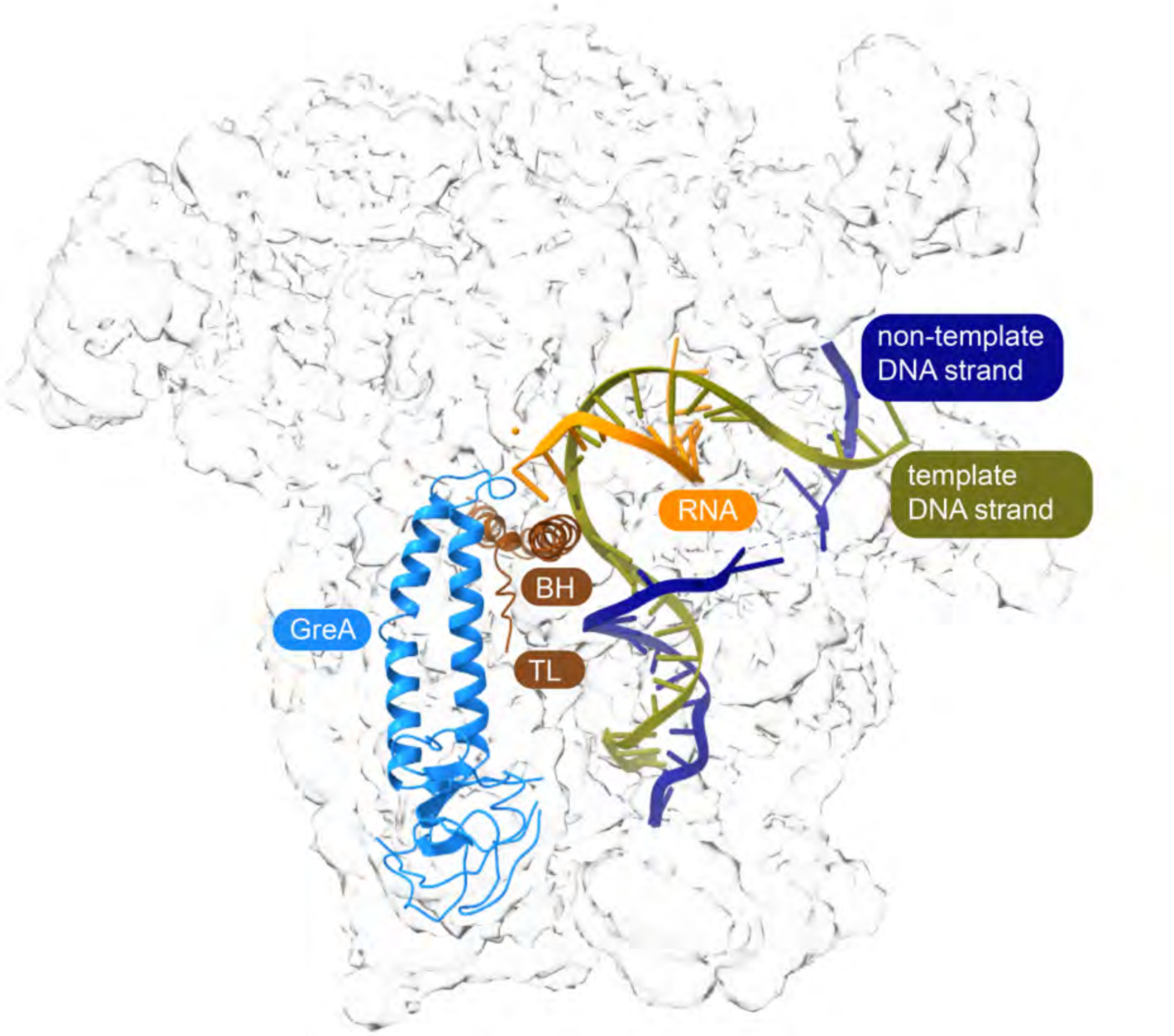


Figure 5.9 – Position of the DNA strands, RNA transcript, BH, TL, and GreA relative to one another in the model of the GreA-bound pre-cleavage complex. The cryo-EM reconstruction is shown in transparent grey, to illustrate the positions of these elements with respect to the complex as a whole.

The BH, implicated in translocation of RNA Polymerase (RNAP) along the template DNA, was observed to be bent, or "kinked", in a direction, which contacts the A-site. This was expected, as the bending of the BH in backtracked states has been described

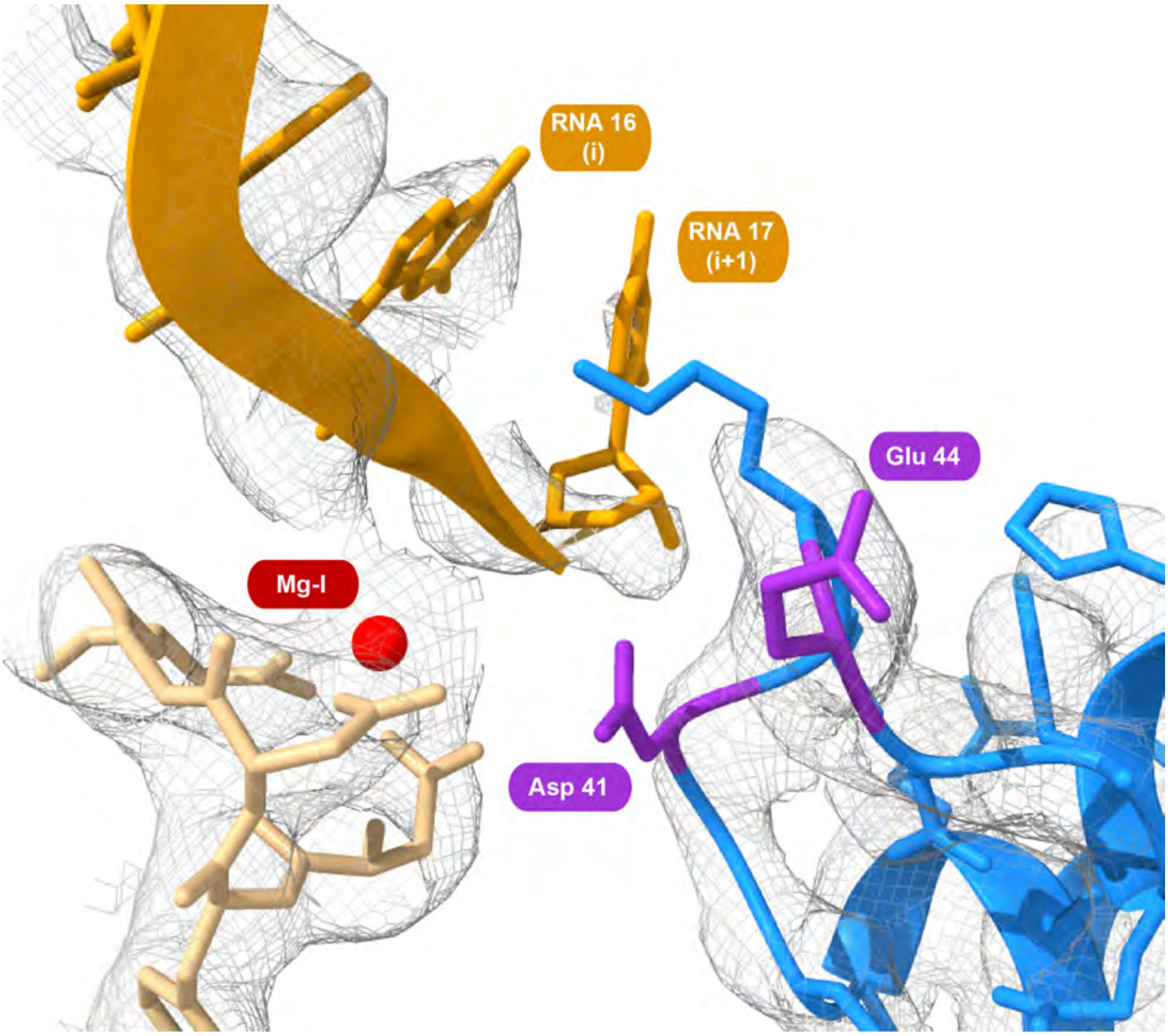


Figure 5.10 – GreA (blue) acidic residues Asp 41 and Glu 44 (highlighted in purple), with Mg-I (red) coordinated by the β' aspartate triad (beige).

and studied well in past papers (Sekine et al. 2015; Bar-Nahum et al. 2005).

Of the two magnesium ions, the binding of Mg-I is more stable, while the second ion, Mg-II, is more transient. In alignment with the expectations that would follow from this –that the density for Mg-I would be more defined than Mg-II –very clear density for Mg-I was observed in all four pre-cleavage structures, with it being the most well-defined in PC-U. However, in the case of Mg-II, there wasn't enough (or any, in some cases) density to reliably model the ion in our structures. In the highest resolution structure, there was some density around the expected position of Mg-II, indicating that it might be present in some fraction of the particles used in the reconstruction. The lack of density for Mg-II did, however, lead us to question another possibility as to why the ion might

Module	***Subunits***	***Residues (chain id (residue numbers))***
Core	2α	A/B 1-234
	β	C 10-26, 514-828, 1071-1235
	β'	D 504-771
Shelf	β	C 1244-1309
	β'	D 346-499, 805-1317, 1358-1407
	ω	all
$\beta 2$	β	C 143-448
Clamp	β	C 1296-1342
	β'	D 1-329, 1321-1344
Lid	β'	D 251-263
Rudder	β'	D 307-326
β'-coiled-coil (clamp helices)	β'	D 264-332
Secondary channel rim helices	β'	D 649-704
Bridge helix	β'	D 770-804
Trigger loop	β'	D 931-$\sim$938, $\sim$1127-1136
E. coli sequence insertion 3 (SI3, also known as β'i6)	β'	D $\sim$945-1130

Table 5.2 – *E. coli* structural modules and corresponding residue numbers used in the structural analysis described in this work

not be seen clearly in the map densities. The RNA used in these complexes contained two phosphorothioate modifications at the 3'-end. This was done to slow down the rate of cleavage as the sulphur in the modified parts of the backbone decreased the coordination of the ions required for cleavage. A possible way around this would be to simply replace the ion in the buffer. Sulphur has a higher affinity for manganese over magnesium, and if the lower density of Mg-II is due to the reduced coordination of magnesium rather than the transient nature of the binding of that ion, replacing magnesium in the EM buffer with manganese would result in better coordination of the second ion. This led to the designing of another set of *in vitro* assays, this time to compare the cleavage of the backtracked RNA in a reaction mix within the standard magnesium-containing Cryo-EM buffer with the cleavage of the same RNA in a reaction mix within an identical buffer that replaced magnesium with manganese. The results for these experiments are described in Section 5.2.

The most interesting observation made on analysis of the structures of the pre-cleavage complexes had to do with the trigger loop. The TL, which in its folded conformation forms the trigger helices (TH) and contacts the A-site, plays an essential role in the catalysis of nucleotide addition, but its exact function during hydrolytic cleavage has not been well-established. In the GreA pre-cleavage complex (GreA-PC) structure, the TL is in an open unfolded conformation. Based on previous structural studies, this general observation was not surprising. What was surprising was that it adopted a very specific open conformation. When comparing the GreA-PC structure with the GreB pre-cleavage complex (GreB-PC) structure, it was very clear to see that the TL in each of them adopted very distinct conformations (Figure 5.11). In the case of the TL in the GreA-PC structure, the domain seemed to align more towards the transcription factor in the SC. Importantly, it seemed to specifically adopt this open conformation in the presence of GreA, firstly because the densities for the TL residues in the highest resolution PC-U structure were well-defined, ruling out the possibility that the domain might have been incorrectly placed in a disordered region during refinements. Secondly, the TL in each of the remaining three pre-cleavage structures was in the same conformation. This showed that for a 1nt backtracked elongation complex, the binding of GreA likely influences the conformation of the TL.

Apart from appearing to influence the positioning of the TL, the GreA coiled-coil NTD was positioned within the SC so that it could contact the A-site. The two main residues of the tip of the GreA NTD, highlighted in Figure 5.10, Glu 41 and Asp 44, were oriented . It should be important to state here that the exact positioning of the two acidic residues can be made primarily from the densities of the protein backbone. Owing to the negative charges of these residues, obtaining reliable densities in maps generated from interactions with the negatively charged electron beam in a transmission electron microscopy (TEM) is often challenging (Wang and Moore 2017).

Another interesting observation that I made during analysis of the maps and models did not have anything to do with the active site of proofreading, but is worth pointing out. At three positions around the β-subunit, marked in black in Figure 5.12, I noticed additional density in the map of the 2.8 Å GreA-PC which did not correspond to any side

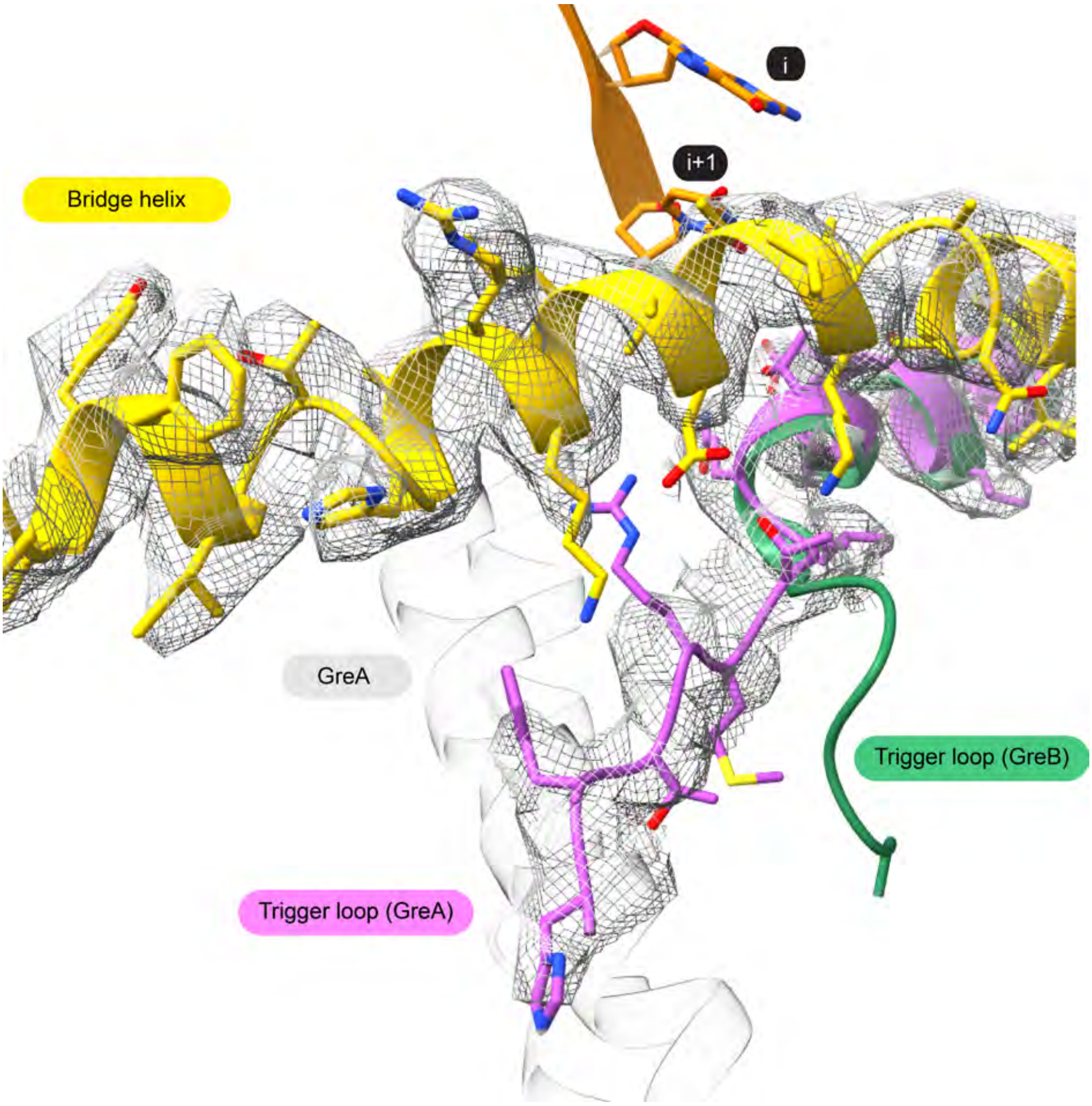


Figure 5.11 – Trigger loop open conformation in GreA pre-cleavage structure

chains within the enzyme. These additional densities were of the detergent Chapso [1], used to eliminate preferential orientation in *E. coli* RNAP complexes. This was confirmed by comparing the observed positions with three known Chapso-binding positions in RNAP, described by Chen et al. (2019).

The total length of the backtracked transcript RNA was 17 bases. The first five 5'-terminal bases are not resolved but the bases starting from position 6 (at the end of the RNA exit channel) to 16 (in the A-site) were well resolved. In the highest resolution

[1] National Center for Biotechnology Information. PubChem Compound Summary for CID 122145, Chapso. https://pubchem.ncbi.nlm.nih.gov/compound/Chapso. Accessed Mar. 4, 2021

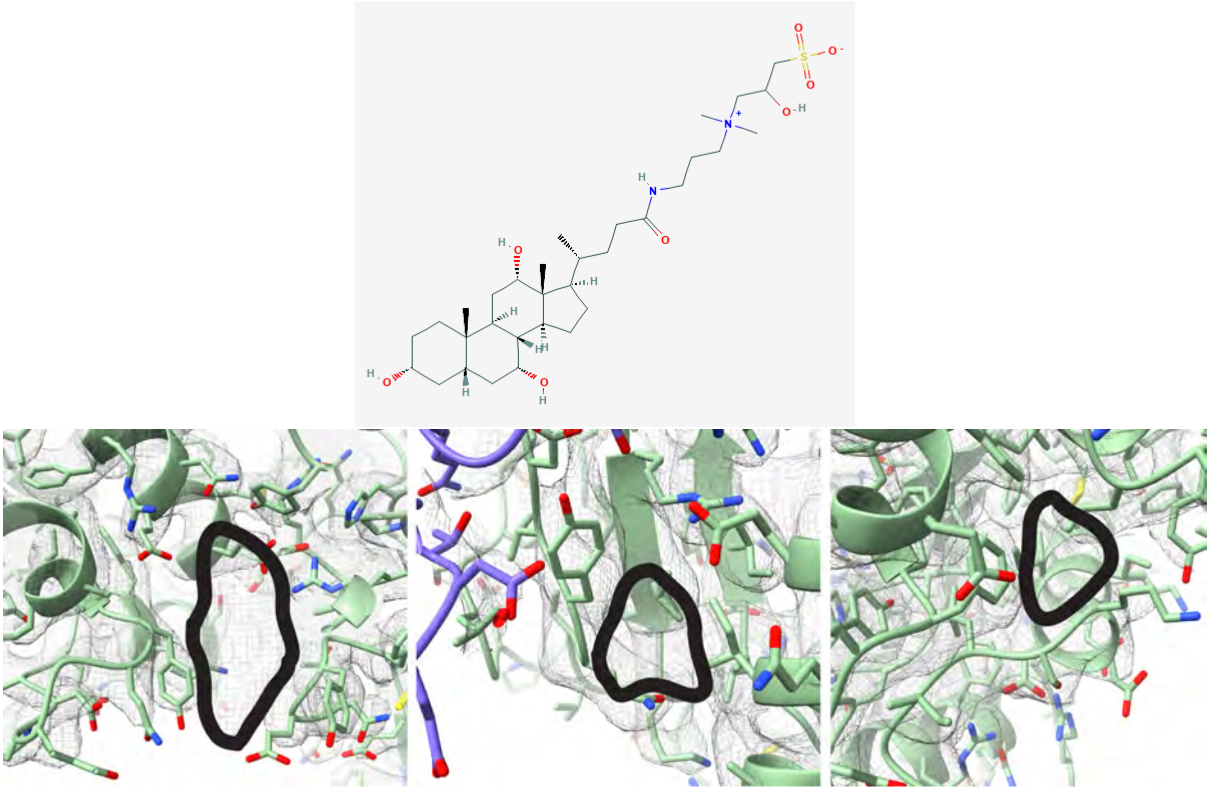


Figure 5.12 – (Top) 2D structure of Chapso molecule[1] ; (Bottom) Chapso binding positions marked in black in the 2.8 Å GreA pre-cleavage map, corresponding to three positions defined in (Chen et al. 2019)

map, the backtracked RNA base (position 17) was not well resolved, with density that got increasingly less well-defined in other maps at lower resolution. Zenkin, Yuzenkova, and Severinov (2006) had proposed a model for cleavage in which the misincorporated backtracked nucleoside monophosphate (NMP) would influence its own cleavage. In the current maps, the backtracked base appears to be quite mobile but with sufficient density to confidently position the backbone.

5.2 Manganese Assays

After failing to see any conclusive density for Mg-II in the maps for the pre-cleavage complexes, a question arose as to whether the reason for that was the inherent transient nature of the binding of that ion, or the presence of the two phosphorothioate modifications in the RNA used for the reconstructions, which are known to lower the affinity for Magnesium. The coordination of the metal ions by these sulphur-containing

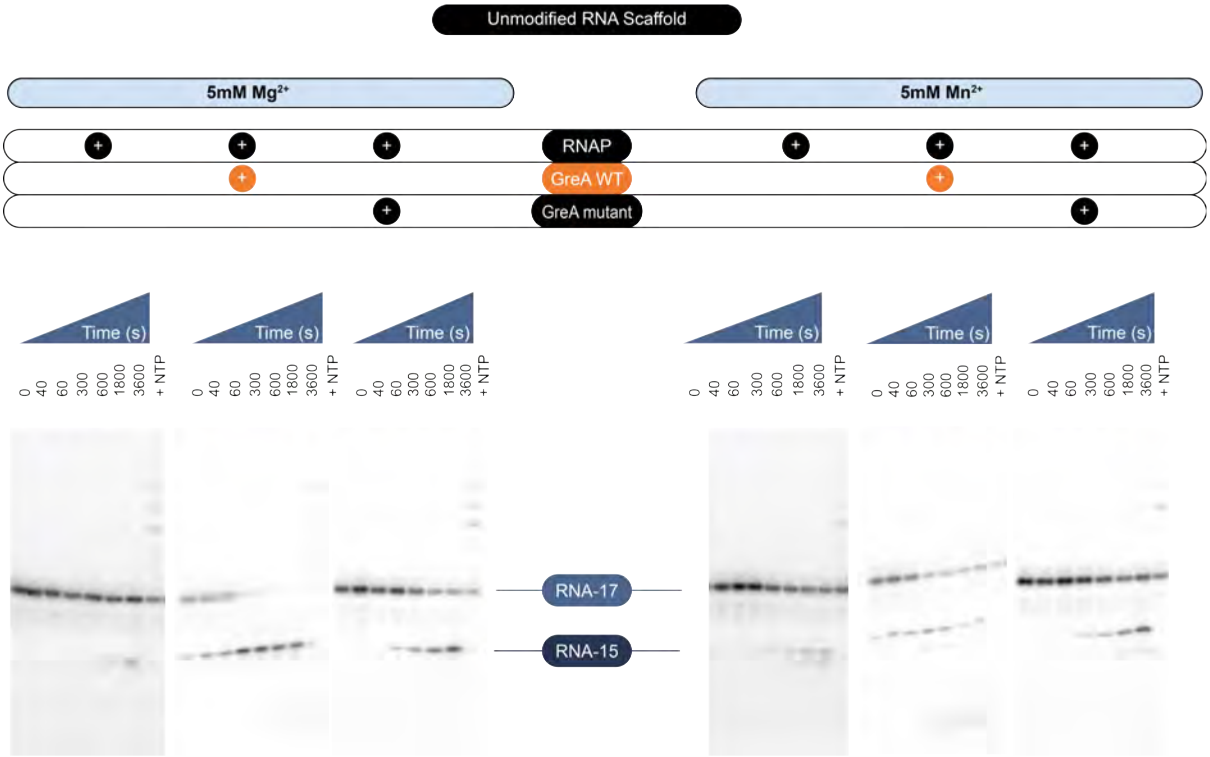


Figure 5.13 – Cleavage Assays comparing Magnesium versus manganese buffers for unmodified Scaffold U

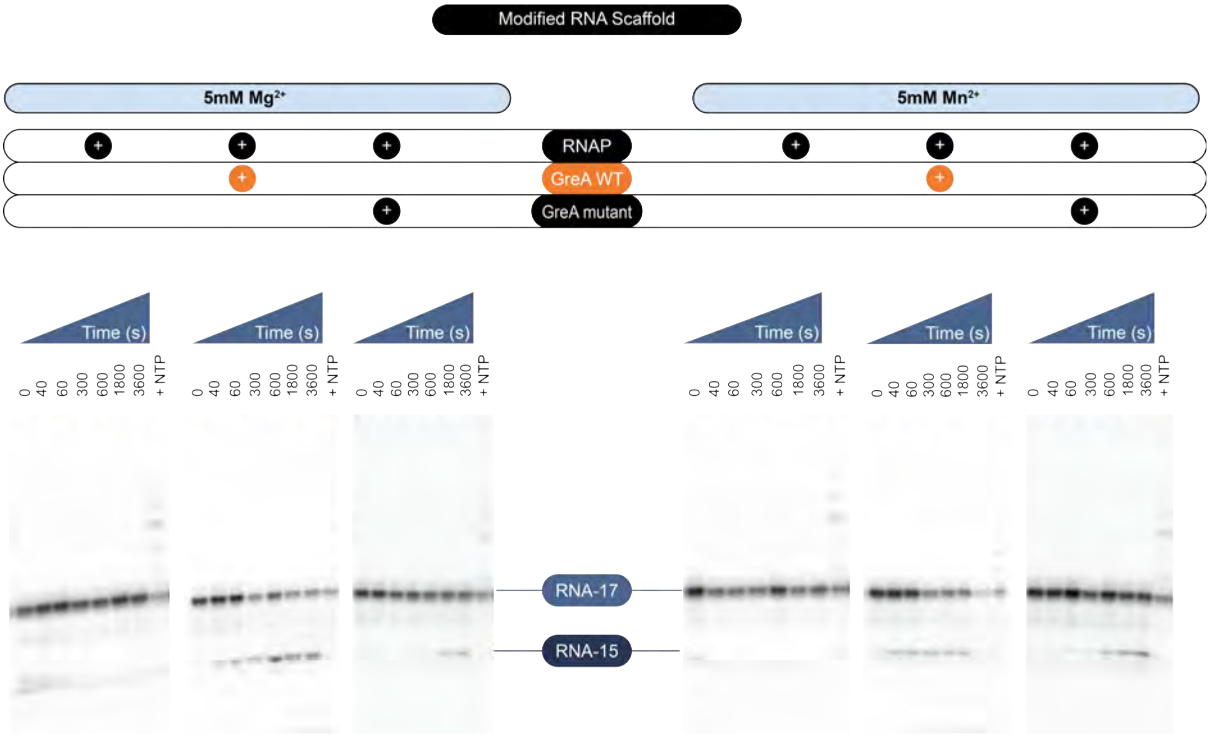


Figure 5.14 – Cleavage Assays comparing Magnesium versus manganese buffers for modified Scaffold U

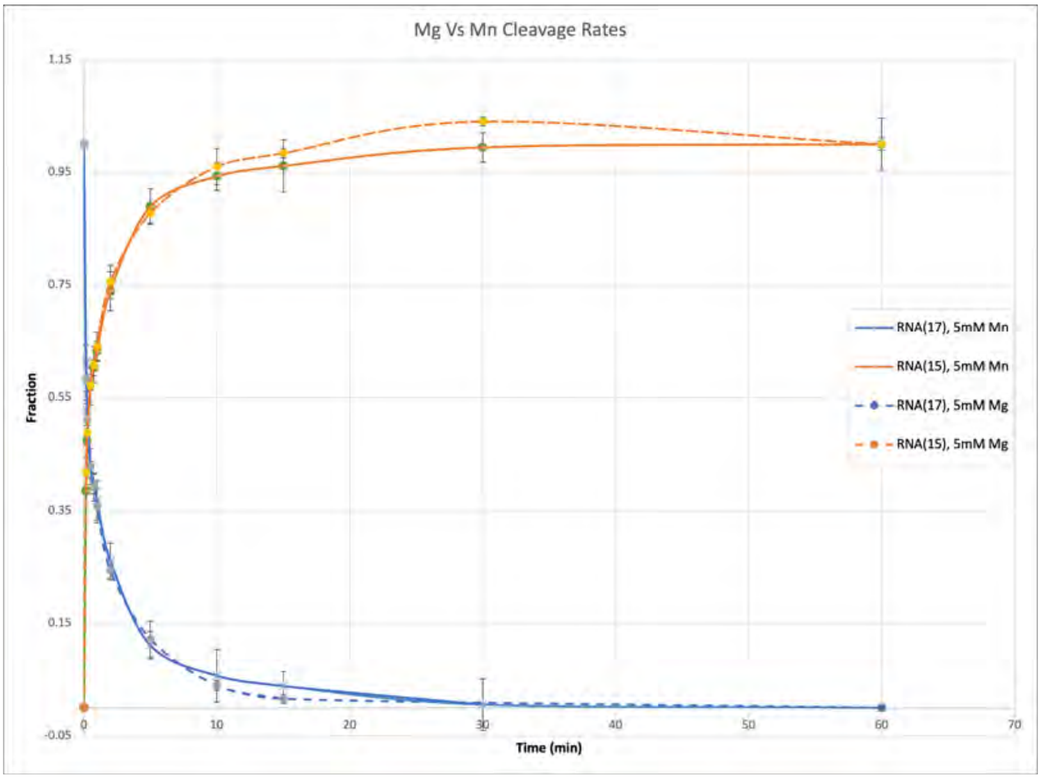


Figure 5.15 – Comparison of cleavage of modified RNA in manganese and magnesium buffers in the presence of WT GreA. The blue lines correspond to the fraction of the uncleaved RNA, with orange corresponding to the cleaved product.

modified RNAs could be increased by replacing magnesium for manganese, which has a higher affinity (Pecoraro, Hermes, and Cleland 1984). Cleavage assays similar to the ones in Section 4.1 were carried out to test the cleavage rates of the same reaction in a magnesium-containing buffer versus a manganese-containing buffer. If, by compensating for the reduced affinity of magnesium by replacing it with manganese in the sample buffer, the rate of cleavage could be restored, there might be a possibility of being able to obtain a cryo-EM reconstruction in which density for the second ion would be well-ordered.

The magnesium buffer was the same standard cryo-EM buffer used for protein storage, transcription assays, and for cryo-EM grids. The reactions under these two conditions were performed with RNAP alone, RNAP with WT GreA, and RNAP with the GreA E41A-D44A mutant. The results for these experiments are illustrated in Figures 5.14. As a control, the same set of experiments was performed for the same scaffold without any RNA modifications, shown in Figure 5.13.

This experiment, in which each set of reactions was carried out in triplicate, also

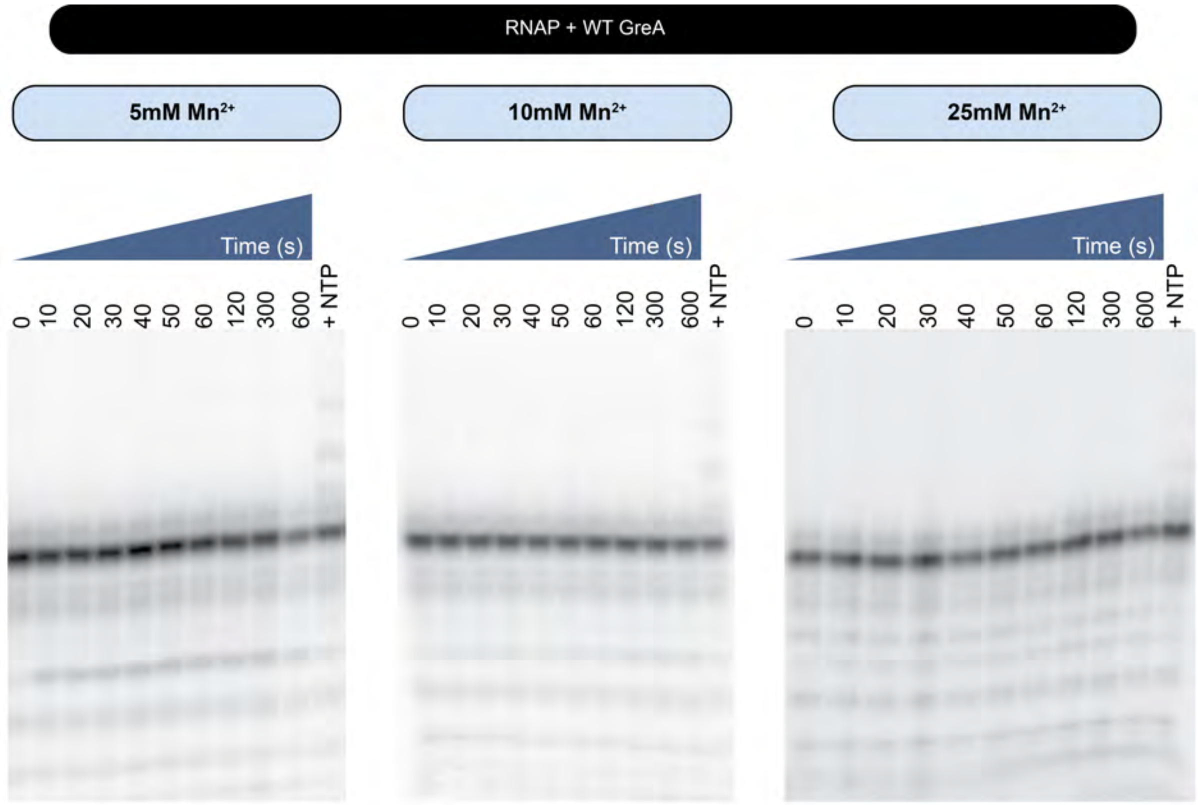

RNAP + WT GreA
5mM Mn²⁺
10mM Mn²⁺
25mM Mn²⁺
Time (s)
Time (s)
Time (s)
0 10 20 30 40 50 60 120 300 600 + NTP
0 10 20 30 40 50 60 120 300 600 + NTP
0 10 20 30 40 50 60 120 300 600 + NTP

Figure 5.16 – Comparison of different concentrations of manganese on GreA-assisted cleavage of modified RNA

allowed me to plot the extent of cleavage for RNAP alone, with WT GreA and with the D41A-E44A mutant GreA Figure 4.4. This would be useful to check the extent cleavage in the presence of mutant GreA in comparison with the reactions with and without WT GreA. In the graph, the data points for cleavage with WT GreA are the upper limit, with the lower limit set by the cleavage in the absence of GreA (RNAP alone). The points for the cleaved fraction in the presence of the GreA mutant clearly lie in between the upper and lower limits, especially for the first half of the time course.

With the modified RNA, no significant difference in cleavage was seen for the manganese buffer in comparison with magnesium (Figure 5.15). As seen previously, cleavage of the modified RNA by RNAP alone barely occurred, with GreA stimulating the cleavage of a small fraction of RNA. Rather than behaving similar to the reaction with RNAP alone, the presence of the GreA mutant did stimulate the cleavage of the modified RNA to a small degree. In the case of the unmodified RNA, the backtracked portion gets almost completely cleaved in the presence of WT GreA in the magnesium buffer, with very little intrinsic cleavage taking place by the final 1 hour time point and the mutant GreA again promoting a larger amount of cleavage than what would be expected. Across all six scaffold + reaction conditions, manganese at a 5mM concentration doesn't appear to promote or impede cleavage. The only exception is the cleavage of the unmodified RNA in the presence of WT GreA, which is slowed down in the presence of Mn^{2+}.

To finally assess whether simply increasing the concentration of manganese in the buffer would promote cleavage of the modified RNA, another set of assays was performed with buffers containing 5mM, 10mM, and 25mM Mn^{2+} in the reaction buffer. Amongst the three conditions, there wasn't an increase in the fraction of cleavage for the higher concentrations of manganese.

Conclusions

Three key observations were made:

- During di-nucleotide cleavage of backtracked RNA in presence of GreA, the trigger loop appears to orient itself in a very specific open conformation.

- GreA is known to increase the efficiency of cleavage through its two conserved acidic residues. However, mutating these residues only slightly reduces the cleavage efficiency of a GreA-RNAP complex, suggesting that GreA might not just work solely through interactions of the acidic residues at its NTD tip with the RNAP active site.

- Replacing magnesium with manganese doesn't appear to compensate for the loss of cleavage activity of phosphorothioate RNA. Therefore, it would be extremely unlikely that a dataset of the complex in the manganese buffer would help me in identifying the presence of the second metal ion.

Going forward, the significance of GreA binding and the TL needed to be explored further. To address these points, the clearest way forward would be to obtain a structure of the 1 nt backtracked complex in the absence of GreA.

Chapter 6

Backtracked Complex

Following the results of the pre-cleavage complexes, it became clear that in order to get a clearer picture of effects of GreA-binding on proofreading in general and on the conformation of the active site in particular, I also needed to obtain a structure of a backtracked complex without GreA bound to the secondary channel. Preliminary comparisons were made between the GreA pre-cleavage complex (GreA-PC) and 3 nucleotide-backtracked complexes (PDB accession codes 6rip and 6ri9, Abdelkareem et al. (2019)). However, this was to just get a general idea of the overall conformation of RNA Polymerase (RNAP). To really understand how GreA influences the enzyme core elements around the acceptor site (A-site), having a 1 nucleotide-backtracked complex instead would be ideal.

6.1 Single-Particle Cryo-EM of BAcktracked Complex

The scaffold used for the backtracked complex (BC) was the same as the one used in the 2.9 Å pre-cleavage structure (backtracked base 'U'). This complex was initially used to test out the feasibility of using graphene oxide (GO)-coated grids with *E. coli* RNAP complexes, for which the datasets either could not be used to generate a high-resolution reconstruction, or resulted in maps with very strong orientation bias. With the given time constraints, it was important to focus on grid-freezing protocols that worked well with the pre-cleavage complexes, i.e., the gold grids (UltrAuFoil 1.2/1.3). A high resolution dataset was first collected in August 2020, but was unusable in the end due to beam shift

1nt Backtracked Complex	
Grid	UltrAuFoil 2/2
Particles Picked	552799 (Topaz)
Particles Used for Consensus Refinement	296710
Raw Micrographs	9798
Pixel Size *(Å/pix)*	0.862
Defocus Range	-0.8 to -2
Voltage *(kV)*	300
Electron Dose *(e-/Å2)*	50.9
Microscope	Titan Krios (IGBMC)
Detector	K3

Table 6.1 – Data Collection parameters for the backtracked complex

issues during the data collection. Another dataset collected in November of the same year resulted in a 3.8 Å reconstruction, shown in Figure 6.1. The data collection parameters are listed in Table 6.1.

3D classification of the extracted particles revealed that there were two distinct classes present in the dataset (Figure 6.2). Of the 296710 particles remaining after 2D classification and clean-up, 118760 (approximately 40%) were put into Class 1, while the remaining 177950 (approximately 60%) in Class2. The two classes refined to 3.9 and 3.6 Å respectively. Comparison of the two maps revealed that the difference between the RNAP conformations in the two classes was primarily in what is called the swivel module. The RNAP structural motifs which make up the swivel module are the RNAP clamp, and shelf modules. The residues which constitute these motifs are listed in Table 5.2. Interestingly, the occupancy of the swivelled state versus the non-swivelled state in the 1 nt BC was found to be similar to what was estimated for the 3 nt BC in Abdelkareem et al. (2019).

In both the backtracked complexes, as well as in the pre-cleavage complexes described in the previous chapter, local resolution maps clearly highlight that RNAP regions towards the periphery were slightly more disordered with poorer resolutions. Of these, the

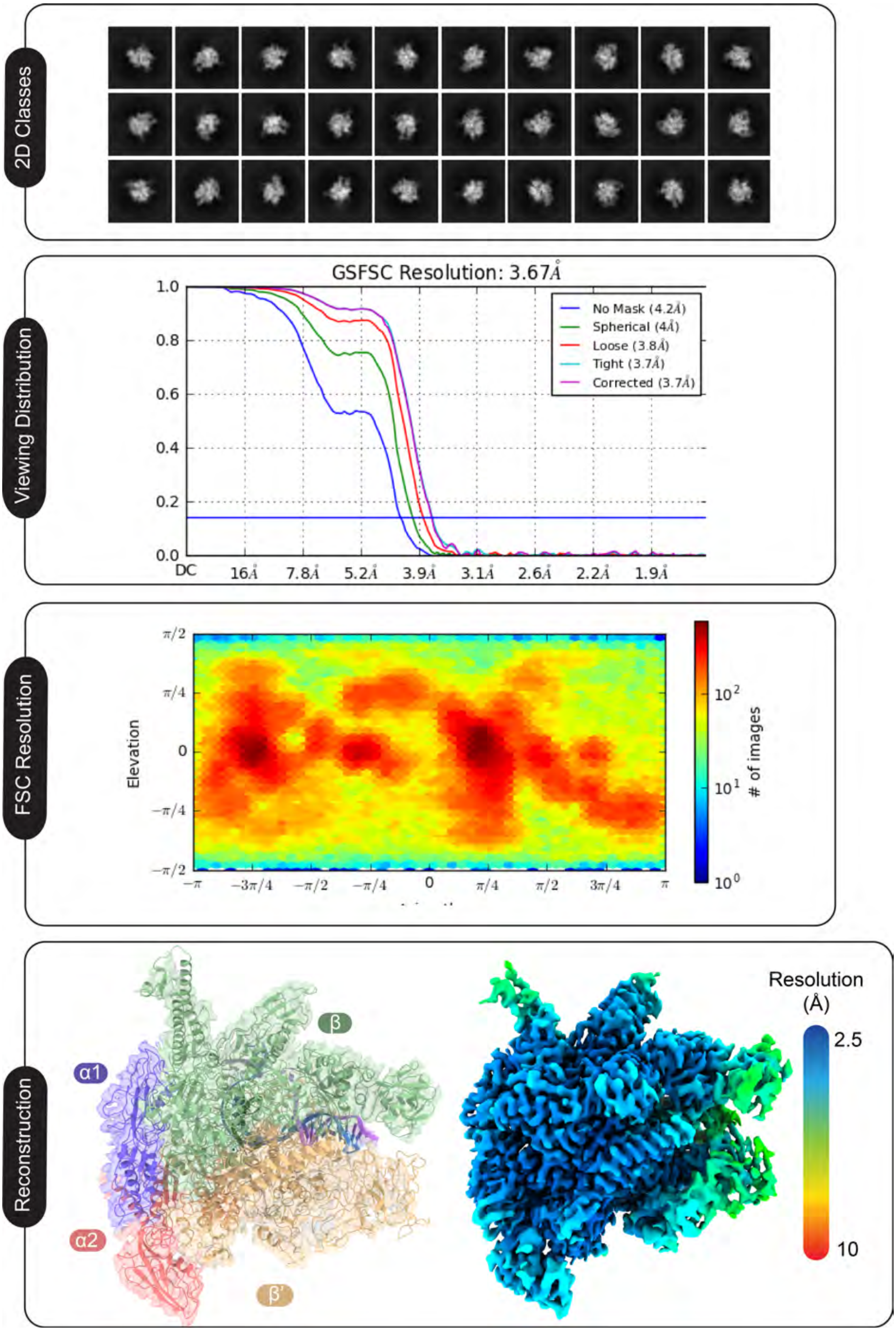


Figure 6.1 – Consensus refinement of the backtracked complex

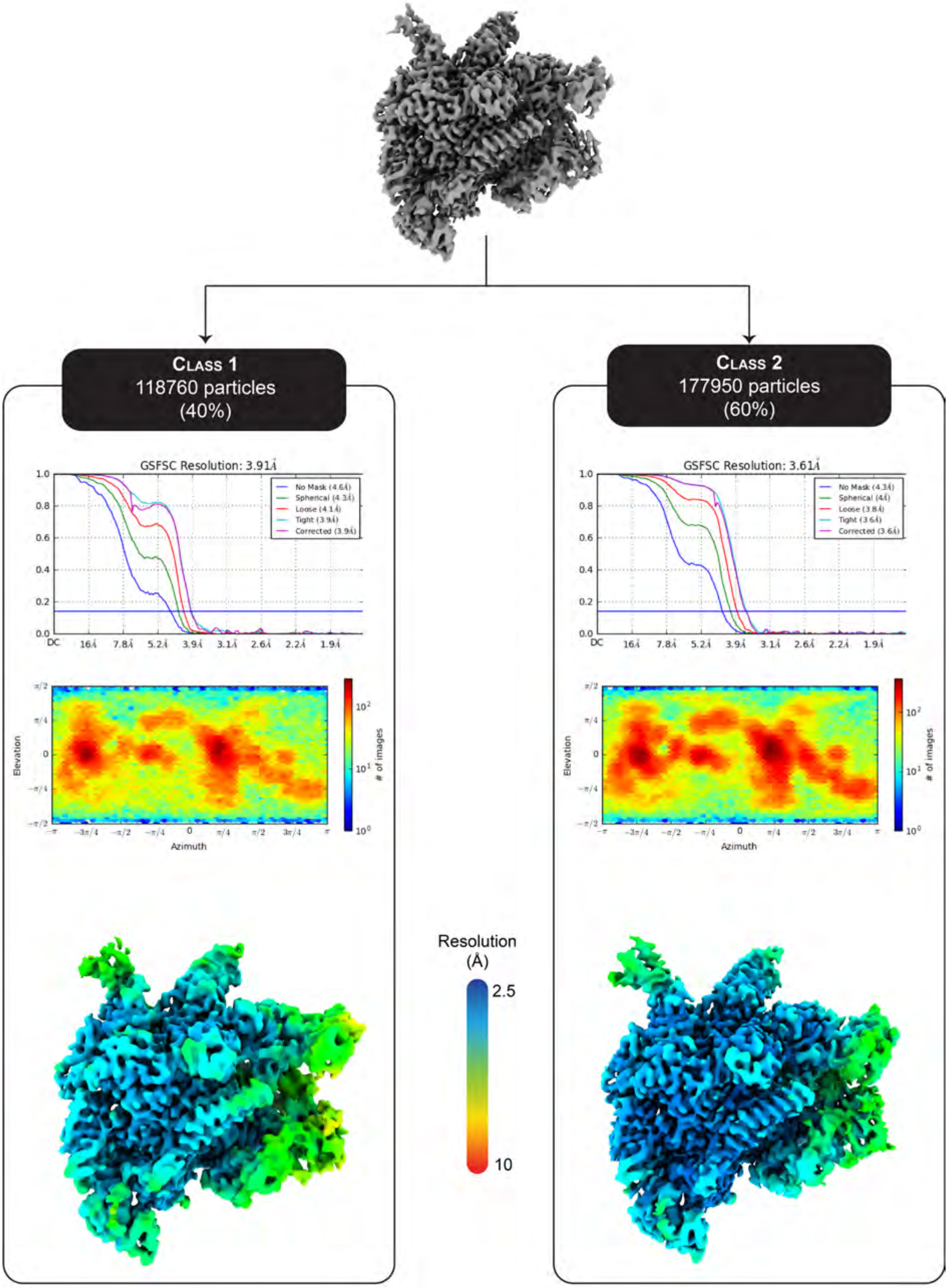


Figure 6.2 – Classification of the backtracked complex

largest disordered region was the SI3 domain, which is a lineage-specific insertion in the trigger loop (TL) of *E. coli* RNAP (Artsimovitch et al. 2003). Density for the backtracked base in both classes was very weak. Although the resolution of the backtracked bases was poor in the GreA-PC complex maps, even the ones with resolutions comparable to the BC maps showed density for at least the backbone.

The main point of interest for the two backtracked classes was the TL. In the GreA-PC structures, the TL in the open conformation was well-resolved up to the residue His 936. In the case of the two backtracked structures obtained after 3D classification, the TL appeared disordered in both. To confirm whether or not this was not simply due to the lower resolution of the BC maps, they were compared with the TL regions in the lower resolution resconstructions of the GreA-PCs. This included the 4.2 Å preliminary reconstruction in Figure 4.10. For each of the GreA-bound complexes, the density of the TL in the novel conformation was clear and well-defined. This suggests the interaction of the TL with GreA in the GreA-PCs stabilizes this novel conformation.

6.2 Comparisons with Pre-Cleavage Complexes

The two classes of the 1 nt BC were compared with the structure of the GreA-PC as well as with the GreB pre-cleavage complex (GreB-PC) and with the swivelled and non-swivelled 3 nt BCs (The GreB complex and the 3 nt complexes from Abdelkareem et al. (2019)). As mentioned previously, the different orientations of the swivel modules in the 1 nt BCs was similar to the different classes of the 3 nt BCs.

Figure 6.3 shows the swivel modules for different sets of backtracked and pre-cleavage complexes. Structural superposition for all models to each other were done using the RNAP core module as a reference, a structurally stable part of RNAP. The extent of swivelling seen in the 1 nt backtracked complex was measured through least squares fitting of the swivel modules without the flexible SI3 domain. The rotation between the non-swivelled and swivelled conformations was measured to be 3.28° and is shown in Figure 6.4.

In addition to the comparisons of the two sets of BCs, the swivel modules of the GreA- and GreB-bound pre-cleavage complexes is also shown (GreB pre-cleavage structure from Abdelkareem et al. (2019)). The aligned structures revealed that while the GreB-PC adopts a non-swivelled conformation, the GreA-PC is in a more swivelled state. In Figure 6.5, the positioning of GreA and GreB and of the SI3 domains for each complex are illustrated. Both the structures in the figure have been aligned by their swivel modules (excluding the SI3 domain). By aligning the swivel modules, the GreA and GreB NTDs also align to one another. Compared to when GreB is bound to the secondary channel (SC), binding of GreA is coupled with the SI3 and the β-lobe, which flank the Gre CTD from two sides, to move towards each other.

GreA does not only influence the positioning of RNAP domains on the surface, but influences the positioning of various active centre elements relative to each other. The differences in specific elements of the two BCs and the GreA-PC active centres are shown in Figure 6.6. In it, GreA appears to influence the positioning of the entire bridge helix (BH) with respect to the A-site. Differences are seen not just for the BH, but also for the TL helices. The difference in the positions of the SI3 domain, which is a sequence insertion within the TL, is also shown. The backtracked RNA base (RNA-17), positioned for cleavage in the GreA-PC structure, clashes with the BH in both the swivelled and non-swivelled backtracked structures.

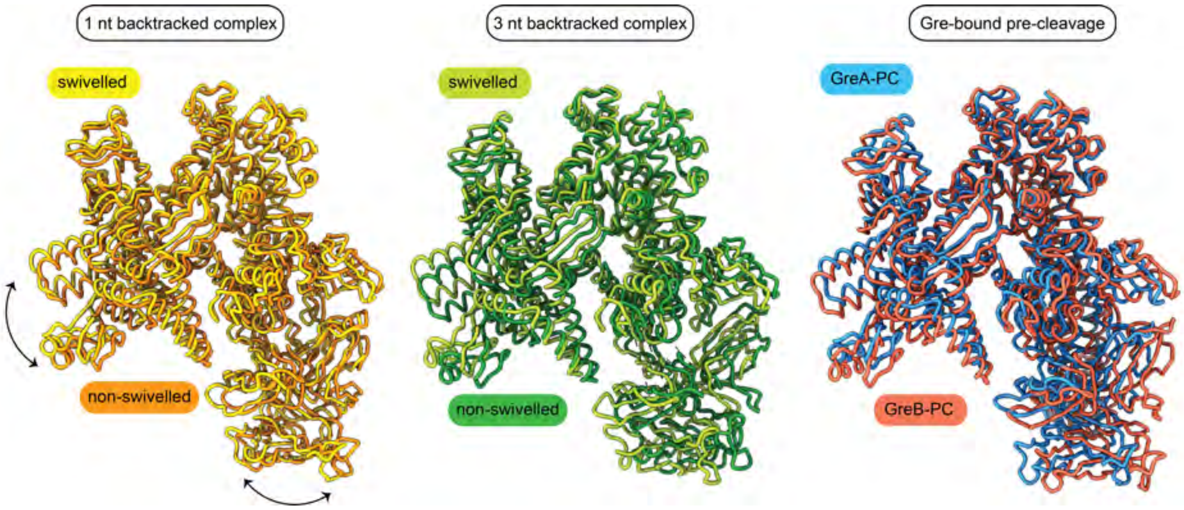


Figure 6.3 – Swivel modules in different complexes aligned to the RNAP core, illustrating that swivelling is seen regardless of the backtracked length, and more importantly that the GreA and GreB backtracked complexes adopt different swivel conformations.

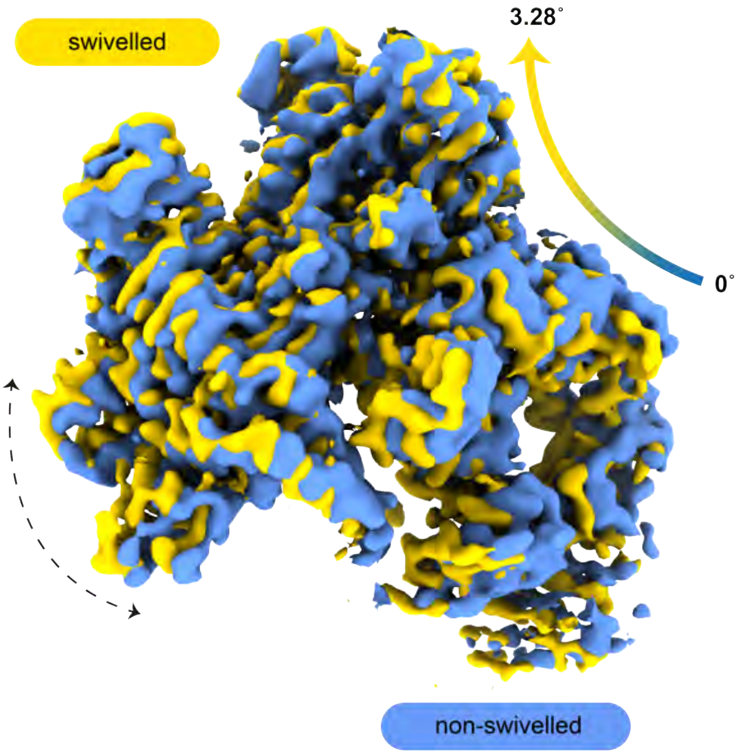


Figure 6.4 – Rotation of the swivel module of the 1 nt backtracked complex

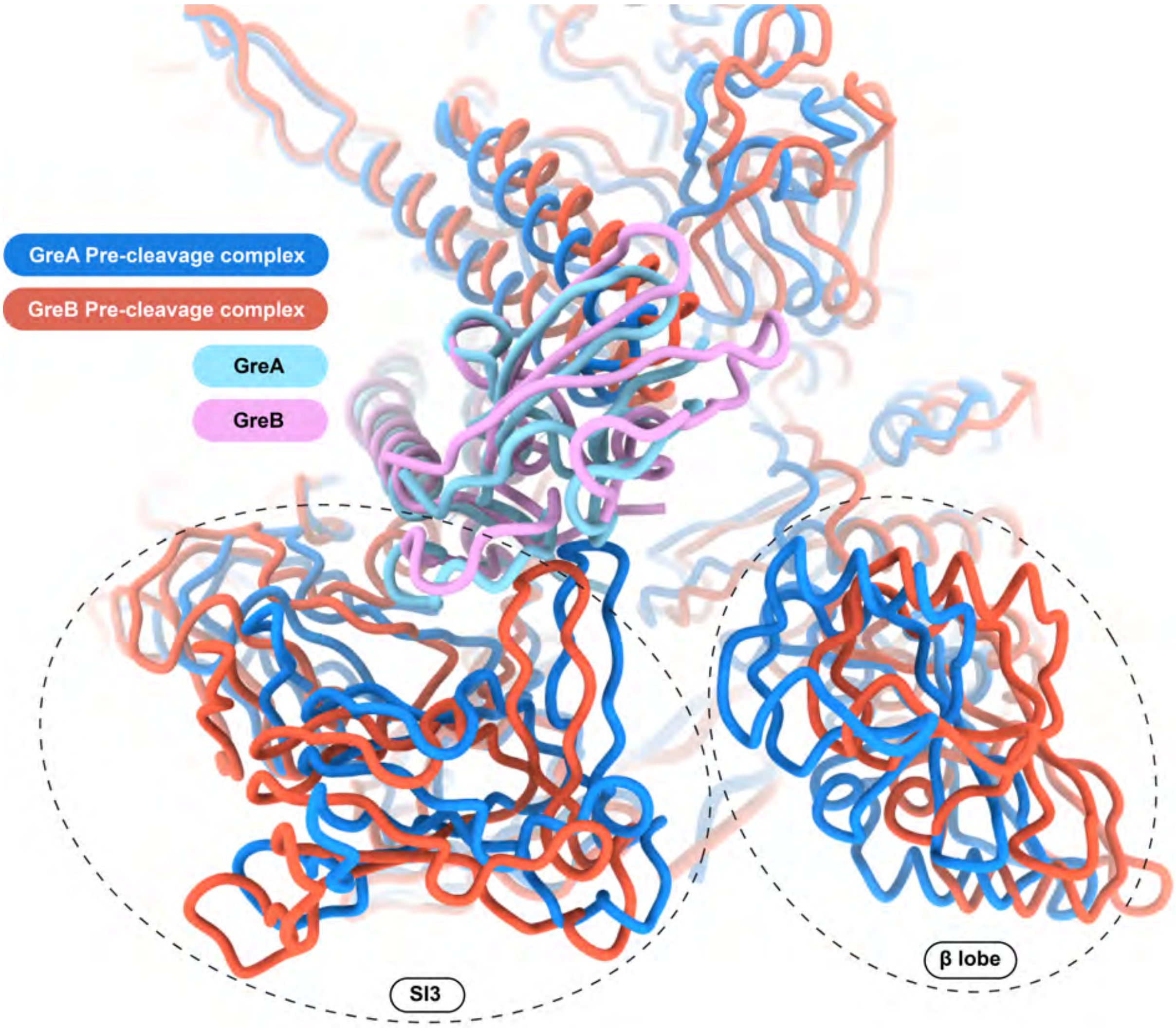


Figure 6.5 – SI3 domain placement in GreA and GreB-bound complexes, aligned by the swivel module excluding the SI3. The SI3 and β-lobe, which form the ends of the pincers in the RNAP crab claw, are circled.

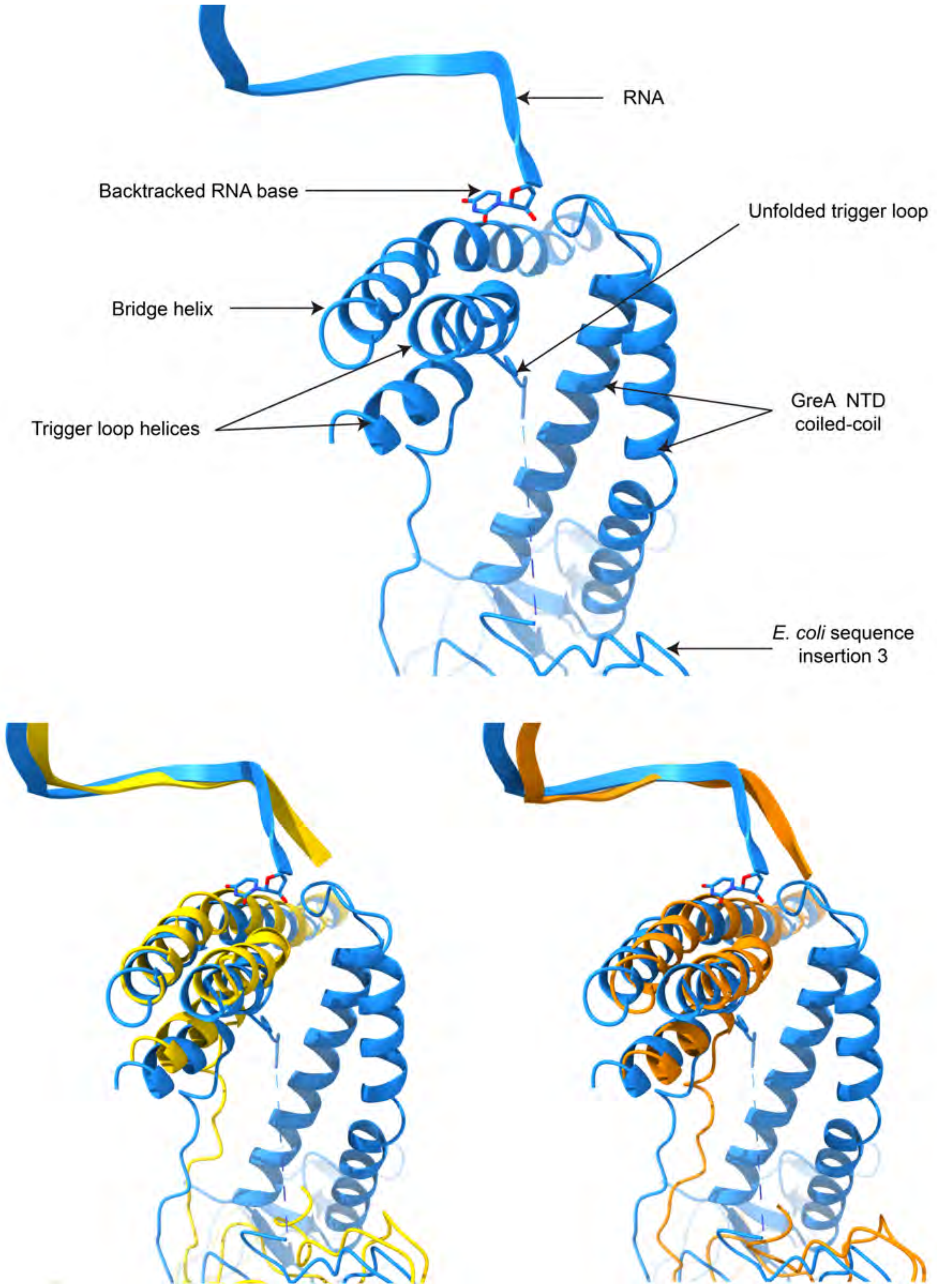


Figure 6.6 – Positions of different active site elements relative to another in the structures with and without GreA. The elements for the GreA pre-cleavage complex are shown in blue, with individual comparisons with the swivelled (yellow) and non-swivelled (orange) backtracked complexes at the bottom.

Chapter 7

Discussion

In total, five high resolution single-particle cryo-EM datasets were collected on *E. coli* RNA Polymerase (RNAP) transcription elongation complex (TEC)s backtracked by 1 nucleotide (nt), four of which included the proofreading factor GreA bound to the secondary channel (SC). In addition to this, different *in vitro* transcription assays were performed to assess the functionality, cleavage positions and cleavage rates of these complexes.

7.1 Swiveling in Backtracked Complexes

Classification of particles picked from the backtracked complex dataset revealed two distinct classes of particles –labelled swivelled and non-swivelled –defined by the relative conformations of the swivel module (RNAP clamp and shelf). 60% of the particles were classified into the swivelled state, with the remaining 40% in the non-swivelled state. Interestingly, this was similar to the ratio of swivelled to non-swivelled states that was seen in the 3 nt backtracked complex (63:37 swivelled to non-swivelled) (Abdelkareem et al. 2019).

The swivel module, which is known to rotate as a single rigid body, can adopt different states depending on the state of the complex. To test whether the two conformations seen were discrete or two points of convergence over a continuum of states, 3D variability analysis was performed in CryoSPARC with the set of particles. Variability analysis did reveal that swivelling takes place more as a range of states and reflects the equilibrium of accessible swivel module positions. However, the classification of particles into the two states does suggest that the swivelled state might be more energetically favourable than the non-swivelled state (i.e. the equilibrium is shifted towards a more swivelled conformation). In contrast with the different states observed for the backtracked complex (i.e. in the absence of GreA), the pre-cleavage complex of RNAP backtracked by 1 nt with GreA did not adopt different conformations, at least not significantly enough to be able to perform 3D classification. The swivel module in the GreA pre-cleavage complex (GreA-PC) structure was more swivelled, although not to the same extent seen in the absence of

GreA. As with the GreA-PC, the GreB pre-cleavage complex (GreB-PC) structure from Abdelkareem et al. (2019) was also observed to be in a single conformational state, but in this case the swivel module was in the non-swivelled position.

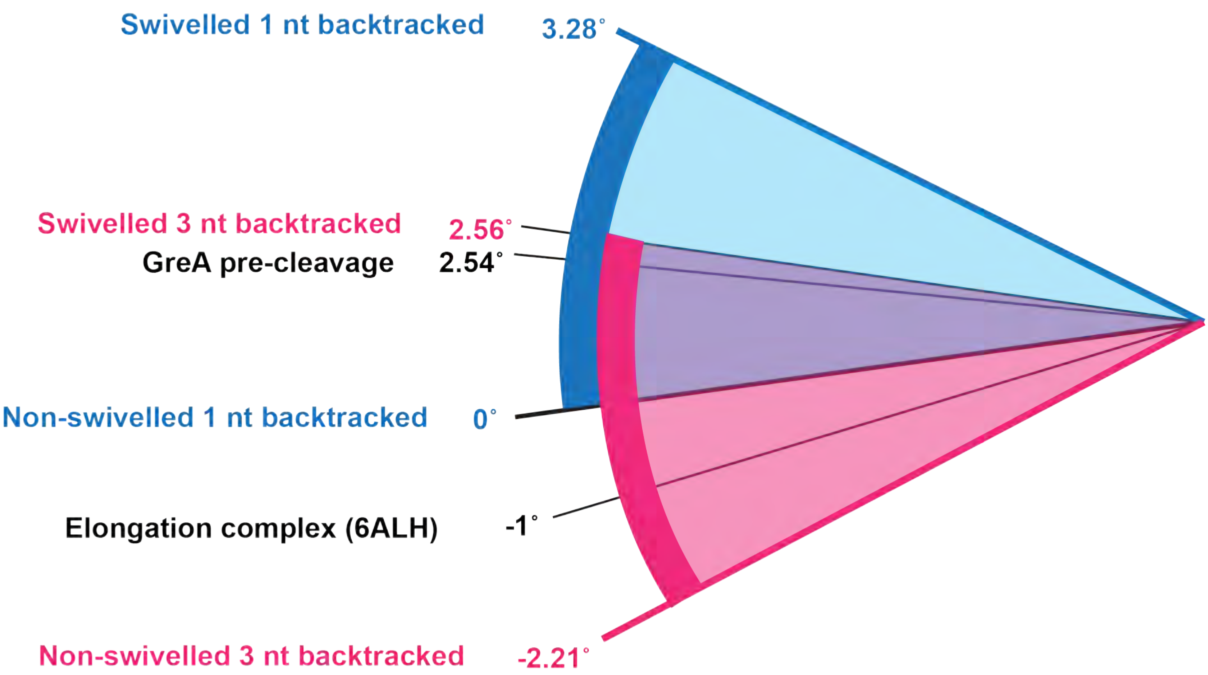


Figure 7.1 – Different rotation angles for swivel modules in different complexes relative to the non-swivelled 1 nucleotide backtracked complex (0°). The swivelling ranges for the 1 and 3 nucleotide backtracked complexes are marked in blue and pink, respectively.

To really understand the implications of swivelling on backtracking and the effects of GreA binding, the rotation angles were measured and compared to each other. Comparison of the swivelling observed for the 1 nt and 3 nt backtracked complexes revealed a difference in both the extent of swivelling and in the range (start and end points), as shown in Figure 7.1. Rotation of the swivel modules was measured for the swivelled and non-swivelled states of the 1 nt and 3 nt backtracked complexes, the GreA-PC, and the elongation complex from Kang et al. (2017). The swivelling degrees for all of the complexes were measured in ChimeraX by first aligning (least squares fitting) the models using only the RNAP core module (residue numbers in Table 5.2) and measuring the rotation angle required to then align the swivel module. The sequence insertion 3 (SI3) domains were excluded from rotation angle measurements as that region is highly flexible while the rest of the swivel module tends to move as a single rigid body. Measurements were all made with respect to the non-swivelled 1nt backtracked complex (BC) (0°). The 1nt BC was able to swivel through an angle of approximately 3.3°, which was slightly less than the swivelling seen in the 3nt BC (4.8°). Importantly, as illustrated in Figure 7.1, the start and end points are shifted depending on the extent of backtracking.

We see that the range of swivelling is different depending on the length of the backtracked RNA, even when the difference is (in this case) 2 nucleotides. In the presence of GreA or GreB, which assist in cleavage of shorter or longer backtracked RNA respectively, the complex also adopts a different swivelled conformation. While the GreB-PC aligns with the 3 nt non-swivelled complex, which in our comparisons shows the least degree of swivelling (the least swivelled) , the GreA-PC aligns close to the 3 nt swivelled conformation, which is in between the swivelled and non-swivelled conformations in the 1 nt BC.

This might begin to explain how a backtracked complex may select for binding of GreA or GreB to the secondary channel, depending on the extent to which the complex is backtracked. In turn, by restricting the movement of the swivel module, binding of either of these cleavage factors into the SC might fix the swivel module in a specific conformation.

7.2 Differentiation Between GreA and GreB

To fully understand the way in which the complex might be able to differentiate between the two Gre factors, it was worth taking a closer look at the interactions of GreA with the SI3 domain, since it appears to interact with the GreA CTD. SI3, also referred to as β'i6, is an *E. coli*-specific sequence insertion within the trigger loop (TL) and extends from the active site to the surface of RNAP. The swivel modules of the backtracked complexes as well as of the GreA-PC and GreB-PC are shown in Figure 7.2, all aligned to the swivel module excluding the flexible SI3 domain. Differences in the positioning of the SI3 domain when the rest of the swivel module was aligned for different complexes further highlighted the differences seen when the complex is backtracked by different lengths, as

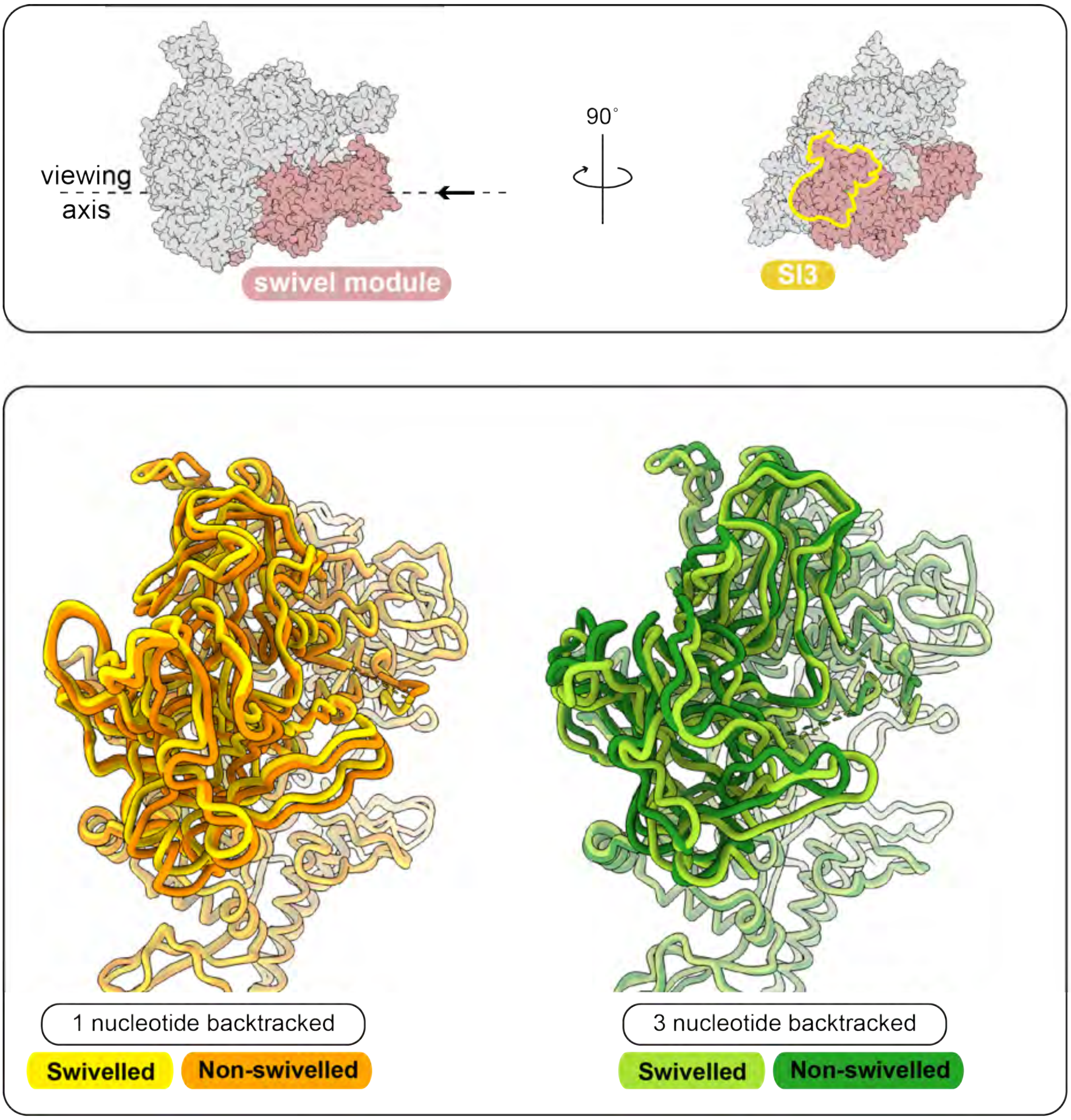


Figure 7.2 – Placement of SI3 domain in different complexes. The SI3 domains for each structure are in the foreground, with the rest of the swivel module in the background. The box on the top shows the viewing direction, which is along the arrow.

well as when a Gre factor binds to its SC.

Between the two 1 nt backtracked complexes (swivelled and non-swivelled), there was no difference seen in the positioning of the SI3 with respect to the rest of the swivel module. The same was seen for the 3 nt backtracked complex (Figure 7.2). However, there was a noticeable shift in the position of the SI3 when comparing the swivelled or non-swivelled 1 nt backtracked complex to the corresponding 3 nt backtracked complex (Figure 7.3(a)). What this showed was that for a specific backtracked length, the swivel module between the swivelled and non-swivelled states including the flexible SI3 domain rotated as a single unit. Once a change in the backtrack length is brought into the picture, the position of the SI3 with respect to the rest of the swivel module changes, but the change is consistent for the two states in the same backtrack length. The SI3 domain, which has been proposed to gate the SC for differentiation between binding of the Gre factors and the initiation factor DksA (Furman et al. 2013), likely also plays a role in gating the SC to select between GreA GreB for different backtracked lengths.

Figure 7.3(b-d) illustrates the difference in SI3 positioning when GreA and GreB are bound. Comparing the swivel modules for the BC with the GreA-PC (Figure 7.3(c)), we see that while the rest of the swivel module is aligned, there is a shift in the SI3 on its own. The specific shift in the domain appears to be towards the GreA CTD. The same is seen when comparing the GreB pre-cleavage complex to the 3nt BC (Figure 7.3(d)). Finally, a comparison of the positioning of the SI3 domains between the GreA and GreB pre-cleavage complexes is shown in Figure 7.3(b). Again, there is a difference in the positioning of the SI3 in each of the complexes. This is possibly a result of the difference in sequence of the carboxyl-terminal domain (CTD)s of these respective factors, leading to a difference in the interaction position with the SI3. While these structures do show that both the extent of backtracking and the binding of different cleavage factors influences the SI3 position, it would not be possible to comment on specific interactions between SI3 and the GreA and GreB CTDs owing to the fact that the density around the flexible SI3 domain is fairly weak and side chain positions cannot be modeled with high confidence.

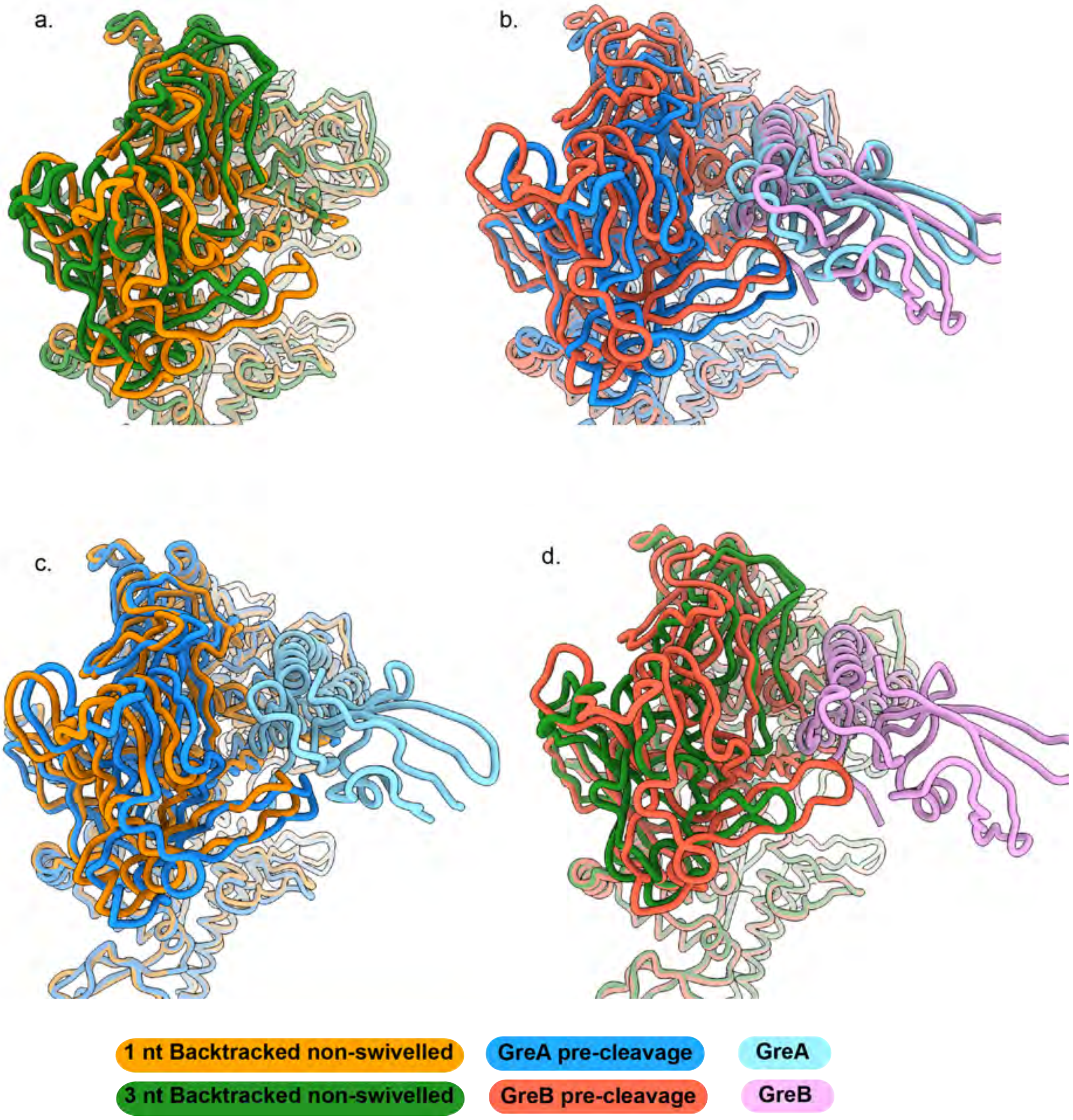

a.
b.
c.
d.
1 nt Backtracked non-swivelled
3 nt Backtracked non-swivelled
GreA pre-cleavage
GreB pre-cleavage
GreA
GreB

Figure 7.3 – Comparison of SI3 in different backtracked complexes, all aligned to the swivel module (excluding the SI3 domain). Figure 7.2 showed that for a single backtracked length, the entire swivel module including the SI3 rotated as a single unit. Here, we see that for different backtracked lengths with or without Gre (a. and b.), the rest of the swivel module is aligned but with a noticeable shift in the SI3.

c. and d. show that for a specific backtrack length, binding of GreA or GreB enduces a lateral shift of the SI3 towards the Gre CTD.

The most important point to note with respect to the observations about the SI3 positioning and its effects on the selection of GreA/B for different backtracked lengths is that the SI3 domain is *E. coli*-specific. Not all bacterial species' genomes encode for multiple Gre factors, and not all of them contain a sequence insertion within the trigger loop. Detailed sequence analyses of different bacterial RNAPs containing multiple cleavage factors will reveal if there is a link between the existence of two separate Gre factors and an SI3 or SI3-like insertion within a species. If there is a connection between the two, that could suggest that the evolution of two functionally different Gre factors within a species might have simply been the consequence of that particular species containing a sequence insertion that allowed it to have a higher degree of selectivity between structurally similar secondary channel-binding factors (SCBF)s. None of this, however, should take away from the significance of the observations on the effects of GreA on cleavage, discussed below. GreA is the transcription factor (TF) that is found universally in all bacterial species, and is extremely important for maintaining the fidelity of transcription. The primary influence of GreA binding to the secondary channel is on the universally conserved active site elements, with the secondary effect having to do with its interactions with the SI3.

7.3 GreA Influences Cleavage on Multiple Levels

GreA has always been known to promote intrinsic hydrolytic cleavage of backtracked RNA through interactions involving the two acidic residues at its tip (Asp 41 and Glu 44 in *E. coli*). If the influence of GreA on cleavage was to occur solely through interactions of these residues, mutating them would then result in a protein capable of binding to the RNAP SC but incapable of influencing the rate of cleavage. Contrary to this, the observations described here point towards the transcription factor affecting cleavage through interactions between RNAP and the GreA NTD and CTD, and by inducing an active site conformation that stimulates RNA cleavage compared to RNAP alone.

Evidence of GreA participating in proofreading in ways other than through its NTD tip was seen in the transcription assays, where mutating the two acidic residues at the tip to alanines (D41A E44A mutant) did not reduce cleavage activity back to intrinsic cleavage levels (with RNAP alone), but simply slowed it down in comparison with WT GreA. The fact that the GreA mutant was active and promoted cleavage, albeit at a reduced capacity, is an extremely fascinating observation, with possible implications in the understanding of factor-assisted transcriptional proofreading in other species. This is because up until this point, it had been all but established that the influence of GreA on cleavage was solely through its conserved acidic residues (Stebbins et al. 1995; Koulich, Nikiforov, and Borukhov 1998). That GreA is still active when these residues as mutated shows that it influences transcript cleavage through a secondary process, which is supported by the structural data obtained for the same complex. Since GreA is a universal cleavage factor in bacteria and since many of the amino-terminal domain (NTD) residues near the tip are conserved, this could have implications in understanding proofreading in other bacterial species. It might also point towards a similar understanding in other domains of life, however it is worth pointing out once again that the only major structural similarity that GreA shares with TFIIS and TFS in eukaryotes and archaea is that they interact with their respective RNAP active sites through conserved acidic residues. Apart from that, the structures of these factors are extremely dissimilar. If any similarities to the GreA pre-cleavage complex are found in Pol-II and archaea, that would certainly be a fascinating step in understanding the evolution of these molecular processes.

Following this, structures of the four pre-cleavage complexes showed a portion of the TL up to the invariant histdine (His 936) positioned in a very specific open conformation, distinct from the open conformations it was observed to be in in similar *E. coli* structures without GreA in the SC or with GreB bound in SC before or after cleavage. The proximity of the GreA NTD and the RNAP TL to one another did point towards GreA playing a direct role in positioning the TL. In particular, the phenylalanine residue within the TL (β' Phe 935) would be capable of forming a stacking pair with another Phe within GreA. Specific interactions of the TL with GreA would explain why we see the same conformation irrespective of the backtracked base, and also why the TL in the BCs without GreA appears to be more disordered. In addition to this, these interactions observed with the NTD might also help in understanding the difference between GreA and GreB interactions in the RNAP core. A noticeable difference between the NTDs in GreA and in GreB is in the size of the basic patches (**Koulich1997DomainFrom**). In GreB, a large basic patch allows for the factor to interact with and stabilise the long extruded RNA in the SC, thereby allowing for cleavage to take place and elongation to resume from a paused state. GreA acts on complexes backtracked by 1-2 nucleotides. In this case, a large basic patch would not be needed since the backtracked RNA isn't long enough to enter the SC and interact with GreA. Instead, residues within the GreA NTD function in stabilising the mobile TL, thus increasing the efficiency of cleavage by positioning the active site elements. The specific residues which are positioned near the TL in the GreA structure are absent in GreB. All of this helps in understanding why GreA and GreB have been shown to exhibit preferences for different cleavage product lengths.

The implications of the novel TL conformation become clear when we look at other key elements within the active site, specifically the backtracked RNA and the bridge helix (BH). Comparisons of the BC (swivelled and non-swivelled) and the GreA-PC (Figure 6.6) show that binding of GreA to the SC repositions active site elements. Density for the backtracked RNA base (RNA-17, position i+1) in the pre-cleavage complex is better resolved than in the backtracked complexes. It is worth pointing out that I am certain the complex is in a state in which no cleavage has occurred, since (i) the biochemical results showed that within the time it took me to freeze the complex on an EM grid, minimal cleavage would have taken place, and (ii) at the very least, I see a DNA-RNA base pair in the acceptor site (A-site) which otherwise would have only been an unpaired DNA had cleavage taken place, transitioning RNAP into the post-translocated state. In its pre-catalytic state, we see the mismatched base positioned outside of the A-site, ready for cleavage to take place. In the backtracked complexes, the backtracked RNA is extremely disordered, with almost no density observed for it in either of the conformational states. The BH, which is bent in all three structures, is shifted towards the A-site. The position adopted by the BH in the backtracked states would very clearly generate a steric clash with the backtracked RNA in its pre-catalytic state.

The TL is known to influence the positioning of the BH (Yuzenkova and Zenkin 2010). In its GreA-stabilised open conformation in the pre-catalytic state, the TL would allow for various elements in the active site to be positioned correctly for efficient hydrolysis of the extruded RNA, in agreement with its proposed role as a positional catalyst rather than a general acid-base catalyst (Mishanina et al. 2017). In the absence of a stabilising element, the TL would then be more disordered, preventing the active site elements from being positioned correctly. This explains why, in the cleavage assays reported here, intrinsic cleavage of the backtracked RNA in the absence of GreA was extremely inefficient in comparison with the rapid cleavage seen in the presence of GreA.

While the GreA NTD both directly and indirectly influences the conformations adopted by core elements of RNAP the CTD would likely influence the complex through its interactions with the surface. The shift in the position of the SI3 domain with respect to the swivel module when GreA is bound to the complex suggests binding between the GreA CTD and the SI3, which would in turn restrict the rotation of the swivel module, essentially locking it a swivelled state. This is consistent with data on GreB, which also requires SI3 for efficient binding to RNAP (Zakharova et al. 1998).

7.4 Towards a Complete Understanding of Proofreading

The structures that were solved allow us to address question of the role of the TL in hydrolysis. By analysing the effects of either deleting the entire TL or mutating the invariant histidine in *Thermus aquaticus*, Yuzenkova and Zenkin (2010) had concluded that the TL was essential for hydrolysis by positioning the mismatched 3'-nucleoside monophosphate (NMP) and by directly participating as a general base. In agreement with their conclusion that the TL is important for hydrolysis, we do see that it does function as a catalyst for hydrolysis, albeit indirectly as a positioning element. The invariant histidine is positioned away from the active site similar to the Pol-II structures (Wang et al. 2009), therefore even in bacteria, the TL would not be able to catalyse hydrolysis of backtracked RNA through His 936 acting as a general base. A mutation of the invariant histidine as described in the *T. aquaticus* paper would likely then disrupt the stability of the surrounding side chains, thereby indirectly impacting the function of the TL in positioning the elements required for cleavage.This also agrees with the results from Zhang, Palangat, and Landick (2010) that the formation of the trigger helices (TH) in *E. coli* RNAP is non-essential for hydrolysis.

The next level in understanding the elements governing mRNA cleavage is the role of GreA. Not just between GreA and GreB, but all the bacterial secondary channel-binding factors (SCBF) show a remarkable extent of structural similarities. Sequence alignments of these transcription factors also shows that most of them contain two acidic residues at the tips of their NTDs. If these SCBFs were to only function through their tip residues, the differences in the rest of their NTDs and their CTDs would primarily govern their binding to the complex. However, with GreA we see that apart from the acidic residues, its NTD coiled-coil also indirectly influences cleavage by positioning the TL to allow for efficient hydrolysis to take place.

The differences in swivelling ranges for different backtracked lengths is the key to understanding how RNAP selects for GreA over GreB when backtracked by 1 nt. Backtracking resulting from a single mismatch allows the complex to adopt a more swivelled state than what has been seen for a similar complex backtracked by more nucleotides. When in a more swivelled conformation, the SI3 domain is positioned closer to the β lobe which forms the other half of the pincers of the RNAP claw. This could restrict entry into the SC, thus allowing only specific proteins to enter and bind to the SC. The GreA CTD is known to be crucial for binding of the factor to RNAP (Koulich, Nikiforov, and Borukhov 1998).

In summary, proofreading likely takes place as follows: RNAP backtracked by 1 nt as a consequence of a misincorporation and the resulting 3'-mismatch would be able to adopt more extreme swivelled conformations compared to longer backtracks (or active

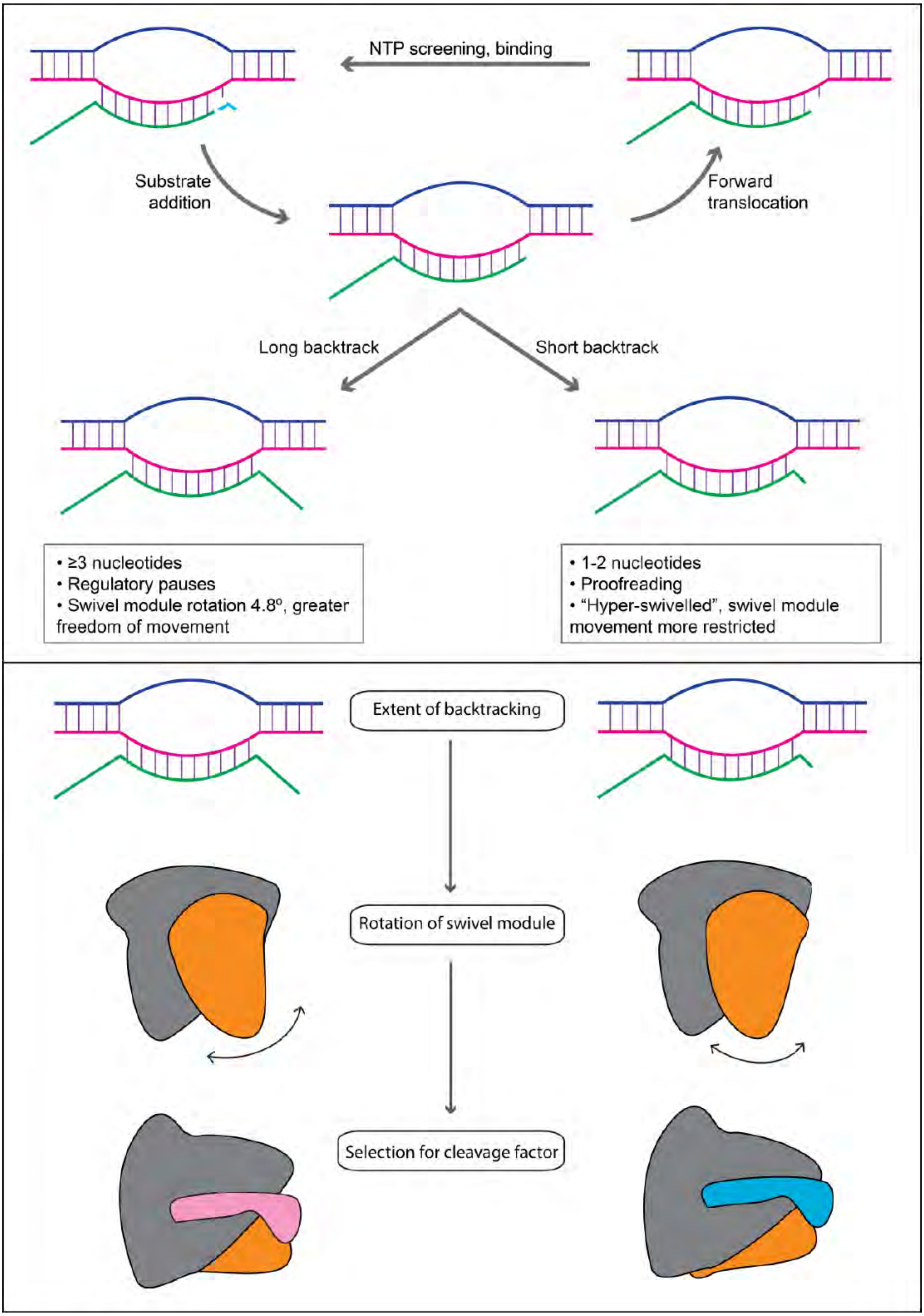


Figure 7.4 – Model for backtrack-dependent factor recruitment. The top half shows the two distinct types of backtracking that can take place during the nucleotide addition cycle (NAC). The bottom half illustrates how the extent of backtracking influences the allowed range of motion of the swivel module (and, by extension, the SI3 domain), which in turn allows for selection of one cleavage factor over another

elongating RNAP). A longer backtrack resulting from the elongation complex (EC) entering an arrested state would be threaded further into the SC, restricting the movement of the swivel module to less swivelled states. Having a long or short backtracked RNA in the SC would also influence the freedom of movement of the TL, and as an extension the SI3 domain, pushing it further away from the β lobe in the case of a longer backtrack, and allowing the SI3 domain to be pulled in towards to the β lobe when the extruded RNA is shorter. The different positions adopted by the SI3 depending on the backtracked length might influence which Gre factor is more likely to bind to the complex. GreA, which is required for rescue of shorter backtracked complexes, would then preferentially bind to the complex and lock it in a fixed swivelled state, restricting elements within the RNAP core to more specific positions that would promote efficient hydrolysis. In addition to this, the GreA NTD coiled-coil also interacts directly with the TL keeping it in a specific open conformation. Without fixing the TL in place, its inherent flexibility in the unfolded state might push the BH to a position that would clash with the extruded RNA. Finally, we see that the backtracked base likely does participate in its own cleavage, owing to its more fixed and less disordered position in the pre-cleavage complex and reduced transcription efficiency on introduction of modifications to the phosphate backbone. However, this interaction would take place solely through the backbone and would not be affected by the nature of the backtracked base itself.

Further experiments would be useful to more conclusively characterise the effects of GreA in proofreading. First, more extensive transcription assays, in particular those comparing the effects of GreA versus GreB on the cleavage of complexes backtracked by long and shorter RNA. Binding assays would also allow us to isolate the specific interactions between the SI3 and GreA CTD. These interactions could not be determined structurally since the SI3 domain, although resolved well enough to determine its position, is still too flexible for many of its side chains to be modelled confidently.

Chapter 8

Publications

Structural Basis of Transcription-Translation Coupling and Collision in Bacteria

In addition to the work on proofreading, I was also able to participate in another project within the lab between November 2019 and January 2020, concerning transcription-translation coupling in *E. coli*.

As discussed in Section 1.2, the idea that the two key processes within gene expression could happen in tandem with one another has been around for decades (Miller, Hamkalo, and Thomas 1970; Byrne et al. 1964). Coupling of the two gene expression machineries is important for multiple reasons: the trailing ribosome affects transcription by releasing elongation complexes from stalled states and increasing the rate of transcription (Proshkin et al. 2010), it influences gene regulation, and can prevent premature termination. The universal transcription factor NusG was proposed to play a role in coupling through specific contacts with RNA Polymerase (RNAP) and the ribosome by Burmann et al. (2010). However, in the case of the ribosome and RNAP directly contacting each other (Kohler et al. 2017), NusG would be incapable of binding to the complex. Close contacts between RNAP and ribosomes were suggested as early as 1970 (Miller, Hamkalo, and Thomas 1970). The complex from Kohler et al. (2017) that suggested direct interactions was formed by colliding the ribosome with a stalled RNAP. A number of questions needed addressing:

- Did the complex from the Kohler paper represent a functional state?
- Is NusG able to couple the two machineries?
- What role does NusG play in coupling?
- Do direct contacts between the ribosome and RNAP play an important role in the coupled processes?

Structures of a transcription-translation coupled complex (expressome) in three states were used to address the questions of how translation influences transcription and the role that NusG plays in this process. The three states were of the expressome in an uncoupled state, in a NusG-coupled state, and in a collided state.

The uncoupled expressome showed a broad range of movement (both translational and rotational) of RNAP relative to the ribosome, which when plotted revealed clusters of enriched orientations. Density for the NusG NTD was observed in one conformational cluster in which the binding sites of the NusG CTD and NTD on the ribosome and RNAP respectively were closest to each other. Increasing the occupancy of NusG in the sample led to the second NusG-coupled expressome structure. Different orientations of RNAP relative to the ribosome were also seen for this complex, albeit with more restricted movement compared with the uncoupled complex. Coupling through NusG, aligns the nascent mRNA with the ribosomal helicase thereby ensuring single stranded RNA entering the ribosome and facilitating smooth transcription and translation. Binding of the NusG CTD to the ribosomal protein uS10 might explain how coupling of translation to transcription would reduce the likelihood of transcription being prematurely terminated, thereby favouring elongation. Shortening the length of the mRNA spanning the gap between the ribosome and RNAP led to a complex in which the mRNA entry channel on the ribosome and the RNA exit channel ofRNAP were aligned to one another (collided expressome). In this collided state, binding of the NusG NTD to RNAP was still observed, but the CTD would be unable to contact the ribosome. The positioning of RNAP relative to the ribosome was significantly different from what was seen in the coupled expressome, and was not dependent on NusG. In the collided state RNAP was restrained in its movement relative to the ribosome but still flexible suggesting no stable interactions take place. Negative stain EM and biochemical experiments of the sample without any nucleic acid scaffolds showed that stable formation of a supramolecular complex requires mRNA, and that ribosome-RNAP contacts are likely not as important for complex formation as previously thought.

The different structures showed that the ribosome and RNAP are linked by the mRNA. Even when coupled to one another, the two units show a wide range of freedom of movement relative of one another, which would be important considering that they generally operate independently. In a coupled state, RNAP would not inhibit binding of translation factors. Conversely, transcription factors which do not require the NusG binding site would also appear to be largely unaffected by coupling.

Contributions

My contribution to this paper was in the atomic modelling of some of the structures, in particular those of the NusG-coupled expressome and of the collided expressome.

STRUCTURAL BIOLOGY

Structural basis of transcription-translation coupling and collision in bacteria

Michael William Webster[1,2,3,4]*, Maria Takacs[1,2,3,4]*, Chengjin Zhu[1,2,3,4], Vita Vidmar[1,2,3,4], Ayesha Eduljee[1,2,3,4], Mo'men Abdelkareem[1,2,3,4], Albert Weixlbaumer[1,2,3,4]†

Prokaryotic messenger RNAs (mRNAs) are translated as they are transcribed. The lead ribosome potentially contacts RNA polymerase (RNAP) and forms a supramolecular complex known as the expressome. The basis of expressome assembly and its consequences for transcription and translation are poorly understood. Here, we present a series of structures representing uncoupled, coupled, and collided expressome states determined by cryo–electron microscopy. A bridge between the ribosome and RNAP can be formed by the transcription factor NusG, which stabilizes an otherwise-variable interaction interface. Shortening of the intervening mRNA causes a substantial rearrangement that aligns the ribosome entrance channel to the RNAP exit channel. In this collided complex, NusG linkage is no longer possible. These structures reveal mechanisms of coordination between transcription and translation and provide a framework for future study.

All organisms express genetic information in two steps. mRNAs are transcribed from DNA by RNA polymerase (RNAP) and then translated by ribosomes to proteins. In prokaryotes, translation begins as the mRNA is synthesized, and the lead ribosome on an mRNA is spatially close to RNAP (*1*, *2*). Coordination of transcription with translation regulates gene expression and prevents premature transcription termination (*3*, *4*). The trailing ribosome inhibits RNAP backtracking, which contributes to the synchronization of transcription and translation rates in vivo and in vitro (*5*–*7*).

Coordination may also involve physical contacts between RNAP and the ribosome. The conserved transcription factor NusG binds RNAP through its N-terminal domain (NusG-NTD) and binds ribosomal protein uS10 through its C-terminal domain (NusG-CTD) both in vitro and in vivo (*8*, *9*). Formation of a NusG-mediated bridge by simultaneous binding has not yet been observed, and the consequences of physical coupling are unknown. RNAP and the ribosome also interact directly (*10*–*12*), and this complex has recently been visualized in situ (*13*). A transcribing-translating expressome complex formed by the collision of ribosomes with stalled RNAP in an in vitro translation reaction was reconstructed at 7.6-Å resolution (*10*). This architecture would not permit a NusG-mediated bridge.

We sought to structurally characterize mechanisms of physical transcription-translation coupling and resolve the relationship between NusG and the collided expressome. Expressomes were assembled by the sequential addition of purified *Escherichia coli* components (70*S* ribosomes, tRNAs, RNAP, and NusG) to a synthetic DNA-mRNA scaffold (fig. S1, A to C). An mRNA with 38 nucleotides separating the RNAP active site from the ribosomal P-site was chosen to imitate a state expected to precede collision (*14*).

A reconstruction of the expressome was obtained at 3.0-Å nominal resolution by cryo–electron microscopy (cryo-EM) (Fig. 1A; fig. S1, D and E; and table S1). RNAP and the ribosome do not adopt a single relative orientation within the expressome, and focused refinement was required to attain a reconstruction of RNAP at 3.8-Å nominal resolution (Fig. 1A and fig. S2; see materials and methods). Refined atomic models collectively present the key steps of prokaryotic gene expression in a single molecular assembly (Fig. 1B).

Direct contacts between RNAP and the ribosome, if they occur, are not stable in this complex, and the mRNA is the only consistent connection. We characterized the dynamics of the complex by plotting the range of RNAP positions relative to the ribosome using the angular assignments of particles from focused reconstructions (Fig. 1C and fig. S3A). RNAP is loosely restrained to a plane perpendicular to an axis connecting the RNAP mRNA exit channel to the ribosomal mRNA entrance channel (movie S1). Within this plane, RNAP rotates freely. Seven clusters represent a series of preferred relative orientations (Fig. 1C and fig. S3A).

RNAP and ribosome models were placed in reconstructions generated from particles in clusters 1 to 6, but a large fraction of cluster 7 was predicted to be incompatible with longer upstream DNA (fig. S3, B to F, and table S2; see materials and methods). Expressome models represent characteristic relative orientations for each cluster, and they collectively suggest a continuous movement of RNAP relative to the ribosome surface involving substantial changes in both rotation (~280°) and translation (~50 Å) (Fig. 1D and movie S1). The closest domain of RNAP to the ribosome is the zinc finger of the β′ subunit (β′-ZF) in all models. In clusters 1 to 3, the β′-ZF sits within a funnel-shaped depression between the head, body, and shoulder domains of the 30*S* subunit, bounded by ribosomal proteins uS3, uS4, and uS5. We predict that RNAP transits from cluster 1 through clusters 2 to 5 to reach positions exemplified by model 6, where the RNAP β′-ZF is between uS3 and uS10 on the 30*S* head domain.

NusG-NTD is bound to RNAP in expressome cluster 6 but not in clusters 1 and 2 (Fig. 1E). We determined that a substantial fraction of the imaged particles lacked NusG because of dissociation during gradient purification (fig. S3G). Notably, the predicted position of the NusG-CTD bound to uS10 (*8*, *9*) is closest to the NusG-NTD bound to RNAP in cluster 6.

An improved reconstruction of the NusG-coupled expressome was obtained from a sample prepared with increased NusG occupancy (Fig. 2A and fig. S4, A and B; see materials and methods). Conformational heterogeneity of the ribosome and RNAP was substantially reduced, but focused refinement was required to obtain well-resolved ribosome and RNAP reconstructions (3.4 and 7.6 Å, respectively) (fig. S4, C to E, and table S1). Continuous density in the unfocused reconstruction confirmed that NusG bridges RNAP and the ribosome (Fig. 2A). We constructed an atomic model of the NusG-coupled expressome by fitting and refining a ribosome model and docking a published RNAP-NusG-NTD model consistent with our map (*15*) into their consensus positions in the unfocused reconstruction (Fig. 2B).

Additional density corresponding to the NusG-CTD bound to uS10 was identified on the ribosome, which otherwise closely resembled that of the uncoupled expressome. The NusG-CTD is a KOW (Kyrpides, Ouzounis, and Woese) domain that consists of a five-stranded β barrel. As in the isolated NusG-uS10 complex, as determined by nuclear magnetic resonance (NMR) (*8*), strand β4 of NusG aligns with strand β4 of uS10, thereby forming an extended intermolecular β sheet (Fig. 2C). However, NusG and uS10 are substantially closer in the expressome than they are in the isolated complex because NusG loops L1 (F141 and F144) and L2 (I164, F165, and R167) insert into a hydrophobic pocket of uS10 that is enlarged by movement of helix α2 (Fig. 2D and fig. S5, A to D). F165 of NusG, in particular, is embedded within uS10. This accounts for its key role in binding uS10, which has been identified by mutational studies (*9*). The altered position of NusG not only increases the area contacting uS10 but avoids clashing with neighboring ribosomal protein uS3 (Fig. 2D).

[1] Department of Integrated Structural Biology, Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC), 67404 Illkirch, France. [2] Université de Strasbourg, 67404 Illkirch, France. [3] CNRS UMR7104, 67404 Illkirch, France. [4] INSERM U1258, 67404 Illkirch, France.
*These authors contributed equally to this work.
†Corresponding author. Email: albert.weixlbaumer@igbmc.fr

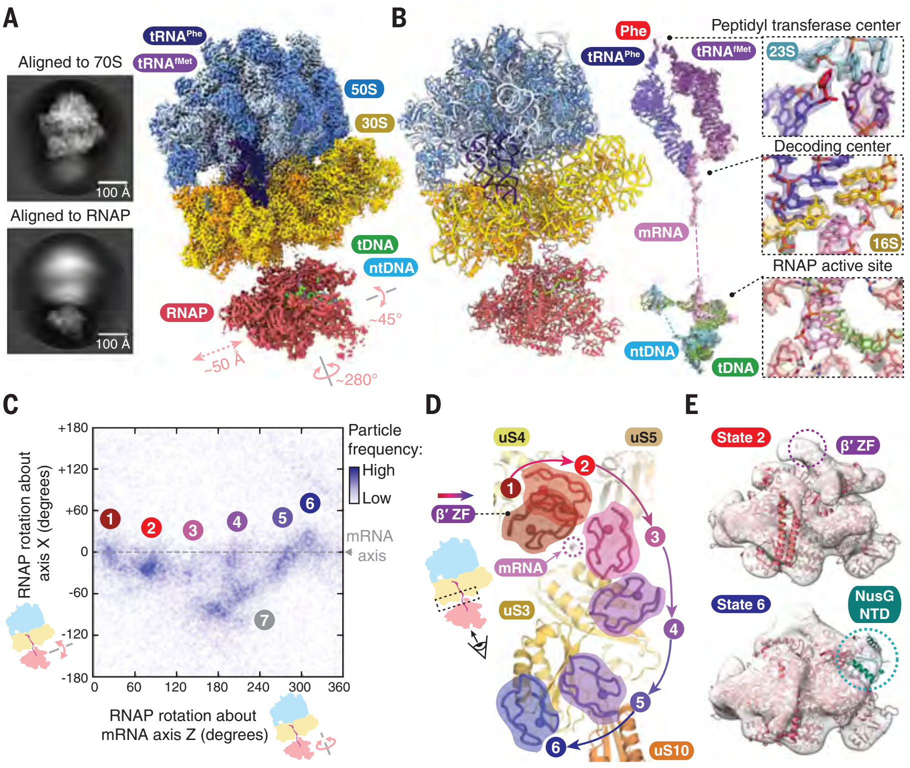


Fig. 1. Structural models of the uncoupled expressome. (**A**) Representative cryo-EM two-dimensional class averages showing conformational variability (left), and cryo-EM maps of the ribosome and RNAP in the uncoupled expressome (right). RNAP is shown in position 2 [see (E)], with measured rotation and translations of RNAP indicated. tDNA, template DNA; ntDNA, nontemplate DNA. (**B**) Atomic model of the uncoupled expressome in ribbon representation (left), and the central steps in gene expression shown by segmented cryo-EM maps with superimposed atomic coordinates (right). (**C**) Plot of RNAP-70*S* relative orientation with clusters indicating a series of orientations (1 to 6) distinguished by rotation of RNAP. Further characterization of expressome particles resembling cluster 6 (Fig. 2) revealed that these are likely physically coupled through NusG. Cluster 7 primarily includes particles with orientations incompatible with longer upstream DNA, but it also includes states that have been characterized by Wang *et al.* (*26*). (**D**) Representative positions of the RNAP β′-ZF in each expressome model relative to the ribosome surface. (**E**) NusG is present in state 6 (dashed green circle) but not in state 2. The position of β′-ZF is shown (dashed purple circle). The focused cryo-EM maps shown are filtered to 20-Å resolution with fitted coordinates.

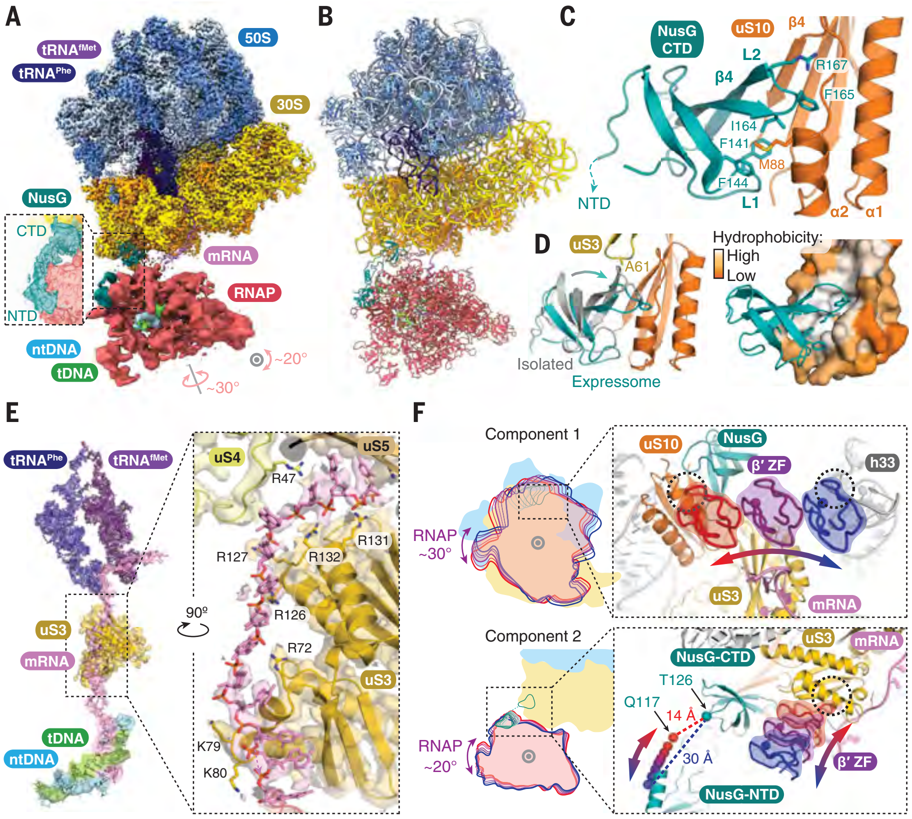


Fig. 2. Structural models of the NusG-coupled expressome. (**A**) Focused cryo-EM maps of the ribosome and RNAP in the NusG-coupled expressome. Inset shows continuous electron density between NusG-NTD and NusG-CTD domains in an unfocused map filtered to 8 Å (slice view). (**B**) Ribbon representations of the NusG-coupled expressome model. (**C**) Interaction of NusG-CTD with ribosomal protein uS10. (**D**) Structural superposition with the isolated NusG-uS10 complex based on alignment to uS10 (gray; PDB code 2KVQ) (left) and hydrophobic pocket created by conformational change of uS10 (right). (**E**) mRNA connecting the ribosome mRNA entrance channel to the RNAP exit channel shown by a cryo-EM map filtered to 4 Å and fitted model. (**F**) The range of RNAP positions relative to the ribosome surface determined by multi-body refinement. Cartoon of two principal components accounting for 44% of variance (left). Component 1 involves rotation in a plane approximately parallel to the surface of the ribosome and is limited by clashes between the β′-ZF of RNAP and either uS10 or h33 (dashed circles). Component 2 is an orthogonal rotation limited by extension of the flexible NusG linker (residues 117 to 126) in one direction (red through purple to blue arrows) and by the clash between β′-ZF and uS3 in the other (dashed circle). Positions of RNAP β′-ZF and NusG residue Q117 indicate trajectories (red through purple to blue arrows). Single-letter abbreviations for the amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr.

The NusG-CTD recruits Rho to terminate the synthesis of untranslated mRNAs (*16*). In the NusG-coupled expressome, NusG binds uS10 with the same interface it binds Rho, which suggests that the events are mutually exclusive (fig. S5E) (*17*). The structure of the expressome thereby explains how the trailing ribosome is sensed by NusG and how transcription termination is consequently reduced.

The binding of the NusG-NTD to RNAP suppresses backtracking by stabilizing the upstream DNA duplex (*15*, *18*). In the expressome, space for the upstream DNA is further restricted by an extended channel formed by uS10 and NusG (fig. S5F). The interaction of the NusG-CTD with uS10 is predicted to reduce dissociation of NusG-NTD from RNAP through increased avidity (*19*). The RNAP-NusG complex within the coupled expressome is likely stabilized by the trailing ribosome, and transcription elongation is consequently favored.

The mRNA exit channel of RNAP is separated from the entrance channel of the ribosome by ~60 Å. Continuous electron density on the solvent side of uS3 allowed the modeling of the intervening 12 mRNA nucleotides, which completed the mRNA path from synthesis to decoding (Fig. 2E and fig. S6, A to C). The interpretability of the electron density varies considerably, however, and this model is considered one of an ensemble of mRNA conformations.

The RNAP mRNA exit channel is adjacent to uS3 residues R72, K79, and K80, and clear electron density for mRNA in this region suggests a relatively stable contact. The path continues to four arginines immediately outside the ribosomal mRNA entrance channel (R126, R127, R131, and R132) (fig. S6A). R131 and R132 have been previously identified as imparting ribosomal helicase activity (*20*). The mRNA path in this region is close to, but different from, that observed previously in structures of mRNA-bound ribosomes (*21*) (fig. S6, D to F).

Binding of the nascent transcript by uS3 likely modulates secondary structure formation. Structured mRNAs can decrease translation rates (*22*), stabilize transcriptional pauses [e.g., the *E. coli his* pause (*23*)], or induce transcription termination (*24*). Although the ribosome can unwind mRNA secondary structure with basic residues in the mRNA entrance channel (*20*), preventing mRNAs folding downstream likely aids translation efficiency. We propose that by positioning RNAP in line with an extended series of basic residues, NusG helps keep nascent mRNAs single stranded and thereby enhances the efficiency of both transcription and translation.

No stable contacts are observed between the core subunits of RNAP and the ribosome in the NusG-coupled expressome. The relative position of RNAP and the ribosome varies between particles, albeit substantially less than the sample with partial NusG occupancy (fig. S4, A and B). Analysis of movement by multibody refinement (*25*) reveals that RNAP is constrained to avoid clashes between β′-ZF and the cavity formed by uS3, uS10, NusG, and helix 33 of 16*S* rRNA (h33) into which it is inserted (Fig. 2F and movie S2). RNAP is also flexibly tethered to the ribosome by NusG, with the length of the NusG linker (residues 117 to 125) varying in the range of 14 to 30 Å.

To test whether lengthening the intervening mRNA alters the architecture of the expressome, we imaged two samples with four additional mRNA nucleotides (42 in total) separating the RNAP active site from the ribosomal P-site (fig. S7A and table S1). Saturation with NusG increased particle frequencies resembling the NusG-coupled expressome with shorter mRNA, including density linking the complexes (fig. S7, B to E). Compared with shorter mRNA, more particles are observed arranged similarly to cluster 7 of the uncoupled expressome (Fig. 1C). This arrangement is termed transcription-translation complex C (TTC-C) by Wang *et al.* (*26*). However, NusG-CTD is bound to uS10 only in cluster 6 but not cluster 7, which indicates that NusG couples in only one arrangement (fig. S7F).

The mRNA spanning the mRNA exit and entrance channels is in an extended conformation in the NusG-coupled expressome. To test whether coupling by NusG is possible when the spanning mRNA is shorter, we obtained a reconstruction of a NusG-containing expressome with an mRNA shortened to 34 nucleotides between the ribosomal P-site and the RNAP active site (Fig. 3A, fig. S8, and table S1). A model was constructed as described for the coupled expressome (Fig. 3B).

In this model, RNAP is positioned close to the ribosome mRNA entrance channel—more than 50 Å from its location in the NusG-coupled expressome. Consistent with this change, RNAP still binds the NusG-NTD but is no longer tethered through the NusG-CTD to uS10 because the NusG linker (residues 117 to 125; maximum extension of ~30 Å) would need to span an 85- to 145-Å distance. We determined the structure of an equivalent sample lacking NusG and confirmed that the position of RNAP is very similar in this case (fig. S9A and table S1). Therefore, the architecture is not NusG-dependent and is similar to particles from clusters 1 and 2 of the uncoupled expressome (fig. S12). We conclude that RNAP coupling to the ribosome through NusG requires the P-site to be >34 nucleotides from the 3′ end of the mRNA.

The rearrangement of RNAP and the ribosome in our structure with shortened mRNA resembles the expressome formed by the collision of translating ribosomes with stalled RNAP [RNAP backbone root mean square deviation (RMSD) ~3 Å based on 16*S* rRNA superposition] (*10*) (fig. S10, A and B, and fig. S12). We therefore refer to this molecular state as the collided expressome. The previous reconstruction was resolved to 7.6 Å, and

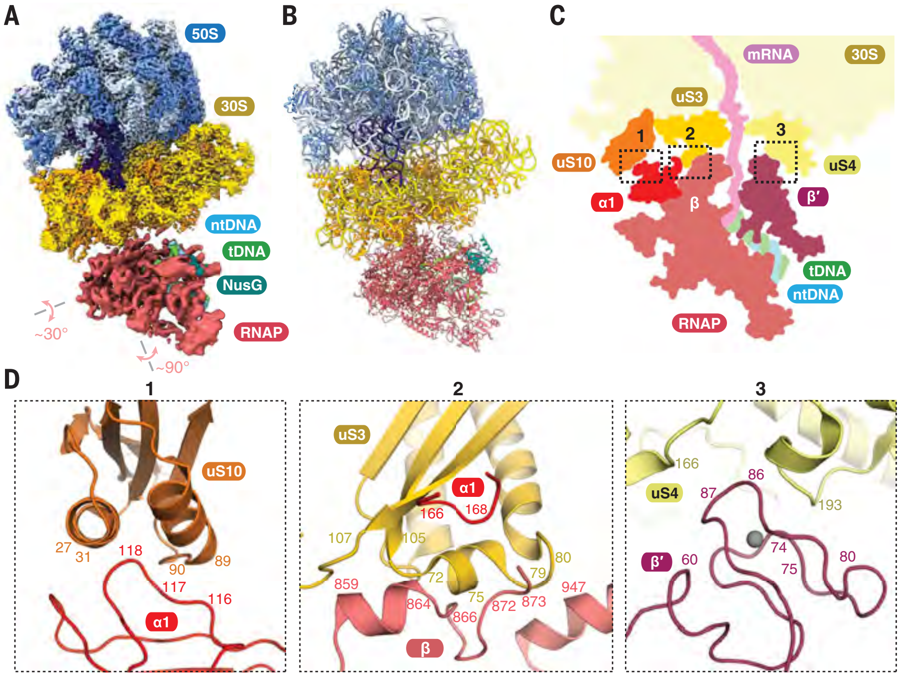


Fig. 3. Structural models of the collided expressome. (**A** and **B**) Cryo-EM map and model of the collided expressome. (**C**) Schematic cross section indicating three regions of close contact between RNAP and ribosome (indicated by dashed rectangles). (**D**) Details of the interaction interfaces of RNAP with the ribosome. Rectangles 1 to 3 correspond to the dashed rectangles in (C).

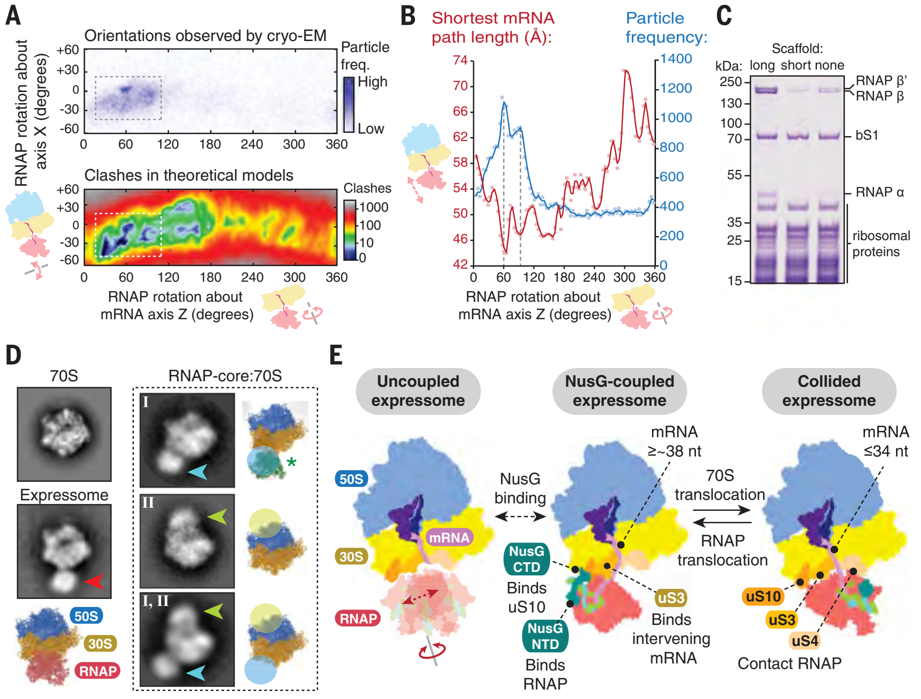


Fig. 4. Formation of RNAP-70*S* complexes. (**A**) Collided expressome RNAP-70*S* relative orientations observed by cryo-EM (top) correspond to a restricted space that avoids steric clashes (bottom). freq., frequency. (**B**) The most-common RNAP positions in the collided expressome (blue line) coincide with the minima of the intervening mRNA path length (red line). (**C**) Gradient copurification of RNAP with 70*S* ribosomes depends on the nucleic acid scaffold. RNAP-70*S* complexes were formed under low-salt conditions with an mRNA long enough to allow ribosome binding (long), or not long enough to allow ribosome binding (short), or without nucleic acids (none). Coomassie-stained SDS–polyacrylamide gel electrophoresis (SDS-PAGE) of the ribosome-containing sucrose gradient peak is shown. (**D**) Negative stain EM class averages of 70*S*-RNAP complexes show distinct binding sites for a core RNAP sample (cyan and lime arrowheads) compared with an expressome sample (red arrowhead). The position of RNAP from the 30*S*-RNAP complex is superimposed (green asterisk). (**E**) Key features and interchange between expressome complexes during transcription-translation coordination. In the uncoupled expressome, RNAP is loosely restrained and adopts various orientations. Coupling by NusG aligns the mRNA with ribosomal protein uS3 and restricts the position of RNAP. Once the ribosome approaches RNAP further, the collided state forms in which the mRNA length is limiting and NusG no longer links the two machineries. nt, nucleotide.

our improved model allowed us to define the interaction surfaces of RNAP and the ribosome in even more detail.

Four regions are in close proximity: uS10 with the NTD of the RNAP α1 subunit, uS3 with RNAP subunits α1 and the β-flap domain, uS4 with β′-ZF, and uS2 with the RNAP ω subunit (Fig. 3, C and D, and fig. S9, B and C). However, density for the ω subunit is very weak, which suggests that partial or complete dissociation occurs upon collision. The contacts bury a total surface area of ~3000 Å^2. However, RNAP moves relative to the ribosome, albeit less than in the samples previously analyzed (fig. S8, B and C). The RNAP-ribosome contacts are likely transient, so the contact area varies. The observed RNAP-ribosome configuration allows notable structural complementarity between the molecular surfaces.

Rotation of RNAP relative to the ribosome beyond the observed position would cause steric clashes (Fig. 4A and fig. S11). We hypothesize that the architecture of the collided expressome is the product of structural complementarity and the energetically favorable minimization of mRNA path length. To test this, we generated ~18,000 hypothetical expressome models representing an exhaustive search of RNAP rotations located about the mRNA axis at a series of distances along it (2° rotational step size, 0.5-Å translational step size). After excluding clashing models, we found that the shortest mRNA path is achieved by the RNAP orientations observed by cryo-EM (Fig. 4B). A simple model is therefore sufficient to explain the observed orientation of RNAP relative to the ribosome: When inserting into the mRNA entrance channel cavity on the ribosome, RNAP adopts an orientation with structural complementarity so that the intervening mRNA spans the shortest distance.

We sought to clarify whether expressome formation is driven by concurrent binding to the same mRNA or whether specific ribosome-RNAP contacts contribute. Copurification of RNAP with ribosomes was substantially reduced when the mRNA did not support concurrent ribosome binding, but RNAP that lacked DNA or mRNA entirely (RNAP-core) bound ribosomes more stably (Fig. 4C and fig. S10, C and D). This observation was previously thought to indicate that expressome formation is not driven by shared mRNA (*10*, *12*).

To examine this, we imaged samples assembled without further purification and lacking nucleic acid scaffolds (RNAP-core-70*S*) by negative stain electron microscopy (EM). No expressomes formed, which suggests that their formation is driven by concurrent mRNA binding and that direct interactions play minor roles. However, we observed at least two alternative RNAP binding sites (Fig. 4D). The sites can be described only approximately from this data, but one (site I) is consistent with an interaction with ribosomal protein uS2 observed in a core RNAP-30*S* complex (*11*). Saturation of ribosomes with ribosomal protein bS1, which has no effect on expressome formation (fig. S13A), abolished the occupancy of site I without affecting the second site (site II). The addition of a nucleic acid scaffold containing just a short mRNA (minimal scaffold) abolished occupancy of site II only, whereas addition of both (short mRNA scaffold and bS1) abolished both (fig. S13). A potential biological role has yet to be shown, but the existence of additional 70*S*-RNAP contact modes highlights the complexity of their interaction.

Thus, the expressome is mRNA-linked and consequently dynamic. A level of structural independence may be required to accommodate internal movements that occur during the reaction cycle of each complex. Coupling by NusG restrains RNAP motions—and happens at variable RNAP ribosome distances (fig. S12)—but not when they collide (Fig. 4E). Relative orientations of the two machineries change in prevalence as a function of their separation (fig. S12). Notably, translation factor binding is compatible with all the observed RNAP orientations. The role of the presented structures in vivo remains to be investigated, and this study provides a basis for elucidating the role of coupling in gene expression, and its regulation by transcription factors and regulatory mRNA structures.

REFERENCES AND NOTES

1. R. Byrne, J. G. Levin, H. A. Bladen, M. W. Nirenberg, *Proc. Natl. Acad. Sci. U.S.A.* **52**, 140–148 (1964).
2. O. L. Miller Jr., B. A. Hamkalo, C. A. Thomas Jr., *Science* **169**, 392–395 (1970).
3. C. Yanofsky, *Nature* **289**, 751–758 (1981).
4. J. P. Richardson, *Cell* **64**, 1047–1049 (1991).
5. S. Proshkin, A. R. Rahmouni, A. Mironov, E. Nudler, *Science* **328**, 504–508 (2010).
6. M. Zhu, M. Mori, T. Hwa, X. Dai, *Nat. Microbiol.* **4**, 2347–2356 (2019).

7. F. Stevenson-Jones, J. Woodgate, D. Castro-Roa, N. Zenkin, *Proc. Natl. Acad. Sci. U.S.A.* **117**, 8462–8467 (2020).
8. B. M. Burmann *et al*., *Science* **328**, 501–504 (2010).
9. S. Saxena *et al*., *Mol. Microbiol.* **108**, 495–504 (2018).
10. R. Kohler, R. A. Mooney, D. J. Mills, R. Landick, P. Cramer, *Science* **356**, 194–197 (2017).
11. G. Demo *et al*., *eLife* **6**, e28560 (2017).
12. H. Fan *et al*., *Nucleic Acids Res.* **45**, 11043–11055 (2017).
13. F. J. O'Reilly *et al*., *Science* **369**, 554–557 (2020).
14. D. Castro-Roa, N. Zenkin, *Nucleic Acids Res.* **40**, e45 (2012).
15. J. Y. Kang *et al*., *Cell* **173**, 1650–1662.e14 (2018).
16. S. L. Sullivan, M. E. Gottesman, *Cell* **68**, 989–994 (1992).
17. M. R. Lawson *et al*., *Mol. Cell* **71**, 911–922.e4 (2018).
18. M. Turtola, G. A. Belogurov, *eLife* **5**, e18096 (2016).
19. G. Vauquelin, S. J. Charlton, *Br. J. Pharmacol.* **168**, 1771–1785 (2013).
20. S. Takyar, R. P. Hickerson, H. F. Noller, *Cell* **120**, 49–58 (2005).
21. H. Amiri, H. F. Noller, *RNA* **25**, 364–375 (2019).
22. X. Qu *et al*., *Nature* **475**, 118–121 (2011).
23. C. L. Chan, R. Landick, *J. Biol. Chem.* **264**, 20796–20804 (1989).
24. I. Gusarov, E. Nudler, *Mol. Cell* **3**, 495–504 (1999).
25. T. Nakane, D. Kimanius, E. Lindahl, S. H. Scheres, *eLife* **7**, e36861 (2018).
26. C. Wang *et al*., *Science* **369**, 1359–1365 (2020).

ACKNOWLEDGMENTS

We thank J. Ortiz, C. Crucifix, X. Guo, and T. C. Cheng for help with data collection at the IGBMC. We thank W. Hagen and F. Weis for help with data collection at the EMBL in Heidelberg, Germany. We thank V. Ramakrishnan and A. C. Kelley for their valuable help in ribosome and tRNA purification. This work has been supported by iNEXT PID 6979, funded by the Horizon 2020 program of the European Union. We acknowledge the European Synchrotron Radiation Facility for the provision of microscope time on CM01, and we thank G. Effantin and M. Hons for their assistance. We thank members of the Weixlbaumer laboratory for critical reading of the manuscript. **Funding:** The authors were supported by the French Infrastructure for Integrated Structural Biology (FRISBI ANR-10-INBS-05, Instruct-ERIC, and grant ANR-10-LABX-0030-INRT, a French State fund managed by the Agence Nationale de la Recherche under the program Investissements d'Avenir ANR-10-IDEX-0002-02). The work was supported by an EMBO long-term fellowship to M.W.W. and the European Research Council starting grant TRANSREG (679734) to A.W. **Author contributions:** M.W.W. and M.T. performed experiments, binding assays, electron microscopy, and data analysis. C.Z. participated in data analysis and purification of NusG. M.W.W., M.T., V.V., A.E., M.A., and A.W. built and refined atomic models. A.W. designed and supervised research. M.W.W., M.T., and A.W. prepared the manuscript with input from all authors. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** Electron density maps for uncoupled, coupled, and collided expressomes (with and without NusG) were deposited in the Electron Microscopy Data Bank (EMD-11418, EMD-11419, EMD-11420, EMD-11421, EMD-11422, EMD-11423, and EMD-11426). Refined coordinates were deposited in the Protein Data Bank database under accession codes 6ZTN (NusG-coupled expressome with RNA-42); 6ZTO, 6ZU1, and 6ZTP (models for uncoupled expressome clusters 1 and 2 and NusG-coupled expressome cluster 6 with RNA-38); 6ZTJ (NusG-coupled expressome with RNA-38); 6ZTL (collided expressome with NusG); and 6ZTM (collided expressome without NusG). Materials are available from the authors on request.

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/369/6509/1355/suppl/DC1
Materials and Methods
Figs. S1 to S13
Tables S1 and S2
References (*27–58*)
MDAR Reproducibility Checklist
Movies S1 to S3

View/request a protocol for this paper from *Bio-protocol*.

28 February 2020; accepted 17 July 2020
Published online 20 August 2020
10.1126/science.abb5036

Science

Structural basis of transcription-translation coupling and collision in bacteria

Michael William Webster, Maria Takacs, Chengjin Zhu, Vita Vidmar, Ayesha Eduljee, Mo'men Abdelkareem and Albert Weixlbaumer

Science **369** (6509), 1355-1359.
DOI: 10.1126/science.abb5036originally published online August 20, 2020

Coupling transcription and translation

In bacteria, the rate of transcription of messenger RNA (mRNA) by RNA polymerase (RNAP) is coordinated with the rate of translation by the first ribosome behind RNAP on the mRNA. Two groups now present cryo–electron microscopy structures that show how two transcription elongation factors, NusG and NusA, participate in this coupling. Webster *et al.* found that NusG forms a bridge between RNAP and the ribosome when they are separated by mRNA. With shortened mRNA, NusG no longer links RNAP and the ribosome, but the two are oriented so that newly transcribed mRNA can enter the ribosome. Wang *et al.* provide further insight into the effect of mRNA length on the complex structures. They also include NusA and show that the NusG-bridged structure is stabilized by NusA.

Science, this issue p. 1355, p. 1359

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. The title *Science* is a registered trademark of AAAS.

Chapter 9

Conclusion

Cryo-EM structures of different transcription elongation complexes were solved, which provided insights into the mechanisms involved in proofreading. A complete understanding of proofreading mechanisms within transcription is essential, since it shines a light on how RNA Polymerase (RNAP) corrects for errors in incorporating new NTP substrates to the RNA.

Four pre-cleavage complexes with GreA bound to the RNAP secondary channel (SC), each backtracked by a different base, showed that the specific misincorporation, or backtracked base, does not lead to any structural differences within the active site. Of these, the reconstructed map of one pre-cleavage complex reached a resolution of 2.8Å, revealing a particular element of the enzyme's catalytic core, the trigger loop (TL), to be in a novel conformation. To fully assess the impact of GreA on cleavage, a fifth cryo-EM dataset was collected on a 1 nucleotide-backtracked complex in the absence of GreA. The backtracked complex data revealed a continuous range of rotational angles of a structural motif called the swivel module, with the particles form the dataset converging into two states - swivelled and non-swivelled.

Comparisons of the structures of the GreA pre-cleavage complex and the two states of the backtracked complex with similar structures of 3 nucleotide backtracked complexes from Abdelkareem et al. (2019) revealed that the extent of backtracking might influence the range within which the swivel module can rotate. Not only this, but binding of either GreA or GreB to the SC also restrains the swivel module in a specific state, with the

GreA pre-cleavage complex adopting a more swivelled conformation. GreA also appears to directly and indirectly position the elements around the active site for efficient cleavage to take place.

GreA has been known to function primarily through the conserved acidic residues at the tip of its NTC coiled-coil contacting the active site. The structural and biochemical data described here instead suggest a model for GreA-assisted cleavage in which the transcription factor does not solely participate in cleavage through its acidic residues. Rather, it does so through additional interactions involving both the GreA NTD (with RNAP elements within the enzyme core) and its CTD (with the RNAP surface).

Bibliography

Abdelkareem, Mo'men, Charlotte Saint-André, Maria Takacs, Gabor Papai, Corinne Crucifix, Xieyang Guo, Julio Ortiz, and Albert Weixlbaumer. 2019. "Structural Basis of Transcription: RNA Polymerase Backtracking and Its Reactivation." *Molecular Cell* 75, no. 2 (July): 298–309. https://doi.org/10.1016/j.molcel.2019.04.029.

Adams, Paul D., Pavel V. Afonine, Gábor Bunkóczi, Vincent B. Chen, Ian W. Davis, Nathaniel Echols, Jeffrey J. Headd, et al. 2010. "PHENIX: A comprehensive Python-based system for macromolecular structure solution." *Acta Crystallographica Section D: Biological Crystallography* 66 (2): 213–221. https://doi.org/10.1107/S0907444909052925.

Adrian, Marc, Jacques Dubochet, Jean Lepault, and Alasdair W. McDowall. 1984. "Cryo-electron microscopy of viruses." *Nature* 308 (5954): 32–36. https://doi.org/10.1038/308032a0.

Alic, Nazif, Nayla Ayoub, Emilie Landrieux, Emmanuel Favry, Peggy Baudouin-Cornu, Michel Riva, and Christophe Carles. 2007. "Selectivity and proofreading both contribute significantly to the fidelity of RNA polymerase III transcription." *Proceedings of the National Academy of Sciences of the United States of America* 104, no. 25 (June): 10400–10405. https://doi.org/10.1073/pnas.0704116104.

Andersen, Kasper R., Nina C. Leksa, and Thomas U. Schwartz. 2013. "Optimized E. coli expression strain LOBSTR eliminates common contaminants from His-tag purification." *Proteins: Structure, Function and Bioinformatics* 81, no. 11 (November): 1857–1861. https://doi.org/10.1002/prot.24364.

Artsimovitch, Irina, and Robert Landick. 2000. "Pausing by bacterial RNA polymerase is mediated by mechanistically distinct classes of signals." *Proceedings of the National Academy of Sciences of the United States of America* 97, no. 13 (June): 7090–7095. https://doi.org/10.1073/pnas.97.13.7090.

Artsimovitch, Irina, Vladimir Svetlov, Katsuhiko S. Murakami, and Robert Landick. 2003. "Co-overexpression of Escherichia coli RNA polymerase subunits allows isolation and analysis of mutant enzymes lacking lineage-specific sequence insertions." *Journal of Biological Chemistry* 278, no. 14 (April): 12344–12355. https://doi.org/10.1074/jbc.M211214200.

Bai, Xiao Chen, Israel S. Fernandez, Greg McMullan, and Sjors H.W. Scheres. 2013. "Ribosome structures to near-atomic resolution from thirty thousand cryo-EM particles." *eLife* 2013, no. 2 (February). https://doi.org/10.7554/eLife.00461.

Ban, N., P. Nissen, J. Hansen, P. B. Moore, and T. A. Steitz. 2000. "The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution." *Science* 289, no. 5481 (August): 905–920. https://doi.org/10.1126/science.289.5481.905.

Bar-Nahum, Gil, Vitaly Epshtein, Andrei E. Ruckenstein, Ruslan Rafikov, Arkady Mustaev, and Evgeny Nudler. 2005. "A ratchet mechanism of transcription elongation and its control." *Cell* 120, no. 2 (January): 183–193. https://doi.org/10.1016/j.cell.2004.11.045.

Barba-Aliaga, Marina, Paula Alepuz, and José E. Pérez-Ortín. 2021. *Eukaryotic RNA Polymerases: The Many Ways to Transcribe a Gene,* April. https://doi.org/10.3389/fmolb.2021.663209.

Beese, Lorena S, and Thomas A Steitz. 1991. *Structural basis for the 3' 5' exonuclease activity of Escherichia coli DNA polymerase 1: a two metal ion mechanism.* Technical report 1.

Belogurov, Georgiy A., Rachel A. Mooney, Vladimir Svetlov, Robert Landick, and Irina Artsimovitch. 2009. "Functional specialization of transcription elongation factors." *EMBO Journal* 28, no. 2 (January): 112–122. https://doi.org/10.1038/emboj.2008.268.

Bepler, Tristan, Andrew Morin, Micah Rapp, Julia Brasch, Lawrence Shapiro, Alex J. Noble, and Bonnie Berger. 2019. "Positive-unlabeled convolutional neural networks for particle picking in cryo-electron micrographs." *Nature Methods* 16, no. 11 (November): 1153–1160. https://doi.org/10.1038/s41592-019-0575-8.

Bokori-Brown, Monika, Thomas G. Martin, Claire E. Naylor, Ajit K. Basak, Richard W. Titball, and Christos G. Savva. 2016. "Cryo-EM structure of lysenin pore elucidates membrane insertion by an aerolysin family protein." *Nature Communications* 7 (April). https://doi.org/10.1038/ncomms11293.

Borukhov, Bergei, Valery Bagitov, and Alex Goldfarb. 1993. *Transcript Cleavage Factors from E. coli.* Technical report.

Borukhov, S., A. Polyakov, V. Nikiforov, and A. Goldfarb. 1992. "GreA protein: A transcription elongation factor from Escherichia coli." *Proceedings of the National Academy of Sciences of the United States of America* 89 (19): 8899–8902. https://doi.org/10.1073/pnas.89.19.8899.

Brilot, Axel F., James Z. Chen, Anchi Cheng, Junhua Pan, Stephen C. Harrison, Clinton S. Potter, Bridget Carragher, Richard Henderson, and Nikolaus Grigorieff. 2012. "Beam-induced motion of vitrified specimen on holey carbon film." *Journal of Structural Biology* 177, no. 3 (March): 630–637. https://doi.org/10.1016/j.jsb.2012.02.003.

Browning, Douglas F., and Stephen J.W. Busby. 2016. *Local and global regulation of transcription initiation in bacteria,* 10, October. https://doi.org/10.1038/nrmicro.2016.103.

Bubunenko, Mikhail G., Carolyn B. Court, Alison J. Rattray, Deanna R. Gotte, Maria L. Kireeva, Jorge A. Irizarry-Caro, Xintian Li, et al. 2017. "A Cre transcription fidelity reporter identifies GreA as a major RNA proofreading factor in Escherichia coli." *Genetics* 206, no. 1 (May): 179–187. https://doi.org/10.1534/genetics.116.198960.

Burmann, Björn M., Kristian Schweimer, Xiao Luo, Markus C. Wahl, Barbara L. Stitt, Max E. Gottesman, and Paul Rösch. 2010. "A NusE:NusG complex links transcription and translation." *Science* 328, no. 5977 (April): 501–504. https://doi.org/10.1126/science.1184953.

Burova, E., S. C. Hung, V. Sagitov, B. L. Stitt, and M. E. Gottesman. 1995. "Escherichia coli NusG protein stimulates transcription elongation rates in vivo and in vitro." *Journal of Bacteriology* 177, no. 5 (March): 1388–1392. https://doi.org/10.1128/jb.177.5.1388-1392.1995.

Byrne, R., J. G. Levin, H. A. Bladen, and M. W. Nirenberg. 1964. "The in vitro Formation of a DNA-Ribosome Complex." *Proceedings of the National Academy of Sciences of the United States of* 52, no. 1 (July): 140–148. https://doi.org/10.1073/pnas.52.1.140.

Chakraborty, Anirban, Dongye Wang, Yon W. Ebright, You Korlann, Ekaterine Kortkhonjia, Taiho Kim, Saikat Chowdhury, et al. 2012. "Opening and closing of the bacterial RNA polymerase clamp." *Science* 337, no. 6094 (August): 591–595. https://doi.org/10.1126/science.1218716.

Chen, James, Alex J. Noble, Jin Young Kang, and Seth A. Darst. 2019. "Eliminating effects of particle adsorption to the air/water interface in single-particle cryo-electron microscopy: Bacterial RNA polymerase and CHAPSO." *Journal of Structural Biology: X* 1 (January). https://doi.org/10.1016/j.yjsbx.2019.100005.

Cheung, Alan C.M., and Patrick Cramer. 2011. "Structural basis of RNA polymerase II backtracking, arrest and reactivation." *Nature* 471, no. 7337 (March): 249–253. https://doi.org/10.1038/nature09785.

Churchman, L. Stirling, and Jonathan S. Weissman. 2011. "Nascent transcript sequencing visualizes transcription at nucleotide resolution." *Nature* 469, no. 7330 (January): 368–373. https://doi.org/10.1038/nature09652.

Cramer, P., D. A. Bushnell, and R. D. Kornberg. 2001. "Structural basis of transcription: RNA polymerase II at 2.8 Ångstrom resolution." *Science* 292, no. 5523 (June): 1863–1876. https://doi.org/10.1126/science.1059493.

Cramer, Patrick. 2019. *Organization and regulation of gene transcription,* 7772, September. https://doi.org/10.1038/s41586-019-1517-4.

Crick, F. H. 1958. "On protein synthesis." *Symposia of the Society for Experimental Biology* 12:138–163.

Demo, Gabriel, Aviram Rasouly, Nikita Vasilyev, Vladimir Svetlov, Anna B Loveland, Ruben Diaz-Avalos, Nikolaus Grigorieff, Evgeny Nudler, and Andrei A Korostelev. 2017. "Structure of RNA polymerase bound to ribosomal 30S subunit," https://doi.org/10.7554/eLife.28560.001.

Dubochet, J., and A. W. McDowall. 1981. "Vitrification of Pure Water for Electron Microscopy." *Journal of Microscopy* 124, no. 3 (December): 3–4. https://doi.org/10.1111/j.1365-2818.1981.tb02483.x.

Dutta, Dipak, Konstantin Shatalin, Vitaly Epshtein, Max E. Gottesman, and Evgeny Nudler. 2011. "Linking RNA polymerase backtracking to genome instability in E. coli." *Cell* 146, no. 4 (August): 533–543. https://doi.org/10.1016/j.cell.2011.07.034.

Ebright, R. H. 2000. *RNA polymerase: Structural similarities between bacterial RNA polymerase and eukaryotic RNA polymerase II,* 5, December. https://doi.org/10.1006/jmbi.2000.4309.

Eigen, Manfred. 1971. "Selforganization of matter and the evolution of biological macromolecules." *Die Naturwissenschaften* 58 (10): 465–523. https://doi.org/10.1007/BF00623322.

Ekland, E. H., and D. P. Bartel. 1996. "RNA-catalysed RNA polymerization using nucleoside triphosphates." *Nature* 382 (6589): 373–376. https://doi.org/10.1038/382373a0.

Emsley, P, B Lohkamp, W G Scott, and K Cowtan. n.d. "Biological Crystallography Features and development of Coot," https://doi.org/10.1107/S0907444910007493.

Epshtein, Vitaly, Francine Toulmé, A. Rachid Rahmouni, Sergei Borukhov, and Evgeny Nudler. 2003. "Transcription through the roadblocks: The role of RNA polymerase cooperation." *EMBO Journal* 22, no. 18 (September): 4719–4727. https://doi.org/10.1093/emboj/cdg452.

Erickson, H P, and A K Lug. 1971. "Measurement and compensation of defocusing and aberrations by Fourier processing of electron micrographs." *Philosophical Transactions of the Royal Society of London. B, Biological Sciences* 261, no. 837 (May): 105–118. https://doi.org/10.1098/rstb.1971.0040.

Fan, Haitian, Adam B Conn, Preston B Williams, Stephen Diggs, Joseph Hahm, Howard B Gamper, Ya-Ming Hou, Seán E O'leary, Yinsheng Wang, and Gregor M Blaha. 2017. "Transcription-translation coupling: direct interactions of RNA polymerase with ribosomes and ribosomal subunits." *Nucleic Acids Research* 45:11043–11055. https://doi.org/10.1093/nar/gkx719.

Frank, J., W. Goldfarb, D. Eisenberg, and T. S. Baker. 1978. "Reconstruction of glutamine synthetase using computer averaging." *Ultramicroscopy* 3, no. C (January): 283–290. https://doi.org/10.1016/S0304-3991(78)80038-2.

Furman, Ran, Oleg V. Tsodikov, Yuri I. Wolf, and Irina Artsimovitch. 2013. "An insertion in the catalytic trigger loop gates the secondary channel of RNA polymerase." *Journal of Molecular Biology* 425, no. 1 (January): 82–93. https://doi.org/10.1016/j.jmb.2012.11.008.

Gentry, D., H. Xiao, R. Burgess, and M. Cashel. 1991. *The omega subunit of Escherichia coli K-12 RNA polymerase is not required for stringent RNA control in vivo,* 12. https://doi.org/10.1128/jb.173.12.3901-3903.1991.

Gilbert, Walter. 1986. "The RNA world." *Nature* 319:618–undefined.

Goddard, Thomas D., Conrad C. Huang, Elaine C. Meng, Eric F. Pettersen, Gregory S. Couch, John H. Morris, and Thomas E. Ferrin. 2018. "UCSF ChimeraX: Meeting modern challenges in visualization and analysis." *Protein Science* 27, no. 1 (January): 14–25. https://doi.org/10.1002/pro.3235.

Goodfellow, Sarah J., and Joost C.B.M. Zomerdijk. 2013. "Basic mechanisms in RNA polymerase I transcription of the ribosomal RNA genes." *Sub-Cellular Biochemistry* 61:211–236. https://doi.org/10.1007/978-94-007-4525-4{_}10.

Gruber, Tanja M., and Carol A. Gross. 2003. *Multiple Sigma Subunits and the Partitioning of Bacterial Transcription Space,* November. https://doi.org/10.1146/annurev.micro.57.030502.090913.

Guo, Xieyang, Alexander G. Myasnikov, James Chen, Corinne Crucifix, Gabor Papai, Maria Takacs, Patrick Schultz, and Albert Weixlbaumer. 2018. "Structural Basis for NusA Stabilized Transcriptional Pausing." *Molecular Cell* 69, no. 5 (March): 816–827. https://doi.org/10.1016/j.molcel.2018.02.008.

Hanske, Jonas, Yashar Sadian, and Christoph W. Müller. 2018. *The cryo-EM resolution revolution and transcription complexes,* October. https://doi.org/10.1016/j.sbi.2018.07.002.

Hausner, Winfried, Udo Lange, and Meike Musfeldt. 2000. "Transcription factor S, a cleavage induction factor of the archaeal RNA polymerase." *Journal of Biological Chemistry* 275, no. 17 (April): 12393–12399. https://doi.org/10.1074/jbc.275.17.12393.

Izban, M. G., and D. S. Luse. 1993. "SII-facilitated transcript cleavage in RNA polymerase II complexes stalled early after initiation occurs in primarily dinucleotide increments." *Journal of Biological Chemistry* 268, no. 17 (June): 12864–12873. https://doi.org/10.1016/s0021-9258(18)31467-4.

Jakobi, Arjen J., Matthias Wilmanns, and Carsten Sachse. 2017. "Model-based local density sharpening of cryo-EM maps." *eLife* 6 (October). https://doi.org/10.7554/eLife.27131.

Joyce, Gerald F. 2002. *The antiquity of RNA-based evolution,* 6894, July. https://doi.org/10.1038/418214a.

Kang, Jin Young, Paul Dominic, B Olinares, James Chen, Elizabeth A Campbell, Arkady Mustaev, Brian T Chait, Max E Gottesman, and Seth A Darst. 2017. "Structural basis of transcription arrest by coliphage HK022 Nun in an Escherichia coli RNA polymerase elongation complex," https://doi.org/10.7554/eLife.25478.001.

Kang, Jin Young, Eliza Llewellyn, James Chen, Paul Dominic B. Olinares, Joshua Brewer, Brian T. Chait, Elizabeth A. Campbell, and Seth A. Darst. 2021. "Structural basis for transcription complex 1 disruption by the mfd translocase." *eLife* 10 (January): 1–86. https://doi.org/10.7554/eLife.62117.

Kang, Jin Young, Tatiana V. Mishanina, Michael J. Bellecourt, Rachel Anne Mooney, Seth A. Darst, and Robert Landick. 2018. "RNA Polymerase Accommodates a Pause RNA Hairpin by Global Conformational Rearrangements that Prolong Pausing." *Molecular Cell* 69, no. 5 (March): 802–815. https://doi.org/10.1016/j.molcel.2018.01.018.

Kang, Jin Young, Tatiana V. Mishanina, Robert Landick, and Seth A. Darst. 2019. *Mechanisms of Transcriptional Pausing in Bacteria,* 20, September. https://doi.org/10.1016/j.jmb.2019.07.017.

Kang, Jin Young, Rachel Anne Mooney, Yuri Nedialkov, Jason Saba, Tatiana V. Mishanina, Irina Artsimovitch, Robert Landick, and Seth A. Darst. 2018. "Structural Basis for Transcript Elongation Control by NusG Family Universal Regulators." *Cell* 173, no. 7 (June): 1650–1662. https://doi.org/10.1016/j.cell.2018.05.017.

Kettenberger, Hubert, Karim Jean Armache, and Patrick Cramer. 2003. "Architecture of the RNA polymerase II-TFIIS complex and implications for mRNA cleavage." *Cell* 114, no. 3 (August): 347–357. https://doi.org/10.1016/S0092-8674(03)00598-1.

Kettenberger, Hubert, Karim Jean Armache, and Patrick Cramer. 2004. "Complete RNA polymerase II elongation complex structure and its interactions with NTP and TFIIS." *Molecular Cell* 16, no. 6 (December): 955–965. https://doi.org/10.1016/j.molcel.2004.11.040.

Kireeva, Maria L., and Mikhail Kashlev. 2009. "Mechanism of sequence-specific pausing of bacterial RNA polymerase." *Proceedings of the National Academy of Sciences of the United States of America* 106, no. 22 (June): 8900–8905. https://doi.org/10.1073/pnas.0900407106.

Kohler, R, R A Mooney, D J Mills, R Landick, and P Cramer. 2017. "Architecture of a transcribing-translating expressome." *Science* 356:194–197.

Komissarova, Natalia, and Mikhail Kashlev. 1997. "Transcriptional arrest: Escherichia coli RNA polymerase translocates backward, leaving the 3′ end of the RNA intact and extruded." *Proceedings of the National Academy of Sciences of the United States of America* 94, no. 5 (March): 1755–1760. https://doi.org/10.1073/pnas.94.5.1755.

Korkhin, Yakov, Ulug M Unligil, Otis Littlefield, Pamlea J Nelson, David I Stuart, Paul B Sigler, Stephen D Bell, and Nicola G. A Abrescia. 2009. "Evolution of Complex RNA Polymerases: The Complete Archaeal RNA Polymerase Structure." Edited by Martin Egli. *PLoS Biology* 7, no. 5 (May): e1000102. https://doi.org/10.1371/journal.pbio.1000102.

Koulich, Dmitry, Vadim Nikiforov, and Sergei Borukhov. 1998. *Distinct Functions of N and C-terminal domains of GreA, an Escherichia coli Transcript Cleavage Factor.* Technical report.

Koulich, Dmitry, Marianna Orlova, Arun Malhotra, Andrej Sali, Seth A Darst, and Sergei Borukhov. 1997. *Domain Organization of Escherichia coli Transcript Cleavage Factors GreA and GreB.* Technical report 11.

Krummel, Barbara, and Michael J. Chamberlin. 1992. "Structural analysis of ternary complexes of Escherichia coli RNA polymerase. Deoxyribonuclease I footprinting of defined complexes." *Journal of Molecular Biology* 225, no. 2 (May): 239–250. https://doi.org/10.1016/0022-2836(92)90918-A.

Kühlbrandt, Werner. 2014. *The resolution revolution,* 6178, March. https://doi.org/10.1126/science.1251652.

Landick, R., J. Carey, and C. Yanofsky. 1985. "Translation activates the paused transcription complex and restores transcription of the trp operon leader region." *Proceedings of the National Academy of Sciences of the United States of America* 82, no. 14 (July): 4663–4667. https://doi.org/10.1073/pnas.82.14.4663.

Lane, William J., and Seth A. Darst. 2010a. “Molecular Evolution of Multisubunit RNA Polymerases: Sequence Analysis.” *Journal of Molecular Biology* 395, no. 4 (January): 671–685. https://doi.org/10.1016/j.jmb.2009.10.062.

Lane, William J., and Seth A. Darst. 2010b. “Molecular Evolution of Multisubunit RNA Polymerases: Structural Analysis.” *Journal of Molecular Biology* 395, no. 4 (January): 686–704. https://doi.org/10.1016/j.jmb.2009.10.063.

Lange, Udo, and Winfried Hausner. 2004. “Transcriptional fidelity and proofreading in Archaea and implications for the mechanism of TFS-induced RNA cleavage.” *Molecular Microbiology* 52, no. 4 (May): 1133–1143. https://doi.org/10.1111/j.1365-2958.2004.04039.x.

Langer, Doris, and Wolfram Zillig. 1993. *Putative tflls gene of sulfolobus acidocaldarius encoding an archaeal transcription elongation factor is situated directly downstream of the gene for a small subunit of DNA-dependent RNA polymerase,* 9, May. https://doi.org/10.1093/nar/21.9.2251.

Laptenko, Oleg, Jookyung Lee, Ivan Lomakin, and Sergei Borukhov. 2003. “Transcript cleavage factors GreA and GreB act as transient catalytic components of RNA polymerase.” *EMBO Journal* 22, no. 23 (December): 6322–6334. https://doi.org/10.1093/emboj/cdg610.

Larson, Matthew H., Rachel A. Mooney, Jason M. Peters, Tricia Windgassen, Dhananjaya Nayak, Carol A. Gross, Steven M. Block, William J. Greenleaf, Robert Landick, and Jonathan S. Weissman. 2014. “A pause sequence enriched at translation start sites drives transcription dynamics in vivo.” *Science* 344, no. 6187 (May): 1042–1047. https://doi.org/10.1126/science.1251871.

Li, J., R. Horwitz, S. McCracken, and J. Greenblatt. 1992. “NusG, a new Escherichia coli elongation factor involved in transcriptional antitermination by the N protein of phage λ.” *Journal of Biological Chemistry* 267, no. 9 (March): 6012–6019. https://doi.org/10.1016/s0021-9258(18)42655-5.

Li, Xueming, Paul Mooney, Shawn Zheng, Christopher R. Booth, Michael B. Braunfeld, Sander Gubbens, David A. Agard, and Yifan Cheng. 2013. “Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM.” *Nature Methods* 10, no. 6 (April): 584–590. https://doi.org/10.1038/nmeth.2472.

Liebschner, Dorothee, Pavel V. Afonine, Matthew L. Baker, Gábor Bunkoczi, Vincent B. Chen, Tristan I. Croll, Bradley Hintze, et al. 2019. “Macromolecular structure determination using X-rays, neutrons and electrons: Recent developments in Phenix.” *Acta Crystallographica Section D: Structural Biology* 75, no. 10 (October): 861–877. https://doi.org/10.1107/S2059798319011471.

Linn, T., and J. Greenblatt. 1992. “The NusA and NusG proteins of Escherichia coli increase the in vitro readthrough frequency of a transcriptional attenuator preceding the gene for the β subunit of RNA polymerase.” *Journal of Biological Chemistry* 267, no. 3 (January): 1449–1454. https://doi.org/10.1016/s0021-9258(18)45966-2.

Liu, Kebin, Yuying Zhang, Konstantin Severinov, Asis Das, and Michelle M. Hanna. 1996. “Role of Escherichia coli RNA polymerase alpha subunit in modulation of pausing, termination and anti-termination by the transcription elongation factor NusA.” *EMBO Journal* 15, no. 1 (January): 150–161. https://doi.org/10.1002/j.1460-2075.1996.tb00343.x.

Mason, S. W., J. Li, and J. Greenblatt. 1992. “Host factor requirements for processive antitermination of transcription and suppression of pausing by the N protein of bacteriophage λ.” *Journal of Biological Chemistry* 267, no. 27 (September): 19418–19426. https://doi.org/10.1016/s0021-9258(18)41792-9.

Mazumder, Abhishek, Miaoxin Lin, Achillefs N. Kapanidis, and Richard H. Ebright. 2020. “Closing and opening of the RNA polymerase trigger loop.” *Proceedings of the National Academy of Sciences of the United States of America* 117, no. 27 (June): 15642–15649. https://doi.org/10.1073/pnas.1920427117.

Miller, O. L., Barbara A. Hamkalo, and C. A. Thomas. 1970. “Visualization of bacterial genes in action.” *Science* 169 (3943): 392–395. https://doi.org/10.1126/science.169.3943.392.

Mishanina, Tatiana V., Michael Z. Palo, Dhananjaya Nayak, Rachel A. Mooney, and Robert Landick. 2017. “Trigger loop of RNA polymerase is a positional not acid-base, catalyst for both transcription and proofreading.” *Proceedings of the National Academy of Sciences of the United States of America* 114, no. 26 (June): E5103–E5112. https://doi.org/10.1073/pnas.1702383114.

Morin, Paul E., Donald E. Awrey, Aled M. Edwards, and Cheryl H. Arrowsmith. 1996. “Elongation factor TFIIS contains three structural domains: Solution structure of domain II.” *Proceedings of the National Academy of Sciences of the United States of America* 93, no. 20 (October): 10604–10608. https://doi.org/10.1073/pnas.93.20.10604.

Mukherjee, Kakoli, Hiroki Nagai, Nobuo Shimamoto, and Dipankar Chatterji. 1999. “GroEL is involved in activation of escherichia coli RNA polymerase devoid of the ω subunit in vivo.” *European Journal of Biochemistry* 266, no. 1 (November): 228–235. https://doi.org/10.1046/j.1432-1327.1999.00848.x.

Murakami, Katsuhiko, Makoto Kimura, Jeffrey T. Owens, Claude F. Meares, and Akira Ishihama. 1997. “The two α subunits of Escherichia coli RNA polymerase are asymmetrically arranged and contact different halves of the DNA upstream element.” *Proceedings of the National Academy of Sciences of the United States of America* 94, no. 5 (March): 1709–1714. https://doi.org/10.1073/pnas.94.5.1709.

Murakami, Katsuhiko S., Shoko Masuda, and Seth A. Darst. 2002. “Structural basis of transcription initiation: RNA polymerase holoenzyme at 4 Å resolution.” *Science* 296, no. 5571 (May): 1280–1284. https://doi.org/10.1126/science.1069594.

Noeske, Jonas, Michael R. Wasserman, Daniel S. Terry, Roger B. Altman, Scott C. Blanchard, and Jamie H.D. Cate. 2015. “High-resolution structure of the Escherichia coli ribosome.” *Nature Structural and Molecular Biology* 22, no. 4 (April): 336–341. https://doi.org/10.1038/nsmb.2994.

Nudler, Evgeny, Alex Goldfarb, and Mikhail Kashlev. 1994. “Discontinuous mechanism of transcription elongation.” *Science* 265, no. 5173 (August): 793–796. https://doi.org/10.1126/science.8047884.

Nudler, Evgeny, Arkady Mustaev, Evgeny Lukhtanov, and Alex Goldfarb. 1997. “The RNA-DNA hybrid maintains the register of transcription by preventing backtracking of RNA polymerase.” *Cell* 89, no. 1 (April): 33–41. https://doi.org/10.1016/s0092-8674(00)80180-4.

Orlicky, Stephen M., Phan T. Tran, Michael H. Sayre, and Aled M. Edwards. 2001. “Dissociable Rpb4-Rpb7 Subassembly of RNA Polymerase II Binds to Single-strand Nucleic Acid and Mediates a Post-recruitment Step in Transcription Initiation.” *Journal of Biological Chemistry* 276, no. 13 (March): 10097–10102. https://doi.org/10.1074/jbc.M003165200.

Orlova, Marianna, Janet Newlands, Asis Das, Alex Goldfarb, and Sergei Borukhov. 1995. “Intrinsic transcript cleavage activity of RNA polymerase.” *Proceedings of the National Academy of Sciences of the United States of America* 92, no. 10 (May): 4596–4600. https://doi.org/10.1073/pnas.92.10.4596.

Palovcak, Eugene, Feng Wang, Shawn Q. Zheng, Zanlin Yu, Sam Li, Miguel Betegon, David Bulkley, David A. Agard, and Yifan Cheng. 2018. “A simple and robust procedure for preparing graphene-oxide cryo-EM grids.” *Journal of Structural Biology* 204, no. 1 (October): 80–84. https://doi.org/10.1016/j.jsb.2018.07.007.

Paul, Brian J., Melanie M. Barker, Wilma Ross, David A. Schneider, Cathy Webb, John W. Foster, and Richard L. Gourse. 2004. “DksA: A critical component of the transcription initiation machinery that potentiates the regulation of rRNA promoters by ppGpp and the initiating NTP.” *Cell* 118, no. 3 (August): 311–322. https://doi.org/10.1016/j.cell.2004.07.009.

Pecoraro, Vincent L., Jeffrey D. Hermes, and W. W. Cleland. 1984. “Stability Constants of Mg2+ and Cd2+ Complexes of Adenine Nucleotides and Thionucleotides and Rate Constants for Formation and Dissociation of MgATP and MgADP.” *Biochemistry* 23 (22): 5262–5271. https://doi.org/10.1021/bi00317a026.

Penczek, Pawel A., Robert A. Grassucci, and Joachim Frank. 1994. “The ribosome at improved resolution: New techniques for merging and orientation refinement in 3D cryo-electron microscopy of biological particles.” *Ultramicroscopy* 53, no. 3 (March): 251–270. https://doi.org/10.1016/0304-3991(94)90038-8.

Perederina, Anna, Vladimir Svetlov, Marina N Vassylyeva, Tahir H Tahirov, Shigeyuki Yokoyama, Irina Artsimovitch, and Dmitry G Vassylyev. 2004. *Regulation through the Secondary Channel-Structural Framework for ppGpp-DksA Synergism during Transcription.* Technical report.

Pettersen, Eric F., Thomas D. Goddard, Conrad C. Huang, Elaine C. Meng, Gregory S. Couch, Tristan I. Croll, John H. Morris, and Thomas E. Ferrin. 2021. “UCSF ChimeraX: Structure visualization for researchers, educators, and developers.” *Protein Science* 30, no. 1 (January): 70–82. https://doi.org/10.1002/pro.3943.

Proshkin, Sergey, A. Rachid Rahmouni, Alexander Mironov, and Evgeny Nudler. 2010. “Cooperation between translating ribosomes and RNA polymerase in transcription elongation.” *Science* 328, no. 5977 (April): 504–508. https://doi.org/10.1126/science.1184939.

Punjani, Ali, John L. Rubinstein, David J. Fleet, and Marcus A. Brubaker. 2017. “CryoSPARC: Algorithms for rapid unsupervised cryo-EM structure determination.” *Nature Methods* 14, no. 3 (February): 290–296. https://doi.org/10.1038/nmeth.4169.

Ramakrishnan, V. 2002. *Ribosome structure and the mechanism of translation,* 4, February. https://doi.org/10.1016/S0092-8674(02)00619-0.

Roghanian, Mohammad, Yulia Yuzenkova, and Nikolay Zenkin. 2011. “Controlled interplay between trigger loop and Gre factor in the RNA polymerase active centre.” *Nucleic Acids Research* 39, no. 10 (May): 4352–4359. https://doi.org/10.1093/nar/gkq1359.

Rohou, Alexis, and Nikolaus Grigorieff. 2015. “CTFFIND4: Fast and accurate defocus estimation from electron micrographs.” *Journal of Structural Biology* 192, no. 2 (November): 216–221. https://doi.org/10.1016/j.jsb.2015.08.008.

Ruff, Emily F., M. Thomas Record, and Irina Artsimovitch. 2015. *Initial events in bacterial transcription initiation,* 2, May. https://doi.org/10.3390/biom5021035.

Russo, Christopher J, and Lori A Passmore. 2016a. *Specimen preparation for high-resolution cryo-EM.* Technical report.

Russo, Christopher J., and Lori A. Passmore. 2016b. “Ultrastable gold substrates: Properties of a support for high-resolution electron cryomicroscopy of biological specimens.” *Journal of Structural Biology* 193, no. 1 (January): 33–44. https://doi.org/10.1016/j.jsb.2015.11.006.

Sanchez-Garcia, Ruben, Josue Gomez-Blanco, Ana Cuervo, Jose Maria Carazo, Carlos Oscar S. Sorzano, and Javier Vargas. 2020. *DeepEMhancer: A deep learning solution for cryo-EM volume post-processing,* June. https://doi.org/10.1101/2020.06.12.148296.

Santos-Beneit, Fernando. 2015. *The Pho regulon: A huge regulatory network in bacteria,* APR, April. https://doi.org/10.3389/fmicb.2015.00402.

Schluenzen, Frank, Ante Tocilj, Raz Zarivach, Joerg Harms, Marco Gluehmann, Daniela Janell, Anat Bashan, et al. 2000. “Structure of functionally activated small ribosomal subunit at 3.3 Å resolution.” *Cell* 102, no. 5 (September): 615–623. https://doi.org/10.1016/S0092-8674(00)00084-2.

Schmidt, Martin C., and Michael J. Chamberlin. 1987. “nusA Protein of Escherichia coli is an efficient transcription termination factor for certain terminator sites.” *Journal of Molecular Biology* 195, no. 4 (June): 809–818. https://doi.org/10.1016/0022-2836(87)90486-4.

Sekine, Shun ichi, Yuko Murayama, Vladimir Svetlov, Evgeny Nudler, and Shigeyuki Yokoyama. 2015. “The ratcheted and ratchetable structural states of RNA polymerase underlie multiple transcriptional functions.” *Molecular Cell* 57, no. 3 (February): 408–421. https://doi.org/10.1016/j.molcel.2014.12.014.

Selby, Christopher P, and Aziz Sancar. 1993. “Molecular Mechanism of Transcription-Repair Coupling.” *Science* 260:53–58.

Severinov, Konstantin, Rachel Mooney, Seth A. Darst, and Robert Landick. 1997. “Tethering of the large subunits of Escherichia coli RNA polymerase.” *Journal of Biological Chemistry* 272, no. 39 (September): 24137–24140. https://doi.org/10.1074/jbc.272.39.24137.

Sosunov, V., Ekaterina Sosunova, Arkady Mustaev, Irina Bass, Vadim Nikiforov, and Alex Goldfarb. 2003. “Unified two-metal mechanism of RNA synthesis and degradation by RNA polymerase.” *The EMBO Journal* 22, no. 9 (May): 2234–2244. https://doi.org/10.1093/emboj/cdg193.

Sosunova, Ekaterina, Vasily Sosunov, Vitaly Epshtein, Vadim Nikiforov, and Arkady Mustaev. 2013. “Control of transcriptional fidelity by active center tuning as derived from RNA polymerase endonuclease reaction.” *Journal of Biological Chemistry* 288, no. 9 (March): 6688–6703. https://doi.org/10.1074/jbc.M112.424002.

Sosunova, Ekaterina, Vasily Sosunov, Maxim Kozlov, Vadim Nikiforov, Alex Goldfarb, and Arkady Mustaev. 2003. “Donation of catalytic residues to RNA polymerase active center by transcription factor Gre.” *Proceedings of the National Academy of Sciences of the United States of America* 100, no. 26 (December): 15469–15474. https://doi.org/10.1073/pnas.2536698100.

Stebbins, Charles E., Sergei Borukhov, Marianna Orlova, Andrey Polyakov, Alex Goldfarb, and Seth A. Darst. 1995. “Crystal structure of the GreA transcript cleavage factor from Escherichia coli.” *Nature* 373:636–640.

Sutherland, Catherine, and Katsuhiko S. Murakami. 2018. “An Introduction to the Structure and Function of the Catalytic Core Enzyme of Escherichia coli RNA Polymerase.” *EcoSal Plus* 8, no. 1 (February). https://doi.org/10.1128/ecosalplus.esp-0004-2018.

Symersky, Jindrich, Anna Perederina, Marina N. Vassylyeva, Vladimir Svetlov, Irina Artsimovitch, and Dmitry G. Vassylyev. 2006. “Regulation through the RNA polymerase secondary channel: Structural and functional variability of the coiled-coil transcription factors.” *Journal of Biological Chemistry* 281, no. 3 (January): 1309–1312. https://doi.org/10.1074/jbc.C500405200.

Taylor, Kenneth A., and Robert M. Glaeser. 1974. “Electron diffraction of frozen, hydrated protein crystals.” *Science* 186, no. 4168 (December): 1036–1037. https://doi.org/10.1126/science.186.4168.1036.

Tegunov, Dimitry, and Patrick Cramer. 2019. “Real-time cryo-electron microscopy data preprocessing with Warp.” *Nature Methods* 16, no. 11 (November): 1146–1152. https://doi.org/10.1038/s41592-019-0580-y.

Tjhung, Katrina F., Maxim N. Shokhirev, David P. Horning, and Gerald F. Joyce. 2020. “An RNA polymerase ribozyme that synthesizes its own ancestor.” *Proceedings of the National Academy of Sciences of the United States of America* 117, no. 6 (February): 2906–2913. https://doi.org/10.1073/pnas.1914282117.

Toulokhonov, Innokenti, and Robert Landick. 2006. "The Role of the Lid Element in Transcription by E. coli RNA Polymerase." *Journal of Molecular Biology* 361, no. 4 (August): 644–658. https://doi.org/10.1016/j.jmb.2006.06.071.

Toulokhonov, Innokenti, Jinwei Zhang, Murali Palangat, and Robert Landick. 2007. "A Central Role of the RNA Polymerase Trigger Loop in Active-Site Rearrangement during Transcriptional Pausing." *Molecular Cell* 27, no. 3 (August): 406–419. https://doi.org/10.1016/j.molcel.2007.06.008.

Tripathi, Lakshmi, Yan Zhang, and Zhanglin Lin. 2014. *Bacterial sigma factors as targets for engineered or synthetic transcriptional control,* SEP, September. https://doi.org/10.3389/fbioe.2014.00033.

Vassylyev, Dmitry G., Marina N. Vassylyeva, Jinwei Zhang, Murali Palangat, Irina Artsimovitch, and Robert Landick. 2007. "Structural basis for substrate loading in bacterial RNA polymerase." *Nature* 448, no. 7150 (July): 163–168. https://doi.org/10.1038/nature05931.

Vinella, Daniel, Katarzyna Potrykus, Helen Murphy, and Michael Cashel. 2012. "Effects on growth by changes of the balance between GreA, GreB, and DksA suggest mutual competition and functional redundancy in Escherichia coli." *Journal of Bacteriology* 194, no. 2 (January): 261–273. https://doi.org/10.1128/JB.06238-11.

Vvedenskaya, Irina O., Hanif Vahedian-Movahed, Jeremy G. Bird, Jared G. Knoblauch, Seth R. Goldman, Yu Zhang, Richard H. Ebright, and Bryce E. Nickels. 2014. "Interactions between RNA polymerase and the "core recognition element" counteract pausing." *Science* 344, no. 6189 (June): 1285–1289. https://doi.org/10.1126/science.1253458.

Wang, Dong, David A. Bushnell, Xuhui Huang, Kenneth D. Westover, Michael Levitt, and Roger D. Kornberg. 2009. "Structural basis of transcription: Backtracked RNA Polymerase II at 3.4 Angstrom Resolution." *Science* 324, no. 5931 (May): 1203–1206. https://doi.org/10.1126/science.1168729.

Wang, Dong, David A. Bushnell, Kenneth D. Westover, Craig D. Kaplan, and Roger D. Kornberg. 2006. "Structural Basis of Transcription: Role of the Trigger Loop in Substrate Specificity and Catalysis." *Cell* 127, no. 5 (December): 941–954. https://doi.org/10.1016/j.cell.2006.11.023.

Wang, Jimin, and Peter B. Moore. 2017. "On the interpretation of electron microscopic maps of biological macromolecules." *Protein Science* 26, no. 1 (January): 122–129. https://doi.org/10.1002/pro.3060.

Weixlbaumer, Albert, Katherine Leon, Robert Landick, and Seth A. Darst. 2013. "Structural basis of transcriptional pausing in bacteria." *Cell* 152, no. 3 (January): 431–441. https://doi.org/10.1016/j.cell.2012.12.020.

Werner, Finn, and Dina Grohmann. 2011. "Evolution of multisubunit RNA polymerases in the three domains of life." *Nature Reviews Microbiology* 9, no. 2 (February): 85–98. https://doi.org/10.1038/nrmicro2507.

Wimberly, Brian T., Ditlev E. Brodersen, William M. Clemons, Robert J. Morgan-Warren, Andrew P. Carter, Clemens Vonrheln, Thomas Hartsch, and V. Ramakrishnan. 2000. "Structure of the 30S ribosomal subunit." *Nature* 407, no. 6802 (September): 327–339. https://doi.org/10.1038/35030006.

Yuzenkova, Yulia, and Nikolay Zenkin. 2010. "Central role of the RNA polymerase trigger loop in intrinsic RNA hydrolysis." 107 (24). https://doi.org/10.1073/pnas.0914424107/-/DCSupplemental.

Zakharova, Natalya, Irina Bass, Elena Arsenieva, Vadim Nikiforov, and Konstantin Severinov. 1998. "Mutations in and monoclonal antibody binding to evolutionary hypervariable region of Escherichia coli RNA polymerase β' subunit inhibit transcript cleavage and transcript elongation." *Journal of Biological Chemistry* 273, no. 38 (September): 24912–24920. https://doi.org/10.1074/jbc.273.38.24912.

Zenkin, Nikolay, Yulia Yuzenkova, and Konstantin Severinov. 2006. "Transcript-assisted transcriptional proofreading." *Science* 313, no. 5786 (July): 518–520. https://doi.org/10.1126/science.1127422.

Zhang, Gongyi, Elizabeth A Campbell, Leonid Minakhin, Catherine Richter, Konstantin Severinov, and Seth A Darst. 1999. *Crystal Structure of Thermus aquaticus Core RNA Polymerase at 3.3 A ° Resolution DNA template, and finally releases itself and the com-pleted transcript from the DNA when a specific termina-tion signal is encountered. The current view is that the.* Technical report.

Zhang, Jinwei, Murali Palangat, and Robert Landick. 2010. "Role of the RNA polymerase trigger loop in catalysis and pausing." *Nature Structural and Molecular Biology* 17 (1): 99–105. https://doi.org/10.1038/nsmb.1732.

Zheng, Shawn Q., Eugene Palovcak, Jean Paul Armache, Kliment A. Verba, Yifan Cheng, and David A. Agard. 2017. *MotionCor2: Anisotropic correction of beam-induced motion for improved cryo-electron microscopy,* 4, February. https://doi.org/10.1038/nmeth.4193.

Zhou, Ming, and Julie A. Law. 2015. *RNA Pol IV and V in gene silencing: Rebel polymerases evolving away from Pol II's rules,* October. https://doi.org/10.1016/j.pbi.2015.07.005.

Ayesha Dinshaw EDULJEE

Structural and functional studies on transcriptional proofreading by GreA

Résumé

Universellement, la transcription de l'ADN en ARN, est effectuée par l'ARN polymérase (RNAP). Au cours de ce processus, un substrat NTP incorrect peut être incorporé dans la molécule d'ARN naissante, résultant en la RNAP entre dans un état de rétrogradé et ne peut pas poursuivre l'élongation de l'ARN. Pour se sortir de cet état et reprendre son activité de transcription, la RNAP coupe la partie erronée de l'ARN. L'efficacité de ce processus catalytique de relecture est accrue en présence de facteurs de clivage GreA dans *E. coli*. Les structures ont été obtenues par cryo-EM d'un complexe rétrogradé d'un nucléotide avec et sans le facteur de clivage GreA, montrant l'importance du motif structurel de la RNAP connu sous le nom « trigger loop », et du processus de sélection de GreA parmi d'autres facteurs de transcription structurellement similaires. De plus, les structures ainsi que les résultats des tests de transcription *in vitro* montrent les différentes manières que GreA participe au clivage.

Mots clés : ARN polymérase, transcription, relecture, GreA, cryo-microscopie électronique

Résumé en anglais

The first step of gene expression, transcription of DNA to RNA, is carried out by DNA-dependent RNA polymerase (RNAP). During this process, an incorrect NTP substrate might be incorporated into the growing RNA molecule. When this occurs, RNAP enters a backtracked state in which it cannot continue with elongation of the RNA. To rescue itself from this backtracked state, RNAP cuts off the erroneous portion of the RNA in a catalytic process whose efficiency is increased in the presence of specific cleavage factors (GreA in bacteria, TFIIS in eukaryotes). Cryo-EM structures of a complex backtracked by 1 nucleotide with and without the cleavage factor GreA bound to the secondary channel were used to address questions related to the process of proofreading in E. coli RNAP, specifically those of the importance of the RNAP structural motif known as the trigger loop, and the process of selection for GreA amongst other structurally similar transcription factors. In addition to this, the structural data along with results from in vitro transcription assays show that GreA participates in cleavage in multiple ways.

Key words: RNA polymerase, transcription, proofreading, GreA, cryo-electron microscopy